cogitatio

MEDIA AND COMMUNICATION

AI, Media, and People: The Changing Landscape of User Experiences and Behaviors

Volume 13

2025

Edited by Jeong-Nam Kim and Jaemin Jung

Open Access Journal

ISSN: 2183-2439

Media and Communication, 2025, Volume 13
AI, Media, and People: The Changing Landscape of User Experiences and Behaviors

Published by Cogitatio Press
Rua Fialho de Almeida 14, 2º Esq.,
1070–129 Lisbon
Portugal

Design by Typografia®
http://www.typografia.pt/en/

Cover image: © cottonbro studio from Pexels

Academic Editors
Jeong-Nam Kim (University of Oklahoma)
Jaemin Jung (Korea Advanced Institute of Science and Technology)

Available online at: www.cogitatiopress.com/mediaandcommunication

Table of Contents

Table of Contents

Media and Communication
2025 • Volume 13 • Article 10915
https://doi.org/10.17645/mac.10915

cogitatio

EDITORIAL

Open Access Journal

Rethinking Media Users in the Age of AI and Algorithmic Mediation

Jaemin Jung [1] and Jeong-Nam Kim [1,2]

[1] Moon Soul Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, Republic of Korea
[2] Gaylord College of Journalism and Mass Communication, University of Oklahoma, USA

Correspondence: Jaemin Jung (nettong@kaist.ac.kr)

Submitted: 7 July 2025 **Published:** 30 July 2025

Issue: This editorial is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

We have collected 16 research essays on how artificial intelligence (AI) is reshaping media, communication, and public life. The authors describe and prescribe how people respond to AI in real settings, such as journalists transitioning to algorithmic newsrooms, students utilizing ChatGPT, and policymakers searching for fairness and transparency. Across all articles, trust, ethics, and context should and could surpass AI's technical power. We classify the essays into four groups: AI adoption and professional integration; AI governance, ethics, and societal risk; pseudo-information detection and correction; and data-science methods for opinion and behavior analysis. These essays witness emerging media transformations, hinting at how AI can coevolve with, not replace, human intelligence in everyday mediated and connected life.

Keywords

AI; AI ethics; AI governance; AI trust; algorithm; collective intelligence; media; publics

1. Introduction

AI has evolved from an innovation to a shared habitat, reshaping how we learn, trade, legislate, communicate, and reason with one another. Given this growing influence, we called for original research and critical reflections on how AI benefits its users and how it could and should change the ways in which people and media coevolve.

The 16 essays selected for this thematic issue show that computing power and efficiency alone are not the whole story: Every gain in an AI model's speed or scale is matched by fresh questions of power, trust, and

authenticity. The authors follow journalists navigating "algorithmic newsrooms," students who treat ChatGPT as both coach and crutch, regulators wrestling with spam and bias, and policymakers struggling to anchor the ideal of "transparency" to something measurable. They also expose new chasms: US-centric biases in large language models, rhetorical markers that trafficked Covid-19 lies across French X (previously Twitter), and the imbalance of AI safety regulations in low-income nations. These articles forgo the simplistic tech-utopia or tech-doom forecasts which often surround discussions of AI in media, and instead analyze and propose ways in which people, institutions, and AI can contribute to collective intelligence in ways that create opportunity without sacrificing human values.

We have arranged the collection of articles by subject into four groups, beginning with AI adoption and professional integration. Five articles describe how newsrooms, businesses, universities, and governments adapt and utilize AI tools in everyday work, demonstrating that trust, long-term relationships, and "centaur" skill sets decide who wins and who loses. The second group explores the subjects of governance, ethics, and societal risk. Here, five articles report on policy battles such as global transparency gaps, copyright issues for prompt engineers, AI filtering and moderation of opinion spamming in governmental rulemaking, and safeguards for data-driven persuasion in the Global South. These articles highlight the need for context-sensitive, rights-preserving governance strategies for AI adoption and practice. In the third group, on pseudo-information (J.-N. Kim & Gil de Zúñiga, 2021) detection and correction, three articles examine human–AI collective efforts that could guard against conspiracy theories, false rumors, and pandemic dis—and misinformation. They present and demonstrate the power of real-time information surveillance dashboards, linguistic markers of fake news, and signals of human oversight. The last group covers data-science methods for opinion and behavior analysis of digital publics. Three articles present theory-based development to understand AI-immersed digital publics' new information environments and methodological toolkits in comment mining, synthetic polling, and tensor decomposition. These new theory-method advancements enable researchers and practitioners to map the emotions expressed and embedded in public narratives.

All of these 16 articles and the four subject groups they address articulate a single challenge: the need to construct media systems in which AI or machines amplify rather than undercut human intelligence in everyday democratic life. Below, we highlight the articles' key thoughts, including their prescriptions and proscriptions for the emerging and evolving interactions between AI, media, and people.

2. AI Adoption and Professional Integration

The articles in this section track ways in which journalists, public-sector communicators, businesses, educators, and consumers incorporate AI into their routine work and learning. The studies find widespread optimism concerning efficiency and personalization, but also heightened anxiety over skills gaps, employment security, and ethical drift. Trust, whether in organizations, technologies, or long-term relationships, emerges as the critical currency that turns curiosity into sustained AI use. Hybrid or "centaur" skill sets, transparent design, and two-way engagement are identified as keys to the successful integration of human and machine contributions in these sectors.

S. Oh and Jung (2025) show that journalists now view AI as both a help and a hazard. Worrying that algorithm-based decisions could bleed into editorial judgment and plant bias into coverage, many journalists

are pushing to learn how generative models work and to keep the final story in human hands. They answer with a practical blueprint: Pair coders and journalists working side-by-side at every design step to bake transparency, fairness, and truth-seeking into a "journalistic algorithm" and spread both its cost and knowledge across the industry. By adapting from passive users to active co-creators, the authors argue, journalists and newsrooms can expand the gains offered by AI without surrendering the craft's core values.

Next, H. Lee et al. (2025) map how everyday people approach AI and find that attitudes split into distinct camps: enthusiasts and confident users race to try new tools; balanced and cautious groups weigh gains against missteps; the uninterested tune out. The team tracks how service-specific trust mediates problem recognition and general AI trust, converting curiosity into concrete intent. The study urges firms to ditch one-size-fits-all pitches and instead match messages to each audience segment, as well as to invest in transparent, two-way communication and treat trust as the real engine of uptake, especially in new or foreign markets where proof and reassurance must go hand-in-hand for both businesses and policymakers.

S. K. Lee et al. (2025) show in their article that college students use ChatGPT as both a shortcut and a crutch. Five motivations—novelty, entertainment, guidance, interaction, and peer influence—are linked to actual use, with novelty and entertainment leading but often yielding superficial engagement, while guidance drives deeper tasks. Students rely on ChatGPT most for simplifying complex ideas and less for contentious topics where misinformation persists. Trust in the service, rather than technical skill, best predicts adoption, underscoring the risk of critical dependency. The authors urge instructors to bring structured AI literacy into courses and challenge developers to close reliability gaps, especially for multilingual users, so that ChatGPT augments rather than erodes the process of study.

Tam et al. (2025) trace how people judge companies that roll out AI tools and find that what matters most is not faith in the code but faith in the firm. Trust in the organization magnifies perceived benefits and softens misgivings; without that trust, even a clear payoff cannot calm anxiety. The authors urge companies to spell out why they use AI, what data they touch, and how the system works, and then to back those claims with ethical safeguards. By meeting curiosity with candor, firms can convert hesitant observers into loyal users and keep support alive as algorithms spread across storefronts and apps.

Lastly, Lovari and De Rosa (2025) examine how European government communicators see generative AI as both an opportunity and a minefield. New rules, volatile media cycles, and restless constituents push them to reinvent themselves as "centaur communicators," blending analog judgment with digital precision. This shift demands that they are able to explain what the tools do, flag risks in plain language, and build guardrails that keep transparency and accountability intact. The authors argue that these officials now anchor democratic discourse; by guiding AI rather than trailing it, they can shield citizens from pseudo-information while freeing up time for deeper engagement. Done right, generative AI could allow governments to move faster without surrendering integrity.

3. AI Governance, Ethics, and Societal Risk

Here, AI is treated as a policy object. The included research ranges from Korea's risk amplification around generative AI to cross-regional audits of transparency, from US e-rulemaking experiments to a tiered copyright proposal for AI prompts, and finally to ethical safeguards for AI-powered social media campaigns

in low-income nations. Together, they argue that effective governance must be adaptive, culturally attuned, and explicitly protective of autonomy, equity, and a vibrant public domain.

S. Kim and Jung (2025) track how the discussion of generative AI in Korea has evolved between early 2023 and the middle of 2024. Mining 56,000 news stories and 68,000 user comments and applying the Social Amplification of Risk Framework, they show the two spheres pulled in different directions. Reporters, echoing experts, framed AI as a big-ticket industrial gamble and warned about misinformation, ethics, and sector-wide upheaval in robotics, chips, and smartphones. As pragmatic worries about labor and social misuse outpaced moral panic, the central question of the debate shifted from "what if AI misbehaves?" to "how will AI reshape work?" The study alarms regulators to ditch one-size-fits-all messages and craft policy alongside publics who judge AI through the lenses of their own stakes.

Sebastião and Dias (2025) probe how policymakers frame transparency in AI rules across regions. Their examination of leading ethical charters and draft laws finds near-universal praise for "transparency" but little agreement on its day-to-day meaning. Empty slogans, they warn, widen accountability gaps when coders, vendors, and regulators all share the workload. The authors call for a single yardstick that respects cultural differences but lays out non-negotiable duties: to disclose data inputs, to audit models, and to pin down who takes responsibility when systems fail. The study argues that real AI governance will depend not on rhetoric but on hard, testable standards and the people willing to put them to work.

Next, Perez et al. (2025) put two e-rulemaking prototypes in front of US citizens and measured how different publics react when AI flags and filters opinion spam. The results cut through the hype: AI, by itself, neither raised nor harmed overall approval, but its impact split sharply along problem-solving lines. People who already felt able and motivated to weigh in welcomed the tool; those who sensed constraints or low stakes read the same system as technocratic overreach. The study shows that legitimacy in digital rulemaking rests less on smarter code than on visible, two-way design that treats citizens as partners, not data points.

In the next article, Jon (2025) tackles the copyright gray area around AI-generated works. He classifies prompts by the depth of human creativity (Tier 1—minimal human input; Tier 2—moderate human creativity in prompt design; Tier 3—substantial human creative contribution) and links each tier to a matching level of protection. Jon also flags real-world knots, such as prompt trolling, cross-border enforcement, and the rise of professional prompt engineers, and sketches practical fixes, from simple prompt-registration forms to international cooperation. His model reconciles protecting creativity and fostering innovation, helping lawmakers to adapt old laws to new technologies.

Penh (2025) spotlights the double edge of AI-driven social media in low- and middle-income countries: targeted feeds can spur healthier habits, widen financial access, and stimulate civic action, yet the same algorithms often amplify bias, manipulate opinion, and spread falsehoods where watchdogs are weak. She argues that "do no harm" must shift from slogan to standard and that firms and aid agencies need to audit persuasion tools, invite local voices into rule-setting, invest in AI safety and literacy, and adapt safeguards to each community's politics and culture. Without that groundwork in place, global AI rules risk echoing donor priorities rather than local needs, and vulnerable users may trade autonomy for convenience. With it, however, AI can advance sustainable development goals without losing trust.

4. Pseudo-Information Detection and Correction

Focusing on conspiracy narratives, crisis rumors, and pandemic falsehoods, these articles test the comparative merits of AI and human fact-checkers, introduce a real-time visual analytics platform (SMART 2.0), and isolate linguistic markers of deception on French X. The evidence favors mixed AI and human workflows, human-in-the-loop model refinement, and language-specific heuristics as the most robust shields against information disorder.

First, Lan et al. (2025) tested whether conspiracy-minded readers trust fact checks more when they come from a person or from an algorithm. In a 2 × 2 experiment, human verifications increased and sparked the strongest intent to share corrections, while fully automated stories fared worst. Yet, readers steeped in conspiracies showed higher baseline trust in an AI checker, perhaps because the automated tool seemed less partisan. Positive machine heuristics—shortcuts that label software as objective—fueled this bump but faded once the same readers realized the story itself was machine-written. The pattern suggests that mixed teams, with humans at the front and transparent AI tools in support, can diminish misinformation better than either one alone. To win over skeptics, organizations should tailor this human–machine blend to specific audience traits and give clear signals about who or what wrote each piece.

Hamad et al. (2025) introduce SMART 2.0, a real-time dashboard that pairs social media streams with traffic, weather, and emergency feeds to spot rumors as they form. During the 2024 UK riots, the tool plotted posts on a map, detected sudden bursts of false claims, and traced how rumors jumped from one district to another. Users could tag content on the fly, and the system instantly incorporated those judgments back into its classifier, sharpening accuracy where local slang or context confused automated filters. By cross-checking each claim with official bulletins, SMART 2.0 let reporters, first responders, and researchers separate fact from noise while events unfolded. The team is currently working to implement stronger language models, multi-platform search, network maps of super-spreaders, and multilingual support, steps meant to turn the system from a crisis tracker into an early-warning system for pseudo-information.

To close this section, Chiu et al. (2025) dissect French-language tweets regarding Covid-19 and reveal how word choice telegraphs deception. They flag three tell-tale tactics: hedging phrases that soften claims, pseudo-scientific jargon that dresses them up, and modal verbs that nudge readers without committing the author. This pattern suggests that peddlers of fake news lower the stakes of their assertions, invoke urgency, and lean on French linguistic hierarchies of obligation to slip past suspicion. By tying specific linguistic cues to veracity, the study supplies a filter that works even when ground truths are murky, offering a useful tool for both newsroom monitors and automated detectors. It also equips educators with concrete examples to demonstrate to students how rhetoric, not just facts, shapes what passes for truth online.

5. Data-Science Methods for Opinion and Behavior Analysis

The final group of articles addresses methodological innovation: how large-scale comment mining updates the theory of communicative actions in problem solving's public typologies, LLM-generated synthetic polling reveals both promise and bias, and tensor decomposition makes high-dimensional text patterns interpretable. Each study illustrates how advanced analytics can reveal hidden structures in digital publics, but only if transparency, validation, and cultural calibration keep pace with technical sophistication.

Yeo et al. (2025) push the communicative action in problem solving model into the comment threads of a high-profile entertainment dispute and show it still holds. Analyzing thousands of posts with a theory-guided data science approach, they identify three familiar publics—aware, active, and activist—but upend old assumptions about passivity: When the barrier to entry drops to a click, even aware users argue, curate links, and fend off pseudo-information. Engagement also shifts with time; active publics drive the early burst, then aware users and hard-core activists keep the issue alive. Because these roles blur and evolve, the authors recommend that communicators replace static surveys with real-time analytics and build big-data strategies to tag, track, and talk with publics as they change.

In their article, K. Lee et al. (2025) tested whether large language models can stand in for real polls on South Korea's labor debate. They worked in two ways: one prompted the model to run regressions on actual survey data, while the other asked it to fabricate a full, hypothetical data set. Both identified the broad left-right shape of opinion, yet both also distorted the view. The team argues that careful prompts, local fine-tuning, and full disclosure of AI's role are the price of using these shortcuts. LLMs can speed exploratory work when polls are scarce, but only humans can prevent built-in biases from turning into false headline "findings" that distort public debate.

Finally, Y. Oh and Park (2025) bring tensor decomposition to communication research, turning a black-box task into an accessible one. They feed LIWC features from thousands of online reviews into the PARAFAC2 algorithm and cleanly separate genuine posts from fakes—for instance, heavy use of first-person pronouns often denotes deception attempting to fake intimacy. Unlike standard models, PARAFAC2 handles records of uneven length and still exposes the weights that drive each decision, so scholars can trace how language, emotion, and context interact at scale. They suggest that the same recipe can upgrade social media monitoring, crisis dashboards, and audience research.

6. Conclusion

AI now sits at the core of how news spreads, schools run, governing rules are introduced, and power is contested among people equipped with ICTs and networked broadcasting media. These 16 articles describe the drastic shift that is underway across newsrooms, classrooms, civic forums, and policy institutions. Each proclaims that the speed, scale, and personalization of the evolving interactions between AI, media, and people can lead to benefits only when the people stay in the loop to check facts, question products, and create the tools themselves. Journalists must refine "journalistic algorithms" to protect autonomy; students may tap LLMs for ideas, yet must still reason and know their limits; regulators who utilize AI filters must ensure human co-moderation to secure legitimacy; people can operate dashboards to track rumor cascades in real time. In every case, trust and transparency determine whether AI strengthens or strains digital spheres of public communicative actions, while cultural contexts shape the resulting dynamics, as shown in the K-pop fan communities, stakeholder politicking in US rulemaking, or fake news trafficking in French tweets.

Therefore, in the shifting landscape of experiences and user behaviors, humans are what matter. AI integration must keep humans, not algorithms, in charge of making meaning. We need stronger trust mechanisms: clear disclosure, stricter audits, hybrid professionals (e.g., "centaurs") who both create and critique emerging intelligent systems, and accountable analytics that track bias, tune models to local cues, and allow public correction at every phase.

These moves turn AI from an intimidating, dictating force into a collective intelligence and reduce the risk of runaway polarization, hidden persuasion, and creative lock-in. The path ahead calls for collaboration between engineers, social scientists, journalists, teachers, and lawmakers, and the results will be worth the effort: Media ecosystems will evolve faster and more fairly and offer deeper user experiences without losing their humanity. Three actors—AI, media, and people—will continue to generate complex, hard-to-define interactions. Yet one consensus emerges from the 16 articles: Whatever the process, we must center agency, ethics, and democratic values to ensure that AI enriches, rather than impoverishes, everyday public life.

Acknowledgments

Kim is now affiliated with the Korea Advanced Institute of Science and Technology (KAIST), though most of the work was completed while he was at the University of Oklahoma.

Conflict of Interests

The authors declare no conflict of interests.

LLMs Disclosure

The authors used ChatGPT to outline 16 articles and to ensure grammatical accuracy and proper language usage in the early stages of drafting this guest editorial essay.

References

Chiu, M. M., Morakhovski, A., Wang, Z., & Kim, J. (2025). Detecting Covid-19 fake news on Twitter/X in French: Deceptive writing strategies. *Media and Communication*, *13*, Article 9483. https://doi.org/10.17645/mac.9483

Hamad, M. M., Danala, G., Jentner, W., & Ebert, D. (2025). SMART 2.0: Social media analytics and reporting tool applied to misinformation tracking. *Media and Communication*, *13*, Article 9543. https://doi.org/10.17645/mac.9543

Jon, W. (2025). Prompting creativity: Tiered approach to copyright protection for AI-generated content in the digital age. *Media and Communication*, *13*, Article 9420. https://doi.org/10.17645/mac.9420

Kim, J.-N., & Gil de Zúñiga, H. (2021). Pseudo-information, media, publics, and the failing marketplace of ideas: Theory. *American Behavioral Scientist*, *65*(2), 163–179. https://doi.org/10.1177/000276422095060

Kim, S., & Jung, J. (2025). How generative AI went from innovation to risk: Discussions in the Korean public sphere. *Media and Communication*, *13*, Article 9523. https://doi.org/10.17645/mac.9523

Lan, D., Zhu, Y., Liu, M., & He, C. (2025). AI agency in fact-checking: Role-based machine heuristics and publics' conspiratorial orientation. *Media and Communication*, *13*, Article 9516. https://doi.org/10.17645/mac.9516

Lee, H., Jung, C., Koo, N., Seo, S., Yoo, S., Hong, H., & Jang, Y. (2025). Who wants to try AI? Profiling AI adopters and AI-trusting publics in South Korea. *Media and Communication*, *13*, Article 9639. https://doi.org/10.17645/mac.9639

Lee, K., Park, J., Choi, S., & Lee, C. (2025). Ideology and policy preferences in synthetic data: The potential of LLMs for public opinion analysis. *Media and Communication*, *13*, Article 9677. https://doi.org/10.17645/mac.9677

Lee, S. K., Ryu, J., Jie, Y., & Ma, D. H. (2025). Motivations and affordances of ChatGPT usage for college students' learning. *Media and Communication*, *13*, Article 9508. https://doi.org/10.17645/mac.9508

Lovari, A., & De Rosa, F. (2025). Exploring the challenges of generative AI on public sector communication in Europe. *Media and Communication*, *13*, Article 9644. https://doi.org/10.17645/mac.9644

Oh, S., & Jung, J. (2025). Harmonizing traditional journalistic values with emerging AI technologies: A systematic review of journalists' perception. *Media and Communication*, *13*, Article 9495. https://doi.org/10.17645/mac.9495

Oh, Y., & Park, C. (2025). Unmasking machine learning with tensor decomposition: An illustrative example for media and communication researchers. *Media and Communication*, *13*, Article 9623. https://doi.org/10.17645/mac.9623

Penh, B. (2025). AI-powered social media for development in low- and middle-income countries. *Media and Communication*, *13*, Article 9577. https://doi.org/10.17645/mac.9577

Perez, L. A., Jensen, M. L., Bessarabova, E., Talbert, N., Li, Y., & Zhu, R. (2025). Public segmentation and the impact of AI use in e-rulemaking. *Media and Communication*, *13*, Article 9550. https://doi.org/10.17645/mac.9550

Sebastião, S. P., & Dias, D. F.-M. (2025). AI transparency: A conceptual, normative, and practical frame analysis. *Media and Communication*, *13*, Article 9419. https://doi.org/10.17645/mac.9419

Tam, L., Kim, S., & Gong, Y. (2025). Support for businesses' use of artificial intelligence: Dynamics of trust, distrust, and perceived benefits. *Media and Communication*, *13*, Article 9534. https://doi.org/10.17645/mac.9534

Yeo, S., Kim, J., Kim, J., & Ko, S. (2025). Investigating publics' communicative action in problem solving (CAPS) through data science. *Media and Communication*, *13*, Article 9552. https://doi.org/10.17645/mac.9552

About the Authors

Jaemin Jung (PhD, University of Florida) is dean of the College of Liberal Arts and Convergence Science and a professor at the Moon Soul Graduate School of Future Strategy at the Korea Advanced Institute of Science and Technology (KAIST). His research focuses on media management, media economics, and the impact of AI on journalism and media industries, with a keen interest in exploring how AI technologies are reshaping the landscape of news production and media consumption.

Jeong-Nam Kim (PhD, University of Maryland College Park) is a communication theorist known for the situational theory of problem solving (STOPS) and for his work on cognitive arrest and epistemic inertia. He directs the DaLI (Debiasing & Lay Informatics) Lab, which addresses challenges such as pseudo-information, public bias, and dysfunctional information markets. Currently a KAIST chair professor, he previously held the Gaylord Family Endowed chair at the University of Oklahoma and is a fellow at several international research centers.

Media and Communication
2025 • Volume 13 • Article 9495
https://doi.org/10.17645/mac.9495

cogitatio

ARTICLE

Open Access Journal

Harmonizing Traditional Journalistic Values With Emerging AI Technologies: A Systematic Review of Journalists' Perception

Sangyon Oh [1,2] and Jaemin Jung [2]

[1] Munhwa Broadcasting Corporation, Republic of Korea
[2] Moon Soul Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, Republic of Korea

Correspondence: Jaemin Jung (nettong@kaist.ac.kr)

Submitted: 28 October 2024 **Accepted:** 6 January 2025 **Published:** 24 April 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

This study investigates how news organizations perceive the integration of artificial intelligence (AI) technologies in news production, focusing on the synthesis of traditional journalistic values with AI advancements. By conducting a meta-analysis of 59 scholarly articles published between 2020 and 2024 in the field of journalism, the research examines the perceptions of journalists, editors, and decision-makers regarding AI. The primary research question explores the general findings of previous studies on journalists' perceptions of AI in their workflows and the frameworks used to reconcile AI with journalistic values. The findings indicate that AI is regarded as a transformative tool, enhancing efficiency, effectiveness, and fostering a new organizational culture. However, it raises concerns about costs and job security. Attitudes toward AI are polarized, with optimism about efficiency gains and skepticism due to potential impacts on employment and ethical standards. Three theoretical models—field theory, human–machine communication, and the technology acceptance model—are employed to understand these dynamics, with field theory addressing power shifts and human–machine communication and the technology acceptance model examining human–AI interaction. To effectively integrate AI with journalistic values, the study proposes three strategies: AI technologists should embed journalistic ethics into their processes, journalists should acquire basic AI technical skills, and collaborative platforms should be established to bridge gaps between journalists and technicians. These strategies aim to create a balanced framework where AI-driven news production can uphold essential journalistic standards while embracing technological innovation.

Keywords

artificial intelligence; journalism; journalistic values; newsrooms; organizational culture

1. Introduction

The rapid advancement of artificial intelligence (AI) has significantly influenced various industries, including journalism, by altering organizational strategies and dynamics (de-Lima-Santos & Ceron, 2021; Gelgel, 2020). In journalism, AI encompasses algorithmic processes that automate the creation and distribution of text, images, and videos with minimal human involvement (Carlson, 2015; Moran & Shaikh, 2022). Following the success of tools like the *LA Times*' QuakerBot, which generates earthquake stories in minutes, newsrooms globally are adopting AI-driven automation for tasks such as tagging, story delivery, summarization, and text-to-speech conversion (Motta et al., 2020; Newman, 2021, 2022; Salaverría & de-Lima-Santos, 2020). Due to the advent of AI, agenda-setting, content gathering, production, and news distribution processes have evolved dramatically.(de-Lima-Santos & Ceron, 2021; Örnebring, 2010).

Such technological shifts pose fundamental challenges to the roles and values of journalists. Van Dalen (2012) asserts that "the idea that journalistic tasks can be completely automated clashes with our general understanding of the nature of journalism" (p. 649; see also Moran & Shaikh, 2022; Örnebring, 2010). Automated technologies capable of replacing specific tasks threaten the professional and social identities of human journalists. In the context of newsrooms, AI can be defined as automated systems designed to replicate human cognition (Lindén & Tuulonen, 2019) or as "cognitive technologies" aimed at emulating human intelligence (Chan-Olmsted, 2019, p. 194).

Research on the impact and practical applications of automated algorithms in journalism has gained momentum since the late 2010s (e.g., Lindén, 2017a; Siitonen et al., 2023; Thurman et al., 2017). The 2010s marked the initial phase of AI integration into journalism, primarily emphasizing the technical and procedural aspects of automation. Since then, two major factors have significantly affected the adoption of AI in journalism: the Covid-19 pandemic and the emergence of generative AI. The reduced mobility of people during the Covid-19 period undoubtedly influenced journalism, leading to a greater reliance on AI for news content production. Additionally, the swift development of generative AI tools like ChatGPT and Midjourney, particularly in post-Covid-19, calls for fresh perspectives from journalism organizations and scholars. The commercialization of AI tools such as transcription, translation, and text generation through models like OpenAI's GPT offers innovative ways to integrate technology into journalism (Jones et al., 2022). The new social and technological changes occurring in the 2020s create an environment that necessitates special attention to the adoption of AI in journalism.

Previous research on AI in journalism often compares AI-generated articles to those created by human journalists or examines how AI-related news is framed, typically focusing on AI as a topic within media coverage. This reflects a predominance of technology-oriented studies that highlight AI as a product in journalism. While this approach has generated significant insights, it tends to overlook the human agents—the journalists—who implement AI. The perspectives of newsroom practitioners on AI increasingly shape the evolving values and roles within journalism.

Scholars emphasize the need for empirical data from journalists to comprehend AI's impact on newsroom practices (Carlson, 2015; Kim & Kim, 2018; Lindén, 2017b; Missaoui et al., 2019; Siitonen et al., 2023). Yet, empirical research on journalists' perceptions of AI remains limited (Moran & Shaikh, 2022). This study addresses this gap by systematically reviewing existing research on how journalists have perceived and

adapted to AI since 2020. It synthesizes findings on the practical perceptions, concerns, and challenges journalists encounter in adopting AI, as well as the organizational dynamics that influence skill development, workforce changes, and identity in AI-integrated newsrooms. By consolidating these insights, this study aims to provide a comprehensive understanding that supports journalism's adaptation to AI while upholding its traditional ethics and values.

2. Issues of AI in Journalism From an Organizational Perspective

An organizational perspective in journalism and AI research is essential because AI involves more than just a technology designed for user convenience and ease of use. The values and professional identity of journalism organizations and journalists have long been associated with truthfulness, transparency, and trustworthiness (Komatsu et al., 2020; Kreft et al., 2023; Paik, 2023; Tariq et al., 2024; van Drunen & Fechner, 2022). Therefore, adopting AI as a technology in newsrooms must align with journalists' professional values and their organizational norms.

While integrating AI into newsrooms could benefit journalists, it may also raise fundamental questions regarding the essential role and identity of journalism (Calvo-Rubio & Rojas-Torrijos, 2024; Guanah et al., 2020; Noor & Zafar, 2023; Okocha & Ola-Akuma, 2022). Work-related identity is influenced by the social groups to which people feel they belong and the enactment of specific behaviors typical of those groups, further enhanced by a sense of "social recognition" from society (Ashforth & Schinoff, 2016; Nelson & Irwin, 2014). Therefore, the accepted values of the journalistic profession, journalistic ethics, and journalists' sense of professionalism are issues that must be examined in conjunction with adopting AI in newsrooms.

Given the ethical issues inherent in AI technology, the extent to which automated news stories can faithfully reflect objectivity, autonomy, and the public interest is still being determined. These journalistic values are fundamental in an era where digital technologies significantly impact journalism's ability to fulfill its traditional role. The threat to these values may lead to a crisis in modern society, as Habermas warned, in which the overdevelopment and dominance of instrumental rationality stifle the communicative rationality of the lifeworld (Habermas, 1984).

In this context, this study explores the complexities of integrating AI into journalism by examining the tension between journalistic values and the mechanical nature of AI. The primary goal is to identify frameworks and standards that reconcile the adoption of AI technologies with the values journalism should uphold. To this end, this study primarily relies on an extensive and systematic review of existing research on the topic, given the significant accumulation of excellent studies, particularly in the 2020s. Based on this review, it also aims to help construct an alternative framework facilitating the harmony between AI's technical supremacy and journalistic values. Thus, it seeks to conduct a meta-analysis of current studies to identify overarching findings and suggest some strategies for developing an alternative framework.

The following are the research questions of this study:

> RQ1: What are the general findings of previous studies on journalists' perceptions of AI adoption in their work processes, and why do they favor or oppose its adoption?

RQ2: What frameworks are employed in the previous studies to explain the adoption of AI in journalism, and what strategies could be proposed to construct an alternative framework necessary for integrating AI with journalistic values?

3. Data and Analysis

For a systematic review of current studies, this research collected academic articles on AI in journalism published from 2020 to 2024, utilizing Google Scholar, Web of Science, and Scopus to ensure comprehensive coverage (Calvo-Rubio & Ufarte-Ruiz, 2021; Martín-Martín et al., 2018). To address Google Scholar's less systematic approach (Siitonen et al., 2023), only the top 100 results were included. Searches were performed using keywords like "AI," "artificial intelligence," "automated," "computational," "robot," "algorithm," "technology," and "data," along with "journalism," "journalist," "news," "media," "newsroom," and "news organization" to ensure thematic relevance across diverse topics.

The review period begins in 2020, marking the Covid-19 pandemic as a transformative moment for automated news production. Declared by the WHO in March 2020, the pandemic generated structured data on infection rates that many media outlets utilized as predictable story frames, which accelerated news automation (Danzon-Chambaud, 2021; de-Lima-Santos & Ceron, 2021; Haim, 2022; Kreft et al., 2023; Okocha & Ola-Akuma, 2022; Montaña-Niño & Burgess, 2024). Daily updates on infections and vaccinations further reinforced this shift, providing journalists with abundant data to manage and interpret for public understanding (Burgess et al., 2022; Pentzold et al., 2021).

The initial search yielded numerous studies unrelated to media and journalism. To refine the sample, this study employed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology (Moher et al., 2010). Following the procedures of this method, articles less relevant to the issues of this study were excluded. First, those related to AI and algorithms in business, law, and information systems were screened out by checking their abstracts and keywords. Second, studies that concentrated on news consumers' perspectives were eliminated. Third, meta-analyses and systematic reviews were filtered out. To ensure consistency, conference proceedings and reports were also excluded, focusing solely on peer-reviewed journal articles of empirical studies.

Finally, this study collected 59 empirical studies that offer insights from journalists, experts, and managers directly involved in news production. The selected studies addressed at least one of the following questions:

1. How are AI technologies utilized in newsrooms?
2. What attitudes and evaluations do newsroom members hold regarding AI adoption?
3. How are journalistic values and meanings realized in the context of AI adoption?

Focusing exclusively on empirical research serves dual purposes. First, it grounds those studies as inherently data-driven, ensuring their findings reflect observable phenomena rather than speculative theorization. This is particularly crucial in the context of journalism and AI, where technological adoption and its implications are often context-dependent and shaped by real-world practices. Second, it incorporates insights from news practitioners that are essential to capture the challenges, opportunities, and ethical considerations faced by those at the forefront of AI's integration into journalism. Thus, concentrating on empirical evidence from

previous studies will not only enhance the reliability and relevance of the findings but also contribute to bridging the gap between academic research and industry practices. It highlights the necessity of anchoring scholarly discourse in the lived experiences and operational realities of journalists navigating a rapidly evolving technological landscape.

Figure 1 illustrates the selection process of 59 papers using the PRISMA methodology.

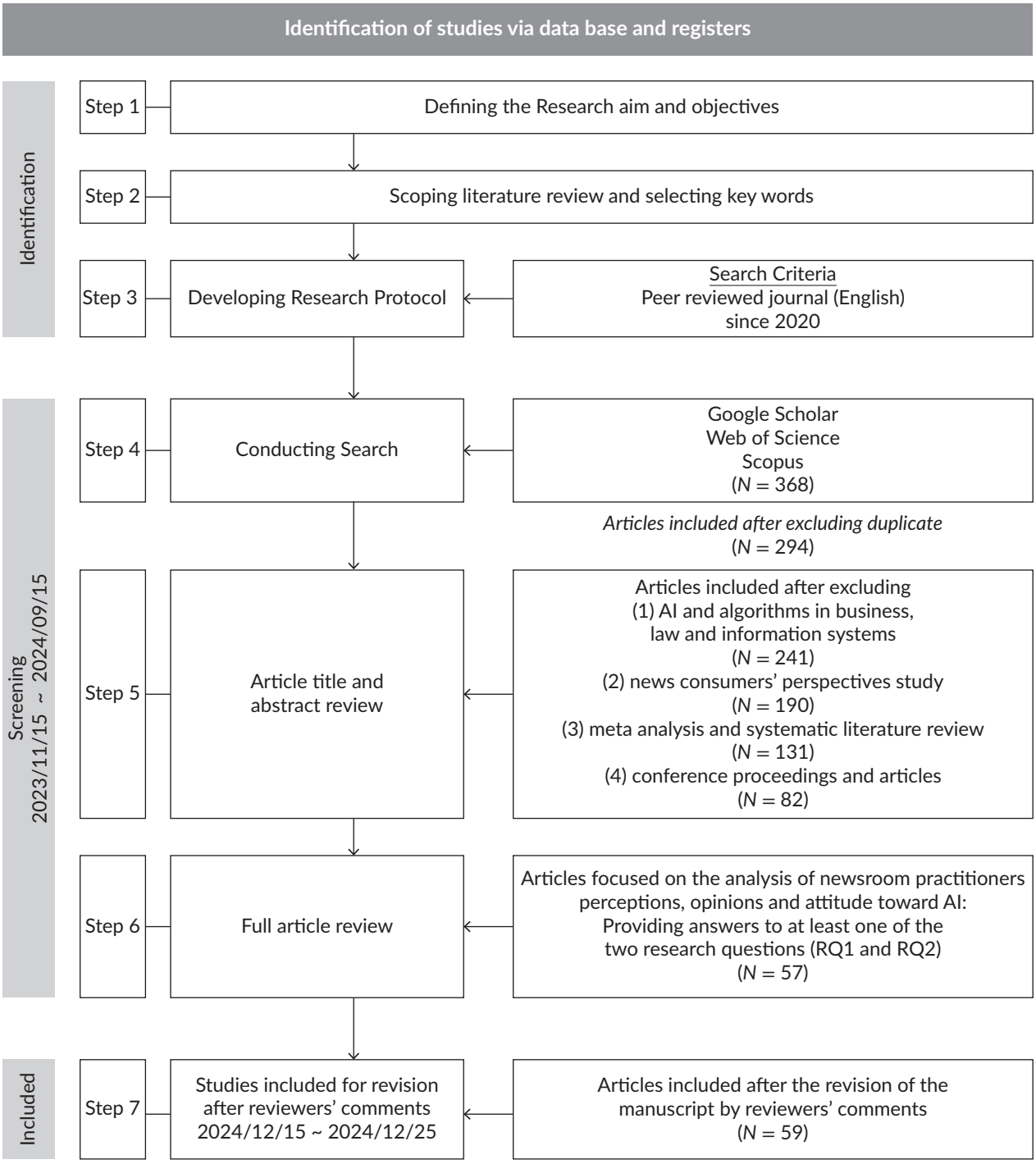


Figure 1. PRISMA flow chart. Source: Moher et al. (2010).

Below is the basic information about the selected papers, including their target countries, publication venues, and years of publication.

In terms of the countries analyzed in the selected papers, the geographical area was broad, encompassing 41 nations. This widespread geographical distribution provides a solid foundation for identifying general patterns in the reviewed studies, demonstrating a global reach that transcends differences in technological infrastructure. While the studies were primarily concentrated in technologically advanced nations such as the US ($N = 6$), Germany ($N = 3$), the UK ($N = 3$), Norway ($N = 3$), and other Western European countries, some also focused on less technologically advanced nations like Pakistan ($N = 5$), Nigeria ($N = 4$), Jordan ($N = 2$), and the UAE ($N = 2$). This division between the two groups can be advantageous for comparisons to identify disparities stemming from the heterogeneous technological infrastructures. Notably, some studies employed a comparative approach, particularly those from Europe and the US, analyzing a range of nations with special emphasis on countries such as China, the UK, Germany, and the US. These multi-country studies ($N = 8$) typically highlighted general features of news organizations rather than focusing on specific constraints related to regional or national contexts.

The reviewed papers were published in various journals: *Digital Journalism* ($N = 15$), *Journalism Practice* ($N = 11$), *Journalism Studies* ($N = 4$), *Journalism* ($N = 3$), *Journalism and Media* ($N = 3$), *Media and Communication* ($N = 2$), *New Media and Society* ($N = 2$), *Communication and Society* ($N = 2$), and *Studies in Media and Communications* ($N = 2$). Most of these journals focus on journalism, communication, and media studies. However, some articles, particularly those discussing newsrooms in less developed countries, appeared in interdisciplinary journals that cover broader fields such as the humanities, social sciences, and geography.

Regarding the timeline, publications were distributed over several years: 2020 ($N = 6$), 2021 ($N = 3$), 2022 ($N = 17$), 2023 ($N = 16$), and early 2024 ($N = 19$), reflecting data collection conducted mid-year. This trend reflects a growing interest in AI's impact on journalism, as evidenced by the substantial increase in publications from 2022 onward.

One of the concerns of this study was to investigate the impact of generative AI on journalism since its introduction in 2022. Notably, more than half of the 36 reviewed papers were published since 2023, which was anticipated due to the rapid growth of interest in generative AI technologies within journalism. However, among the 59 papers analyzed, only two since 2023 specifically addressed journalists' perspectives on ChatGPT, providing valuable early insights into its influence on newsroom dynamics. This trend indicates that a comprehensive investigation into the impact of generative AI on journalism remains beyond the scope of this study and is reserved for future research.

4. Results

Previous studies on the adoption of AI in journalism have typically been conducted at three different levels. Some investigations focused on a micro-level analysis, examining the individual responses of media practitioners to AI technology: their personal dispositions, temperaments, professionalism, technical proficiency, etc. (e.g., Ayyad et al., 2023; de Haan et al., 2022; Jones et al., 2022). These studies explore how journalists perceive the benefits and drawbacks of adopting AI, the influence of their technical expertise, and the ethical and normative challenges they encounter in the integration process.

Other studies focused on the meso-level analysis of mass media organizations, examining how news content was produced using AI technology (e.g., Allam & Hollifield, 2021; Bastian et al., 2021; Møller & Thylstrup, 2024). These studies primarily investigated ownership structure, organizational culture, technical training programs, and the distribution of collective financial resources in journalism organizations—factors that affected the integration of AI in journalism.

The third group of studies conducted a macro-level analysis, investigating the broader social context in which AI technology is developed and the national-level infrastructure for AI and journalism (e.g., Ahmad et al., 2023; Calvo-Rubio & Rojas-Torrijos, 2024; Gondwe, 2024; Munoriyarwa et al., 2021; Yu & Huang, 2021). They sought to identify favorable social conditions for integrating AI into the media landscape, including challenges related to inadequate internet access and limited data availability, the establishment of cultural norms, and national regulations for AI usage.

Despite varying levels of analysis from diverse perspectives, the reviewed studies suggested general findings on the relationship between AI and journalism, highlighting their universal significance. Some findings pertain to fundamental issues, such as the advantages and disadvantages of AI adoption, whereas others address more specific concerns. A recurring theme is the interplay between journalists' professional identity and the ethical considerations involved in AI integration. Reflecting on the evolving role of AI algorithms, these studies expressed concerns about the potential erosion of journalists' professional ideology and values due to AI adoption. The findings from existing studies are summarized below.

4.1. Findings for RQ1

The first research question addressed in this study is how journalists perceive the adoption of AI in journalism and why they maintain specific attitudes toward it. This question is crucial as it enables us to comprehend the factors that influence journalists' positive or negative perspectives on AI adoption, helping us identify what facilitates and hinders this adoption in the newsroom. Previous studies reveal that the discourse on AI in newsrooms is predominantly divided into optimism and pessimism. According to Cave et al. (2018), popular portrayals of AI in the English-speaking West often oscillate between excessive optimism about the technology's potential achievements and melodramatic pessimism. Additionally, it is also noteworthy that journalists exhibit an ambivalent attitude in various facets of journalism regarding the introduction of AI technology in news production.

4.1.1. Positive Perspectives

Many of the studies reviewed in this work revealed that journalists' positive attitudes toward the application of AI technology center around three main issues: enhancing the efficiency and effectiveness of news production and creating a new organizational culture.

4.1.1.1. Efficiency of AI: Save Time and Resources

Most journalists pointed out that the introduction of AI in news organizations is still in its initial stages. However, it is evident that they displayed a very positive attitude toward AI adoption, believing it would enhance their work's efficiency and productivity. AI systems that automatically generate news articles based

on data sets and templates save time and resources for news organizations. Journalists pointed to time savings and increased efficiency as major advantages of AI (Canavilhas, 2022; Cools & Diakopoulos, 2024; Noain-Sánchez, 2022; Tejedor & Vila, 2021).

Automated technologies are particularly advantageous for generating large volumes of articles on specific topics (Ahmad et al., 2023; Haim & Graefe, 2017). Tools that automate transcription, image and video tagging, and story creation in news production can reduce temporal and physical variable costs (Ahmad et al., 2023; Keefe et al., 2021). The benefits of autonomously produced content through algorithms became more apparent in time-sensitive newsroom environments (Ahmad et al., 2023; Wölker & Powell, 2021). Schapals and Porlezza (2020) propose that automated journalism provides valuable support to journalists in managing routine tasks, thus allowing them to focus on more intricate responsibilities that still require the unique expertise of human professionals.

The advantages of AI concerning efficiency are most evident in the context of generative AI. Some studies have specifically examined the impact of generative AI on journalism, viewing it as a means of showcasing state-of-the-art advancements in AI (Cools & Koliska, 2024; Jia et al., 2023; Spyridou & Danezis, 2024).

4.1.1.2. Effectiveness of AI: Automating Computation and Visualization

Journalists have indicated that adopting AI could enhance the effectiveness of their work by improving the quality of their products. AI plays a central role in automating computation-intensive processes, enabling journalists to access and extract critical information that was previously difficult to obtain (Beckett, 2019; de-Lima-Santos, 2022; Fridman et al., 2023). Data journalism utilizes AI technologies to analyze and visualize vast amounts of information. Visualization is vital for presenting complex information in a simple and comprehensible format (Fridman et al., 2023; Rodríguez et al., 2015). By leveraging these tools, journalists can pursue in-depth topics more effectively, contributing to the public discourse through investigative journalism. The adoption of generative AI, in particular, will dramatically improve the quality of news content. For instance, research shows that OpenAI's GPT software series, powered by deep learning, produces text of quality remarkably close to human writing (Floridi & Chiriatti, 2020; Moran & Shaikh, 2022).

4.1.1.3. New Organizational Culture: Fostering Collaborations Among Journalists

The organizational structure and culture of newsrooms significantly influence journalists' perceptions and adoption of AI systems. Organizational culture in media organizations is a critical determinant in executing journalistic innovation (Steensen, 2018; Zaragoza Fuster & García Avilés, 2022). Journalists working for large media groups that prioritize public service and are not under significant financial pressure tend to exhibit relatively positive and proactive attitudes toward AI adoption (Ahmad et al., 2023; Jones et al., 2022; Munoriyarwa et al., 2021; Zaragoza Fuster & García Avilés, 2022).

This aligns with previous research suggesting that technology adoption is influenced by political, social, economic, and cultural environments (Burr, 2015; Pinch & Bijker, 1984; Yu & Huang, 2021). For instance, the BBC in the UK and RTVE in Spain have established specific innovation departments, like media labs, to equip journalists with the knowledge and tools necessary for developing innovations in content production and distribution (Nunes & Mills, 2021). These initiatives cultivate a "collaborative space for innovators from

within and beyond companies to engage with one another or function as a loose network of communities of practice within a specific geographic cluster, brought together to solve a problem, experiment, or play" (Mills & Wagemans, 2021, p. 1469; see also Møller & Thylstrup, 2024; Svensson, 2021; Zaragoza Fuster & García Avilés, 2022).

4.1.2. Negative Perspectives

Journalists who worried about the adoption of AI mainly cited the enormous cost of implementing AI and its impact on the labor market (particularly regarding employment opportunities) in journalism.

4.1.2.1. Burden of Cost: Lack of Financial and Technological Resources

Despite the significant advantages of AI technology, financial resources and environmental assets are prerequisites for reaping the benefits of AI in newsroom organizations. The challenges in securing or supporting resources (funds and personnel for technology adoption, development, and maintenance) are barriers from the initial stages of establishing AI infrastructure (Guanah et al., 2020; Paik, 2023; Yu & Huang, 2021). The essential algorithmic tasks for journalistic organizations include storytelling, layout, headline optimization, and selecting story-related materials such as images and photos (Bold-Erdene, 2020; Munoriyarwa et al., 2021). Nevertheless, implementing the requisite technologies entails substantial costs (de-Lima-Santos, 2022; Litskevich, 2022; Noor & Zafar, 2023; Okocha & Ola-Akuma, 2022; Paik, 2023). While recognizing that AI can enhance the productivity and efficiency of work processes, media companies may find that the enormous cost of new technology diminishes their motivation for investment.

This contradiction—that AI can lower costs in news production and operations but does not attract organizational financial support—may be linked to a lack of knowledge about AI's potential (Canavilhas, 2022; de Haan et al., 2022; Jamil, 2021; Noain-Sánchez, 2022). Journalists frequently shared such concerns in less technologically developed countries like Nigeria, Pakistan, and South Africa, as well as among smaller or regional media organizations even in more developed nations. This stands in contrast with the active algorithmization of news production processes by well-funded media entities in Europe and the US, including *The Guardian*, *The New York Times*, and *Washington Post* (Cools & Koliska, 2024; Jamil, 2021; Jia et al., 2023; Milosavljević & Vobič, 2021; Munoriyarwa et al., 2021; Svensson, 2021; Zaragoza Fuster & García Avilés, 2022). Insufficient funding, a lack of technical expertise, and a rigid institutional environment pose significant obstacles to AI adoption within journalistic organizations (Boczkowski, 2005; de-Lima-Santos & Mesquita, 2021; Krumsvik et al., 2019; Lindblom et al., 2022; Paulussen, 2016).

4.1.2.2. Impact on Employment: Concerns About Job Security

Another skepticism and anxiety of journalists regarding AI have impeded the adoption and diffusion of the technology. Concerns about job security and social status manifest as vague fears about AI technology. While innovative technology and automation can threaten job security across various sectors, the field of journalism faces a unique challenge due to the prevailing journalistic logic in newsrooms. Journalism ideology is often interpreted as "how journalists give meaning to their news work" (Deuze, 2005, p. 444; see also Helberger et al., 2022) and frequently serves as a normative framework in media studies (Danzon-Chambaud & Cornia, 2021; Lindén, 2017b; Usher, 2017). However, the processes behind AI's data

and algorithm formation are technically complex and challenging to understand. This constitutes a perplexing situation for journalists who generally lack awareness of this new technology—one that could ultimately jeopardize their job security. Furthermore, enhancing news production productivity by applying advanced AI technology will significantly reduce job opportunities for journalists, resulting in large-scale layoffs. This structural shift in the labor market will compel journalists to develop negative attitudes toward the adoption of AI in their workplaces.

4.1.3. Ambivalent Attitudes: Integration of AI Technology With Journalistic Values

While some aspects of the advantages and disadvantages of adopting AI in journalism are somewhat expected, others remain uncertain, primarily due to the ambiguous attitudes of media practitioners. Especially important in this sense is that there is little consensus among news organizations about integrating AI technology with journalistic values. Although journalists generally advocate for the inclusion of journalistic values in AI-driven news content, they are divided on whether the current AI technology adequately respects these values. They also identified various challenges that hinder the incorporation of journalistic ethics and values into AI-assisted news stories.

AI can open new avenues for journalistic research and reporting, but such technologies are far from neutral (Ahmad et al., 2023; Bastian et al., 2021; Gondwe, 2023; Jamil, 2023; Moran & Shaikh, 2022; Munoriyarwa et al., 2021). Journalists displayed a relatively ambiguous attitude, expressing both hope and skepticism regarding the introduction of journalistic values. No one presented fixed opinions reflecting pure optimism or pessimism. Instead, they conditionally assessed whether AI would enhance or undermine journalistic values based on specific conditions (de Haan et al., 2022; Jones et al., 2022; Noain-Sánchez, 2022; Sholola et al., 2024; Soto-Sanfiel et al., 2022; Spyridou & Danezis, 2024). The debate on how AI will advance or hinder the normative vision that journalism upholds has spurred extensive scholarly discussions (Carlson, 2015; Gutierrez Lopez et al., 2023; Kothari & Cruikshank, 2021; Lewis et al., 2019; Stray, 2019).

The efforts of news organizations to integrate AI technology with journalistic values have become more pronounced, particularly during the Covid-19 pandemic, when the risk of misinformation increased (Kreft et al., 2023; Montaña-Niño & Burgess, 2024; Sharadga et al., 2022; Túñez-López et al., 2021). However, many algorithm-based tools are fundamentally not designed and developed with journalistic values and norms in mind (de Haan et al., 2022; Diakopoulos, 2019). Journalists suggest that more contextual information is necessary to enhance the quality of AI-generated news content. From a journalistic perspective, providing context that explains the reasons and methods behind news events, enabling readers and viewers to connect the dots, has become increasingly important (Ahmad et al., 2023; Zaid et al., 2022).

This issue of journalistic values and accountability of journalists continues to evolve and intensify with the prevalence of generative AI (Cools & Diakopoulos, 2024; Cools & Koliska, 2024; Paik, 2023). Some studies advocate for developing accountability models to update journalistic ethical standards in the generative AI era, while others delve into the practical risks and opportunities associated with generative AI technologies, stressing the importance of continuous monitoring and evaluation to ensure the ethical and responsible deployment of such tools. This indicates that despite advancements in generative AI, the technology has yet to be regulated and monitored by those at the core of the newsroom operations.

4.2. Findings for RQ2

The second research question pertains to the frameworks that journalism organizations apply to understand the adoption of AI technology and what strategies can be adopted for the development of alternative frameworks facilitating the collaboration of AI with journalistic values. This study found that three theoretical models are primarily applied to explain the relationship between AI technology and journalism: field theory, human–machine communication (HMC), and technology acceptance model (TAM). Field theory is particularly relevant for analyzing power dynamics within journalism, while HMC and TAM elucidate the interaction between humans and AI technology.

4.2.1. Field Theory: Broadening the Boundaries and Power Dynamics of Journalism

With the introduction of AI technology in journalism, the traditional boundaries of the journalism field have become blurred. Various kinds of AI professionals are widely collaborating with journalists in news content production. Developers, programmers, and designers are now regarded as representatives of the journalism profession in newsrooms (Lischka et al., 2022; Møller & Thylstrup, 2024; Olsen, 2023; Schjøtt Hansen & Hartley, 2021). The significance of new entrants to journalistic work in the form of data scientists is growing (Chew & Tandoc, 2022; Lischka et al., 2022; Møller & Thylstrup, 2024). These technology professionals consistently introduce new information technologies into organizations, embodying the avant-garde journalism community (Hepp & Loosen, 2019). Consequently, the perspective that IT experts and developers should be considered key actors in news organizations (Anderson, 2013; Diakopoulos, 2020; Moran & Shaikh, 2022) is gaining traction.

Diverse practitioners within the journalist group play a crucial role in maintaining news operations (Jamil, 2021; Lewis & Westlund, 2015), while traditional journalists are still regarded as the core agents that uphold journalism (Jamil, 2021; Ryfe, 2019). Thus, editorial technologists work "at the intersection of traditional journalist positions and technologically intensive positions that were once generally separate" (Kosterich, 2020, p. 2). The advancement of AI-based digital technologies has prompted the phenomenon of "the blending of journalist-technologists" (Hermida & Young, 2017, p. 171) and created an "intersectional techno-journalistic space" (Ananny & Crawford, 2015, p. 192). Data managers, analysts, algorithm developers, and other newsroom members fill this intersectional space and are now incorporated into the realm of "journalists."

The expanded boundaries of journalism and the influx of new members naturally incur a new power dynamic, often leading to conflicts between traditional journalists and newcomers. The introduction of new members, frequently referred to as insurgents who challenge the status quo, inevitably threatens the power of incumbents striving to maintain the field as it currently exists. It is widely recognized that Pierre Bourdieu's field theory provides the best framework for analyzing the struggles between new and established journalists to accumulate, exchange, and monopolize various power resources. This explains why many studies have relied on this theory to elucidate journalism's complex power dynamics and hierarchies among journalists (Bourdieu & Wacquant, 1992; Lindblom et al., 2022; Lischka et al., 2022; Møller & Thylstrup, 2024; Perreault & Ferrucci, 2020). Although field theory was not originally intended to explore technology-driven organizational changes, research inspired by Bourdieu has rapidly increased, analyzing how digital technology is altering the journalism field (Craft et al., 2016; Hovden, 2008; Lindblom et al., 2022; Schultz, 2007; Vos et al., 2012).

4.2.2. The HMC Model

The human–computer interaction framework, as articulated by communication scholars, defines "humans as communicators" and "machines as intermediaries or facilitators" (Jamil, 2021, p. 1402; see also Barnlund, 2008; Weiswasser, 1997). Jamil (2021) further elaborated, "[The] human-machine communication framework, which is an emerging area of communication research…posits technologies and machines as communicators" (Jamil, 2021, p. 1403). According to Guzman (2018, p. 1), the HMC concept is concretized into three areas: human–computer interaction, human–robot interaction, and human–agent interaction. Within the context of human–computer interaction, it is possible to design systems that verify news sources (including instances where news content is revised and republished over time), measure media bias, and more (Cruz et al., 2020; Diakopoulos, 2020; Evans et al., 2020; Gutierrez Lopez et al., 2023; Jamil, 2021; Jones et al., 2022; Komatsu et al., 2020). Understanding the algorithms that represent AI development, as well as the interactions between machines and human journalists, can enhance the journalistic values of trust, objectivity, and transparency.

4.2.3. TAM

TAM is one of the most influential and widely used theories for analyzing the factors that determine the adoption of new technologies by individuals or groups (Davis, 1989; Venkatesh et al., 2003). The core components of this model are "perceived ease of use" and "perceived usefulness." "Perceived ease of use" refers to the degree to which a person believes that using a particular technology would be free of effort. "Perceived usefulness" denotes the degree to which a person believes that using a specific technology would enhance their job performance. When individuals perceive a technology as both easy to use and useful, they are more likely to develop a positive attitude toward its adoption.

TAM has provided a theoretical foundation for connecting technology and journalism across various cultural contexts (Ayyad et al., 2023; Goni & Tabassum, 2020; Shah et al., 2024; Soto-Sanfiel et al., 2022; Zhou, 2008). With the evolution of new technologies, TAM has been expanded to include various additional variables, resulting in more nuanced models. TAM2 (Venkatesh & Davis, 2000) introduced additional determinants of technology adoption, such as job relevance and social influence factors, bridging the gap between technology adoption and journalism research (Ayyad et al., 2023). TAM later evolved into TAM3 (Venkatesh & Bala, 2008), which detailed variables like computer self-efficacy and experience, and UTAUT (Unified Theory of Acceptance and Use of Technology), which incorporated factors like price value and habit (Venkatesh et al., 2003), and UTAUT2 (Venkatesh et al., 2012). While TAM's broad applicability is advantageous, it has been criticized for providing only general information about users' opinions on systems (Ayyad et al., 2023; Mathieson, 1991). This critique is particularly pertinent when navigating the complex equation of merging technology with journalistic ethics.

5. Considerations for the Establishment of an Alternative Framework

Drawing from an extensive literature review and the general findings from existing studies, this study now intends to suggest some helpful ideas for constructing an alternative framework that aligns AI technology with journalistic values. The fundamental issue here is how human journalists' editorial judgment and ethical values can be incorporated into AI-generated content (Bell et al., 2017; Jamil, 2021; Ward, 2018). In an era

marked by the rise of automated journalism, exemplified by AI, traditional journalists have become increasingly compelled to staunchly defend their work—or what many have referred to as their "craft" (Schapals & Porlezza, 2020, p. 23; see also Hanitzsch & Vos, 2017). Journalists are concerned about whether essential ethics and values can be technically implemented.

Following Ward's (2018) definition of journalistic values as principles and norms guiding public journalism, they can be classified into organization-centered and audience-centered (Bastian et al., 2021). Organization-centered values include objectivity, diversity, transparency, accountability, and other related values that constitute "good" journalism. In contrast, audience-centered values encompass privacy, data protection, user agency, autonomy, and other values that focus on the relationship between the media and the audience. One study indicated that journalists generally prioritize "core" values, such as transparency, diversity, editorial autonomy, personal relevance, and usability, over less essential ones like objectivity, neutrality, and enjoyment of usage (Bastian et al., 2021 p. 855).

Journalists have still not reached a consensus on how to implement their professional values in AI-based news production. Additionally, academic research on this issue is relatively scant compared with an enormous accumulation of studies focused on technical matters and journalists' attitudes toward them. In this context, it is imperative to explore the possibility of an alternative framework to enhance both the efficiency and effectiveness of news production without losing the professional ethic and values of journalism. It is beyond the reach of this study to develop a fully established theoretical framework. Instead, it proposes three essential strategies or approaches for implementing this alternative framework: the AI technologists' side, the journalists' side, and the collaboration between the two.

5.1. Infusing Journalistic Values Into AI Technologists and Data Scientists

The first strategy for implementing journalistic values is to demand that technologists and data scientists learn and incorporate them into their technical work. It is generally believed that these technicians are only interested in collecting and providing data from a purely technical perspective. However, the decisions of which data to collect and how to refine and process them are never free from biases, which critically threaten the objectivity and transparency of news content (Haim, 2022; Jamil, 2021; Lindblom et al., 2022; Noor & Zafar, 2023). Data specialists confess that "datasets are never neutral sources, and almost all of them are biased in some way" since "AI technology is prone to inherit human biases" (Noain-Sánchez, 2022, p. 113).

In this situation, the newsroom relies on the journalist's active involvement in news production to check the fulfillment of journalistic values in the news content. Journalists (not technology) are still accountable for applying these values to their stories. Eventually, it is up to journalists to decide how to incorporate technology into their narratives. Some journalists confess that they trust their own "gut and skills" in determining what stories should be published, without enshrining editorial judgments such as impartiality in the tool itself (Gutierrez Lopez et al., 2023, p. 494).

The monopoly of editorial judgments by journalists may sometimes incur conflict between two groups of specialists participating in news production: journalists and technicians. The contrast between the "hard" practices of data science and the "soft" considerations of journalists generates "science frictions." This friction, however, involves productive tension that reshapes the awareness of data scientists and AI

engineers, prompting them to consider ethics more seriously than some of their previous places of employment (Møller & Thylstrup, 2024, p. 9). They began to acknowledge their social responsibility regarding technology implementation; as one noted, "It is our responsibility to go forth and to make sure that we are presenting the clearest picture possible and that we are presenting it fairly" (Lischka et al., 2022, p. 11).

Renewed awareness of technicians' responsibilities creates an excellent social environment to infuse journalistic values into these technology practitioners. By developing some strategies for blending technological capabilities with editorial requirements, it is now possible to convert them to ethical data scientists and AI programmers. The future of AI journalism hinges on merging algorithms with editorial and ethical parameters (Jamil, 2023; Møller & Thylstrup, 2024; Noain-Sánchez, 2022; Perreault & Ferrucci, 2020; Rydenfelt, 2021; Spyridou & Danezis, 2024). Just like the journalists' gut feeling has safeguarded journalistic values so far, the algorithmic gut feeling, based on the normative orientation of data technologists, will enforce ethical values in AI-driven news stories in the future.

Simultaneously, technologists need to clarify how they process data from a technical perspective, especially how AI assists in making specific recommendations for the journalists (Cools & Diakopoulos, 2024; Cools & Koliska, 2024; Olsen, 2023; van Drunen & Fechner, 2022). This explanation will enhance the transparency of the technical work process, which is a prerequisite to realizing the value of trust and encouraging users to embrace technical innovations. Achieving transparency "can strengthen the legitimacy of the chance to use such a system" (Bastian et al., 2021, p. 849).

5.2. Enhancing Journalists' Adaptability to AI Technology

The second strategy for integrating journalistic values into AI technology is to educate journalists on the technical details of AI. The accumulation of enormous amounts of data used for news production overwhelms journalists who struggle to understand how to process this information. The monopoly of databases by large engineering corporations and the secret management of AI algorithms exacerbate journalists' helplessness in producing AI-based news content. The situation is further complicated, as even their developers of AI systems may find it difficult to clarify how they function: This is known as the black-box problem (Castelvecchi, 2016). The technical complexities of incorporating algorithmic logic into news production lead journalists to adopt a notably passive attitude toward its adoption (Ayyad et al., 2023; Canavilhas, 2022; de Haan et al., 2022; Noain-Sánchez, 2022). This limited understanding of AI specifics within the industry poses a significant threat to journalists' overall performance (Kreft et al., 2023).

Addressing this challenge begins with raising awareness, empowering journalists to actively pursue knowledge and better understand the workings of AI (Basak et al., 2024; de Haan et al., 2022; Heravi & Lorenz, 2020; Olsen, 2023; Trang et al., 2024). Journalists acknowledge that their role in the journalism field requires ongoing advancement in technology, writing, and ethical standards. It is evident that journalism, in the rapidly evolving AI-driven media landscape, now requires a mindset grounded in technological innovation (Ahmad et al., 2023; Lindblom et al., 2022; Montaña-Niño & Burgess, 2024; Olsen, 2023). They should be bold enough to embrace a digital mindset and undergo training on technical tools to control and supervise AI processes (Noain-Sánchez, 2022).

Some studies express optimism by predicting that organizations and individuals can become digitally savvy if they are equipped with only about 30% fluency in technical topics, which offers the minimum digital literacy needed to be digital (Leonardi & Neeley, 2022). Realistically speaking, however, it is never easy for journalists to obtain digital literacy skills without undergoing rigorous training and extensive learning processes. This raises an urgent call for journalistic curricula to meet these requirements to help establish norms and quality standards for data collection, processing, analysis, and visualization for journalism (Ahmad et al., 2023; Cools & Diakopoulos, 2024; Fieiras-Ceide et al., 2022; Haim, 2022). In this sense, the educational model for journalism must be updated to accommodate AI.

Although there is no consensus on how to teach journalists to acquire the technical abilities needed to perform specific tasks in AI engineering, journalists generally agree on the necessity of learning those skills. One journalist testified to this necessity by stating, "We cannot leave this technology managed only by techs. Education and training are essential, and we may focus on the editorial role of algorithms and how decisions made by algorithms can have serious social implications" (Noain-Sánchez, 2022, p. 115). Some even went further to insist that "it is essential that journalists learn how to programme in order to understand what is behind algorithms and the criteria they follow" (Noain-Sánchez, 2022, p. 115).

With enhanced digital literacy capabilities, journalists can audit the products of AI technical experts to determine whether they observe journalistic ethics and values. They can also supervise the work process of AI specialists to check whether ethical codes are implemented and ethical principles are embedded by design. One perfect example is those journalists who have switched careers from journalistic posts to technical ones, acting as conduits between the needs of journalists and the technical teams (Gutierrez Lopez et al., 2023). This career shift not only provided them with the capabilities necessary for algorithm audits but also helped other journalists understand the complexities of AI technology. One of them defined her role as translating what the tool does in an easy-to-understand and meaningful way, so that it is evident to journalists why this helps them. This aptly testifies to the effectiveness of technical training for journalists who will merge their professional values with AI-driven works in news production.

5.3. Facilitating Collaborations Between Journalists and AI Technicians

The final approach is to develop a new collaborative framework for journalists and AI technologists to work together while maintaining their traditional field boundaries. This is a somewhat realistic solution, given the diverse impasses both journalists and technicians encounter when attempting to cross disciplinary boundaries by learning each other's expertise. In fact, journalists and technicians strongly agree that collaboration is necessary with their partners. In this regard, the formation of multidisciplinary teams comprising diverse fields of expertise and the organizational dynamism that encourages both internal and external collaboration emerge as crucial elements within the culture of modern newsrooms (Bailer et al., 2022; de-Lima-Santos, 2022; Grimme & Zabel, 2024).

Collaboration can take various forms, from casual meetings to establishing cooperative organizations. A casual meeting can evolve into a more serious one as the need for collaboration is widely shared among participants. This progression could serve as a first step toward data transparency and a culture of open journalism (Cook, 2021; Dierickx et al., 2023), where increasingly intelligent entities, knowledge sharing, and collaborative thinking may become integral components of a newsroom (Grimme & Zabel, 2024). One

typical collaboration model suggested by Gutierrez Lopez et al. (2023) will help us understand the workflow of collaboration; in this model, participants "undertook several in-house rounds of ideation across their interdisciplinary team to look out for themes in the groups of codes" (Gutierrez Lopez et al., 2023, p. 491). When this occurs, collaboration will advance to a new level, transcending the traditional disciplinary boundaries of each participant.

The collaboration participants generally agree that human involvement is essential. They believe that "automated recommendation technologies are never in place 'instead of' but always in combination with humans who decide what is important news and what is not" (Bastian et al., 2021, p. 846). This signifies the prioritization of human involvement over technical processes in news production; editorial decision-making, with particular attention to journalistic values, should take place before designing algorithms sensitive to these values.

The formalization of journalists' involvement in AI-based news production can create new professional roles, such as intermediaries with technical and editorial expertise who facilitate collaboration between editorial and technical departments. It can also establish new procedures that outline the roles of editors and journalists in algorithmic design (Cools & Koliska, 2024; Jia et al., 2023; Lindblom et al., 2022; van Drunen & Fechner, 2022). *Washington Post* developed a highly effective method to achieve this goal by employing liaisons connecting the two groups of specialists. The *Post* hired two liaisons who served as intermediaries between the newsroom staff and the more technically oriented engineering team. These liaisons possessed knowledge and skills in both journalism and technology, enabling them to translate the newsroom's goals and values into actionable technical requirements for the engineering teams, and vice versa (Cools & Koliska, 2024, p. 675).

One significant obstacle that both journalists and technicians should overcome to facilitate collaboration is the cultural difference between these two groups of professionals. They each developed their expertise in entirely different cultural and organizational contexts, which hinders their mutual understanding. Additionally, technicians and journalists may face conflicts over essential decision-making in the news production process, as the former infiltrates a new territory and challenges the power and authority of journalists who have traditionally dominated the field. Editorial technologists struggle to gain appropriate recognition and sufficient symbolic capital (Lischka et al., 2022). Therefore, it is essential to foster an amicable relationship between the two groups and acknowledge the equal status of all participants in the collaboration, regardless of their career backgrounds. By achieving this collaborative spirit, they can build a solid foundation of mutual understanding and recognition, leading to shared ownership of the products they create.

6. Conclusion

Journalists who participated in the interviews or surveys in the 59 reviewed papers worked in 41 countries, including both advanced and less technologically developed ones. Despite differences in their perceptions of AI based on newsroom organizations and regional contexts, they generally predicted that AI adoption in newsrooms was inevitable. The findings of the reviewed papers indicate that they commonly pointed out the benefits of AI adoption in terms of enhancing the efficiency and effectiveness of their work processes and promoting a collaborative organizational culture. On the other hand, they expressed concerns about the financial burdens and job insecurity that AI adoption could incur. They also recognize that the potential of

generative AI technologies is so vast that they need to learn its internal working to improve the quality of their work.

Empirical studies connecting AI and journalism most frequently reference Bourdieu's field theory, HMC, and TAM. After a thorough review of previous research, this study proposed three strategies for developing an alternative framework that integrates AI into journalism. The key components of this alternative framework include how to assist journalists in adapting to new AI technologies, how to encourage technicians to uphold journalistic values, and how to foster a collaborative relationship between the two professions. These factors pertain to the issue of power dynamics within journalism, which is experiencing a radical transformation due to the rise of AI and its integration into the field, a change that will be further accelerated by the groundbreaking advancements in generative AI.

This study primarily focused on the findings of the general pattern in existing research. These findings and suggestions could establish a good foundation for comparative studies and case studies in the future. The recommendations of this study may serve as a yardstick for evaluating how closely each specific case aligns with or diverges from the general trend. Additionally, it makes significant contributions by identifying universal challenges that journalism faces in the era of AI, such as algorithmic bias, ethical dilemmas, and the global exchange of innovative practices. It also provides a framework for understanding how AI reshapes journalistic values like accuracy, trust, objectivity, and accountability. By doing so, it not only advances scholarly discourse but also equips newsrooms and policymakers with practical strategies for ethical and effective AI integration, ultimately strengthening journalism across various contexts.

This study explores ways to harmonize the adoption of AI with journalistic values and concludes by proposing the concept of a "journalistic algorithm." The term "journalistic" reflects traditional professional values like reliability, fairness, and truth-seeking. Combined with "algorithm," often associated with the enigmatic nature of AI technology, it underscores the need for transparency and ethical integrity in AI systems. Given this context, this study argues that AI algorithms used in journalism must adhere to these ethical and normative principles, making the integration of journalistic values into AI a central objective in the evolving relationship between technology and journalism.

Furthermore, newsrooms can advance beyond simply adopting or using the technology presented to them by actively developing it and engaging in discussions to make algorithms more "explainable." Thus, technologists—a new type of journalist—must open the black box and publicize the algorithms they utilize to produce news content, enhancing the transparency of AI technology in journalism. Conversely, professional journalists who strive to uphold the traditional ethos of their field must learn about AI algorithms to verify that journalistic values are maintained and respected in AI-generated news products. In doing so, they can highlight the unique professional reasoning of human journalists by participating in the "creation of meaning" process.

7. Limitations

Some limitations should be acknowledged when interpreting the results of this study. First, while it attempts to identify general patterns from previous research by emphasizing the universal aspects of AI adoption, it overlooks regional and national variations, only occasionally referencing the differences between technologically advanced countries and those less developed. This limitation is somewhat unavoidable, as

the two objectives—identifying general features and examining regional diversities—cannot be accomplished in a single article. It would be very helpful for future studies to focus on this aspect of regional and national variations.

Second, since this study only included English-language papers relevant to the research questions, the range of reviewed documents may be somewhat limited. This selection bias is driven by practical considerations such as accessibility, global knowledge dissemination, and the role of English as the lingua franca of academia. However, this focus undoubtedly restricts the scope of this research by excluding studies in other languages, particularly those that provide localized or culturally nuanced perspectives on journalism, media, and AI.

Third, literature reviews inherently have certain limitations, which empirical studies should complement. Combining search strings that are optimized for the research topic in order to find highly relevant papers is difficult. In addition, databases are continuously updated, so the final sample may vary slightly depending on the timing, even with the same strings. The concepts and models proposed and highlighted in this study should be substantiated through future empirical research. To evaluate the applicability of these models in journalistic environments, it is highly recommended that both qualitative and quantitative methods be applied simultaneously.

Conflict of Interests

The authors declare no conflict of interests. In this article, editorial decisions were undertaken by Jeong-Nam Kim (University of Oklahoma).

References

Ahmad, N., Haque, S., & Ibahrine, M. (2023). The news ecosystem in the age of AI: Evidence from the UAE. *Journal of Broadcasting & Electronic Media*, *67*(3), 323–352. https://doi.org/10.1080/08838151.2023.2173197

Allam, R., & Hollifield, A. (2021). Factors influencing the use of journalism analytics as a management tool in Egyptian news organizations. *Journalism Practice*, *17*(3), 601–623. https://doi.org/10.1080/17512786.2021.1927803

Ananny, M., & Crawford, K. (2015). A liminal press: Situating news app designers within a field of networked news production. *Digital Journalism*, *3*(2), 192–208. https://doi.org/10.1080/21670811.2014.922322

Anderson, C. W. (2013). Towards a sociology of computational and algorithmic journalism. *New Media & Society*, *15*(7), 1005–1021. https://doi.org/10.1177/1461444812465137

Ashforth, B. E., & Schinoff, B. S. (2016). Identity under construction: How individuals come to define themselves in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 111–137. https://doi.org/10.1146/annurev-orgpsych-041015-062322

Ayyad, K., Ben Moussa, M., & Zaid, B. (2023). Journalists' perception of the adoption of new communication technologies in the UAE's media organizations. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2023.2300277

Bailer, W., Thallinger, G., Krawarik, V., Schell, K., & Ertelthalner, V. (2022). AI for the media industry: Application potential and automation levels. In B. Þór Jónsson, C. Gurrin, M.-T. Tran, D.-T. Dang-Nguyen, A. M.-C. Hu, B. H. T. Thanh, & B. Huet (Eds.), *MultiMedia Modeling: 28th International Conference, MMM 2022* (pp. 109–118). Springer. https://doi.org/10.1007/978-3-030-98358-1_9

Barnlund, D. C. (2008). A transactional model of communication. In C. D. Mortensen (Ed.), *Communication theory* (2nd ed., pp. 47–57). Transaction.

Basak, S., Tabassum, M., Goni, M. A., & Kundu, P. (2024). Artificial intelligence (AI) and future newsrooms: A study on journalists of Bangladesh. *Pacific Journalism Review*: *Te Koakoa*, 30(1/2), 96–110. https://doi.org/10.24135/pjr.v30i1and2.1235

Bastian, M., Helberger, N., & Makhortykh, M. (2021). Safeguarding the journalistic DNA: Attitudes towards the role of professional values in algorithmic news recommender designs. *Digital Journalism*, *9*(6), 835–863. https://doi.org/10.1080/21670811.2021.1912622

Beckett, P. (2019). *Ownership, financial accountability and the law: Transparency strategies and counter-initiatives*. Routledge.

Bell, E., Owen, T., Brown, P., Hauka, C., & Rashidian, N. (2017). *The platform press: How Silicon Valley reengineered journalism*. Tow Center for Digital Journalism, Columbia University. https://doi.org/10.7916/D8R216ZZ

Boczkowski, P. J. (2005). *Digitizing the news: Innovation in online newspapers*. MIT Press.

Bold-Erdene, J. (2020). *Application of algorithms in newsrooms* [Unpublished master's thesis]. University of Missouri. https://hdl.handle.net/10355/79884

Bourdieu, P., & Wacquant, L. (1992). *An invitation to reflexive sociology*. University of Chicago Press.

Burgess, J., Albury, K., McCosker, A., & Wilken, R. (2022). *Everyday data cultures*. Wiley.

Burr, V. (2015). *Social constructionism* (3rd ed.). Routledge.

Calvo-Rubio, L.-M., & Rojas-Torrijos, J.-L. (2024). Criteria for journalistic quality in the use of artificial intelligence. *Communication & Society*, *37*(2), 247–259. https://doi.org/10.15581/003.37.2.247-259

Calvo-Rubio, L.-M., & Ufarte-Ruiz, M.-J. (2021). Artificial intelligence and journalism: Systematic review of scientific production in Web of Science and Scopus (2008–2019). *Communication & Society*, *34*(2), 159–176. https://doi.org/10.15581/003.34.2.159-176

Canavilhas, J. (2022). Artificial intelligence and journalism: Current situation and expectations in the Portuguese sports media. *Journalism and Media*, *3*(3), 510–520. https://doi.org/10.3390/journalmedia3030035

Carlson, M. (2015). The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism*, *3*(3), 416–431. https://www.doi.org/10.1080/21670811.2014.976412

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., & Taylor, L. (2018). *Portrayals and perceptions of AI and why they matter.* The Royal Society.

Chan-Olmsted, S. M. (2019). A review of artificial intelligence adoptions in the media industry. *International Journal on Media Management*, *21*(3/4), 193–215. https://doi.org/10.1080/14241277.2019.1695619

Chew, M., & Tandoc, E. C., Jr. (2022). Media startups are behaving more like tech startups—Iterative, multi-skilled, and journalists that "hustle." *Digital Journalism*, *12*(2), 191–211. https://doi.org/10.1080/21670811.2022.2040374

Cook, C. E. (2021). Assessing conditions for inter-firm collaboration as a revenue strategy for politically pressured news media. *Journal of Media Business Studies*, *20*(1), 52–71. https://doi.org/10.1080/16522354.2021.2002106

Cools, H., & Diakopoulos, N. (2024). Uses of generative AI in the newsroom: Mapping journalists' perceptions of perils and possibilities. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2024.2394558

Cools, H., & Koliska, M. (2024). News automation and algorithmic transparency in the newsroom: The case of the *Washington Post*. *Journalism Studies*, *25*(6), 662–680. https://doi.org/10.1080/1461670x.2024.2326636

Craft, S., Vos, T. P., & Wolfgang, D. J. (2016). Reader comments as press criticism: Implications for the journalistic field. *Journalism*, *17*(6), 677–693. https://doi.org/10.1177/1464884915579332

Cruz, A. F., Rocha, G., & Cardoso, H. L. (2020). On document representations for detection of biased news articles. In *SAC '20: Proceedings of the 35th Annual ACM Symposium on Applied Computing* (pp. 892–899). Association for Computing Machinery. https://doi.org/10.1145/3341105.3374025

Danzon-Chambaud, S. (2021, August 6). Covering COVID-19 with automated news. *Columbia Journalism Review*. https://www.cjr.org/tow_center_reports/covering-covid-automated-news.php

Danzon-Chambaud, S., & Cornia, A. (2021). Changing or reinforcing the "rules of the game": A field theory perspective on the impacts of automated journalism on media practitioners. *Journalism Practice*, *17*(2), 174–188. https://doi.org/10.1080/17512786.2021.1919179

Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

de Haan, Y., van den Berg, E., Goutier, N., Kruikemeier, S., & Lecheler, S. (2022). Invisible friend or foe? How journalists use and perceive algorithmic-driven tools in their research process. *Digital Journalism*, *10*(10), 1775–1793. https://doi.org/10.1080/21670811.2022.2027798

de-Lima-Santos, M.-F. (2022). ProPublica's data journalism: How multidisciplinary teams and hybrid profiles create impactful data stories. *Media and Communication*, *10*(1), 5–15. https://doi.org/10.17645/mac.v10i1.4433

de-Lima-Santos, M.-F., & Ceron, W. (2021). Artificial intelligence in news media: Current perceptions and future outlook. *Journalism and Media*, *3*(1), 13–26. https://doi.org/10.3390/journalmedia3010002

de-Lima-Santos, M.-F., & Mesquita, L. (2021). In a search for sustainability: Digitalization and its influence on business models in Latin America. In R. Salaverría & M.-F. de-Lima-Santos (Eds.), *Journalism, data, and technology in Latin America* (pp. 55–96). https://doi.org/10.1007/978-3-030-65860-1_3

Deuze, M. (2005). What is journalism? Professional identity and ideology of journalists reconsidered. *Journalism*, *6*(4), 442–464. https://doi.org/10.1177/1464884905056815

Diakopoulos, N. (2019). Towards a design orientation on algorithms and automation in news production. *Digital Journalism*, *7*(8), 1180–1184. https://doi.org/10.1080/21670811.2019.1682938

Diakopoulos, N. (2020). Computational news discovery: Towards design considerations for editorial orientation algorithms in journalism. *Digital Journalism*, *8*(7), 945–967. https://doi.org/10.1080/21670811.2020.1736946

Dierickx, L., Lindén, C.-G. C, & Opdahl, A. L. (2023). The ethical dimensions of data quality for automated fact-checking. In B. Heravi (Ed.), *The Joint Computation + Journalism European Data & Computational Journalism Conference* 2023 (pp. 22–24). C+J Symposium; DataJConf. https://difusion.ulb.ac.be/vufind/Record/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/366095/Details

Evans, N., Edge, D., Larson, J., & White, C. (2020). News provenance: Revealing new text reuse at web-scale in an augmented news search experience. In *CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3334480.3375225

Fieiras-Ceide, C., Vaz-Álvarez, M., & Túñez-López, M. (2022). Artificial intelligence strategies in European public broadcasters: Uses, forecasts and future challenges. *Profesional de la Información*, *31*(5), Article e310518. https://doi.org/10.3145/epi.2022.sep.18

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Fridman, M., Krøvel, R., & Palumbo, F. (2023). How (not to) run an AI project in investigative journalism. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2023.2253797

Gelgel, N. M. (2020). Will technology take over journalism? *Informasi*, *50*(2), 5–10. https://doi.org/10.21831/informasi.v50i2.36379

Gondwe, G. (2023). ChatGPT and the Global South: How are journalists in sub-Saharan Africa engaging with generative AI? *Online Media and Global Communication*, *2*(2), 228–249. https://doi.org/10.1515/omgc-2023-0023

Gondwe, G. (2024). Artificial Intelligence, journalism, and the Ubuntu robot in sub-Saharan Africa: Towards a normative framework. *Digital Journalism*. Advance online publication. https://doi.org/10.1080/21670811.2024.2311258

Goni, A., & Tabassum, M. (2020). Artificial intelligence (AI) in journalism: Is Bangladesh ready for it? A study on journalism students in Bangladesh. *Athens Journal of Mass Media and Communications*, *6*(4), 209–228. https://doi.org/10.30958/ajmmc.6-4-1

Grimme, M., & Zabel, C. (2024). AI in the newsroom: A collective case study about newsworker–AI collaboration in the German newspaper industry. *Journal of Media Business Studies*. Advance online publication. https://doi.org/10.1080/16522354.2024.2380120

Guanah, J. S., Agbanu, V. N., & Obi, I. (2020). Artificial intelligence and journalism practice in Nigeria: Perception of journalists in Benin City, Edo State. *International Review of Humanities Studies*, *5*(2), Article 16. https://scholarhub.ui.ac.id/irhs/vol5/iss2/16

Gutierrez Lopez, M., Porlezza, C., Cooper, G., Makri, S., MacFarlane, A., & Missaoui, S. (2023). A question of design: Strategies for embedding AI-driven tools into journalistic work routines. *Digital Journalism*, *11*(3), 484–503. https://doi.org/10.1080/21670811.2022.2043759

Guzman, A. L. (2018). What is human–machine communication, anyway? In A. L. Guzman (Ed.), *Human–machine communication: Rethinking communication, technology, and ourselves* (pp. 1–28). Peter Lang.

Habermas, J. (1984). *The theory of communicative action: Reason and the rationalization of society* (Vol. 1). Beacon Press.

Haim, M. (2022). The German data journalist in 2021. *Journalism Practice*, *18*(6), 1378–1397. https://doi.org/10.1080/17512786.2022.2098523

Haim, M., & Graefe, A. (2017). Automated news: Better than expected? *Digital Journalism*, *5*(8), 1044–1059. https://doi.org/10.1080/21670811.2017.1345643

Hanitzsch, T., & Vos, T. P. (2017). Journalistic roles and the struggle over institutional identity: The discursive constitution of journalism. *Communication Theory*, *27*(2), 115–135. https://doi.org/10.1111/comt.12112

Helberger, N., van Drunen, M., Moeller, J., Vrijenhoek, S., & Eskens, S. (2022). Towards a normative perspective on journalistic AI: Embracing the messy reality of normative ideals. *Digital Journalism*, *10*(10), 1605—1626. https://doi.org/10.1080/21670811.2022.2152195

Hepp, A., & Loosen, W. (2019). Pioneer journalism: Conceptualizing the role of pioneer journalists and pioneer communities in the organizational re-figuration of journalism. *Journalism*, *22*(3), 577–595. https://doi.org/10.1177/1464884919829277

Heravi, B. R., & Lorenz, M. (2020). Data journalism practices globally: Skills, education, opportunities, and values. *Journalism and Media*, *1*(1), 26–40. https://doi.org/10.3390/journalmedia1010003

Hermida, A., & Young, M. L. (2017). Finding the data unicorn: A hierarchy of hybridity in data and computational journalism. *Digital Journalism*, *5*(2), 159–176. http://doi.org/10.1080/21670811.2016.1162663

Hovden, J. F. (2008). *Profane and sacred: A study of the Norwegian journalistic field* [Unpublished doctoral dissertation]. University of Bergen. https://bora.uib.no/bora-xmlui/handle/1956/2724

Jamil, S. (2021). Artificial intelligence and journalistic practice: The crossroads of obstacles and opportunities

for the Pakistani journalists. *Journalism Practice*, *15*(10), 1400–1422. https://doi.org/10.1080/17512786.2020.1788412

Jamil, S. (2023). Automated journalism and the freedom of media: Understanding legal and ethical implications in competitive authoritarian regime. *Journalism Practice*, *17*(6), 1115–1138. https://doi.org/10.1080/17512786.2021.1981148

Jia, C., Riedl, M. J., & Woolley, S. (2023). Promises and perils of automated journalism: Algorithms, experimentation, and "teachers of machines" in China and the United States. *Journalism Studies*, *25*(1), 38–57. https://doi.org/10.1080/1461670x.2023.2289881

Jones, B., Jones, R., & Luger, E. (2022). AI 'everywhere and nowhere': Addressing the AI intelligibility problem in public service journalism. *Digital Journalism*, *10*(10), 1731–1755. https://doi.org/10.1080/21670811.2022.2145328

Keefe, J., Zhou, Y., & Merrill, J. (2021, May 12). The present and potential of AI in journalism. *Knight Foundation Journalism*. https://knightfoundation.org/articles/the-present-and-potential-of-ai-in-journalism

Kim, D., & Kim, S. (2018). Newspaper journalists' attitudes towards robot journalism. *Telematics and Informatics*, *35*(2), 340–357. https://doi.org/10.1016/j.tele.2017.12.009

Komatsu, T., Lopez, G. M., Makri, S., Porlezza, C., Cooper, G., MacFarlane, A., & Missaoui, S. (2020). AI should embody our values: Investigating journalistic values to inform AI technology design. In *NordiCHI '20: Proceedings of the 11th Nordic Conference on Human–Computer Interaction: Shaping Experiences, Shaping Society* (Article 11). Association for Computing Machinery. https://doi.org/10.1145/3419249.3420105

Kosterich, A. (2020). Managing news nerds: Strategizing about institutional change in the news industry. *Journal of Media Business Studies*, *17*(1), 51–68. https://doi.org/10.1080/16522354.2019.1639890

Kothari, A., & Cruikshank, S. A. (2021). Artificial intelligence and journalism: An agenda for journalism research in Africa. *African Journalism Studies*, *43*(1), 17–33. https://doi.org/10.1080/23743670.2021.1999840

Kreft, J., Boguszewicz-Kreft, M., & Fydrych, M. (2023). (Lost) pride and prejudice. Journalistic identity negotiation versus the automation of content. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2023.2289177

Krumsvik, A. H., Milan, S., Bhroin, N. N., & Storsul, T. (2019). Making (sense of) media innovations. In M. Deuz & M. Prenger (Ed.), *Making media: Production, practices, and professions* (pp. 193–206). Amsterdam University Press. https://doi.org/10.1515/9789048540150-014

Leonardi, P., & Neeley, T. (2022). *The digital mindset: What it really takes to thrive in the age of data, algorithms, and AI*. Harvard Business Review Press.

Lewis, S. C., Guzman, A. L., & Schmidt, T. R. (2019). Automation, journalism, and human–machine communication: Rethinking roles and relationships of humans and machines in news. *Digital Journalism*, *7*(4), 409–427. https://doi.org/10.1080/21670811.2019.1577147

Lewis, S. C., & Westlund, O. (2015). Actors, actants, audiences, and activities in cross-media news work: A matrix and a research agenda. *Digital Journalism*, *3*(1), 19–37. https://doi.org/10.1080/21670811.2014.927986

Lindblom, T., Lindell, J., & Gidlund, K. (2022). Digitalizing the journalistic field: Journalists' views on changes in journalistic autonomy, capital and habitus. *Digital Journalism*, *12*(6), 894–913. https://doi.org/10.1080/21670811.2022.2062406

Lindén, C.-G. (2017a). Algorithms for journalism: The future of news work. *The Journal of Media Innovations*, *4*(1), 60–76. https://doi.org/10.5617/jmi.v4i1.2420

Lindén, C.-G. (2017b). Decades of automation in the newsroom. *Digital Journalism*, *5*(2), 123–140. https://doi.org/10.1080/21670811.2016.1160791

Lindén, C.-G., & Tuulonen, H. (Eds.). (2019). *News automation: The rewards, risks and realities of 'machine journalism.'* WAN-IFRA.

Lischka, J. A., Schaetz, N., & Oltersdorf, A.-L. (2022). Editorial technologists as engineers of journalism's future: Exploring the professional community of computational journalism. *Digital Journalism*, *11*(6), 1026–1044. https://doi.org/10.1080/21670811.2021.1995456

Litskevich, V. (2022, October 24). How much does artificial intelligence cost in 2021? *Azati*. https://azati.ai/how-much-does-it-cost-to-utilize-machine-learning-artificial-intelligence

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, D. E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/j.joi.2018.09.002

Mathieson, K. (1991). Predicting user intentions: Comparing the technology acceptance model with the theory of planned behavior. *Information Systems Research*, *2*(3), 173–191. https://doi.org/10.1287/isre.2.3.173

Mills, J., & Wagemans, A. (2021). Media labs: Constructing journalism laboratories, innovating the future: How journalism is catalysing its future processes, products and people. *Convergence: The International Journal of Research Into New Media Technologies*, *27*(5), 1462–1487. https://doi.org/10.1177/1354856521994453

Milosavljević, M., & Vobič, I. (2021). 'Our task is to demystify fears': Analysing newsroom management of automation in journalism. *Journalism*, *22*(9), 2203–2221. https://doi.org/10.1177/1464884919861598

Missaoui, S., Gutierrez-Lopez, M., MacFarlane, A., Makri, S., Porlezza, C., & Cooper, G. (2019). How to blend journalistic expertise with artificial intelligence for research and verifying news stories. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://openaccess.city.ac.uk/id/eprint/22996

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*, *8*(5), 336–341. https://doi.org/10.1016/j.ijsu.2010.02.007

Møller, H. J., & Thylstrup, N. B. (2024). The algorithmic gut feeling—Articulating journalistic doxa and emerging epistemic frictions in AI-driven data work. *Digital Journalism*. Advance online publication. https://doi.org/10.1080/21670811.2024.2319641

Montaña-Niño, S. X., & Burgess, J. (2024). Beyond the 'critical incident': Covid-19, data journalism and the slow road to editorial automation in Australian newsrooms. *New Media & Society*, *26*(3), 1315–1332. https://doi.org/10.1177/14614448231201644

Moran, R. E., & Shaikh, S. J. (2022). Robots in the news and newsrooms: Unpacking meta-journalistic discourse on the use of artificial intelligence in journalism. *Digital Journalism*, *10*(10), 1756–1774. https://doi.org/10.1080/21670811.2022.2085129

Motta, M., Stecula, D., & Farhart, C. E. (2020). How right-leaning media coverage of Covid-19 facilitated the spread of misinformation in the early stages of the pandemic. *Canadian Journal of Political Science*, *53*(2), 335–342. https://doi.org/10.1017/S0008423920000396

Munoriyarwa, A., Chiumbu, S., & Motsaathebe, G. (2021). Artificial intelligence practices in everyday news production: The case of South Africa's mainstream newsrooms. *Journalism Practice*, *17*(7), 1374–1392. https://doi.org/10.1080/17512786.2021.1984976

Nelson, A. J., & Irwin, J. (2014). "Defining what we do—all over again": Occupational identity, technological change, and the librarian/internet-search relationship. *Academy of Management Journal*, *57*(3), 892–928. https://doi.org/10.5465/amj.2012.0201

Newman, N. (2021). *Journalism, media, and technology trends and predictions 2021*. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2021

Newman, N. (2022). *Journalism, media, and technology trends and predictions 2022*. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2022

Noain-Sánchez, A. (2022). Addressing the impact of artificial intelligence on journalism: The perception of experts, journalists and academics. *Communication & Society*, *35*(3), 105–121. https://doi.org/10.15581/003.35.3.105-121

Noor, R., & Zafar, H. (2023). Use of artificial intelligence in Pakistani journalism: Navigating challenges and future paths in TV newsrooms. *Journal of Asian Development Studies*, *12*(3), 131–150. https://doi.org/10.62345/jads.2023.12.3.131

Nunes, A. C. B., & Mills, J. (2021). Journalism innovation: How media labs are shaping the future of media and journalism. *Brazilian Journalism Research*, *17*(3), 652–679. https://doi.org/10.25200/BJR.v17n3.2021.1440

Okocha, D. O., & Ola-Akuma, R. O. (2022). Journalistic metamorphosis: Robot journalism adoption in Nigeria in a digital age. *IGWEBUIKE: An African Journal of Arts and Humanities*, *8*(1), 255–267. https://doi.org/10.13140/RG.2.2.29105.45929

Olsen, G. R. (2023). Enthusiasm and alienation: How implementing automated journalism affects the work meaningfulness of three newsroom groups. *Journalism Practice*, *19*(2), 304–320. https://doi.org/10.1080/17512786.2023.2190149

Örnebring, H. (2010). Technology and journalism-as-labour: Historical perspectives. *Journalism*, *11*(1), 57–74. https://doi.org/10.1177/1464884909350644

Paik, S. (2023). Journalism ethics for the algorithmic era. *Digital Journalism*. Advance online publication. https://doi.org/10.1080/21670811.2023.2200195

Paulussen, S. (2016). Innovation in the newsroom. In T. Witschge, C. W. Anderson, D. Domingo, & A. Hermida (Eds.), *The Sage handbook of digital journalism* (pp. 192–206). Sage.

Pentzold, C., Fechner, D. J., & Zuber, C. (2021). "Flatten the curve": Data-driven projections and the journalistic brokering of knowledge during the Covid-19 crisis. *Digital Journalism*, *9*(9), 1367–1390. https://doi.org/10.1080/21670811.2021.1950018

Perreault, G. P., & Ferrucci, P. (2020). What is digital journalism? Defining the practice and role of the digital journalist. *Digital Journalism*, *8*(10), 1298–1316. https://doi.org/10.1080/21670811.2020.1848442

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Rodríguez, M. T., Nunes, S., & Devezas, T. (2015). Telling stories with data visualization. In *NHT '15: Proceedings of the 2015 Workshop on Narrative & Hypertext* (pp. 7–11). Association for Computing Machinery. https://doi.org/10.1145/2804565.2804567

Rydenfelt, H. (2021). Transforming media agency? Approaches to automation in Finnish legacy media. *New Media & Society*, *24*(12), 2598–2613. https://doi.org/10.1177/1461444821998705

Ryfe, D. (2019). Institutional theory and journalism. In T. P. Vos & F. Hanusch (Eds.), *The international encyclopedia of journalism studies*. https://doi.org/10.1002/9781118841570.iejs0037

Salaverría, R., & de-Lima-Santos, M.-F. (2020). Towards ubiquitous journalism: Impacts of IoT on news. In J. Vázquez-Herrero, S. Direito-Rebollal, A. Silva-Rodríguez, & X. López-García (Eds.), *Journalistic metamorphosis: Media transformation in the digital age* (pp. 1–15). Springer. https://doi.org/10.1007/978-3-030-36315-4_1

Schapals, A. K., & Porlezza, C. (2020). Assistance or resistance? Evaluating the intersection of automated

journalism and journalistic role conceptions. *Media and Communication*, *8*(3), 16–26. https://doi.org/10.17645/mac.v8i3.3054

Schjøtt Hansen, A., & Hartley, J. M. (2021). Designing what's news: An ethnography of a personalization algorithm and the data-driven (re)assembling of the news. *Digital Journalism*, *11*(6), 924–942. https://doi.org/10.1080/21670811.2021.1988861

Schultz, I. (2007). The journalistic gut feeling: Journalistic doxa, news habitus, and orthodox news values. *Journalism Practice*, *1*(2), 190–207. https://doi.org/10.1080/17512780701275507

Shah, M. H. A., Khoso, I. A., & Dharejo, N. (2024). Journalist perceptions and views towards the integration of AI-based applications in the journalism industry in Pakistan: Expansion of the UTAUT model. *Annals of Human and Social Sciences*, *5*(2), 317–326. https://doi.org/10.35484/ahss.2024(5-ii)30

Sharadga, T. M., Tahat, Z., & Safori, A. O. (2022). Journalists' perceptions towards employing artificial intelligence techniques in Jordan TV's newsrooms. *Studies in Media and Communication*, *10*(2), 239–248. https://doi.org/10.11114/smc.v10i2.5749

Sholola, Y. A., Banjo, Y. O., Saliu-Yusuf, M. J., Ogundeyi, T. S., & Ayantunji, K. A.-A. (2024). Perceived effect of artificial intelligence on ethical journalism among journalists in Kwara State. *International Journal of Research and Innovation in Social Science*, *8*(2), 2051–2063. https://doi.org/10.47772/ijriss.2024.802145

Siitonen, M., Laajalahti, A., & Venäläinen, P. (2023). Mapping automation in journalism studies 2010–2019: A literature review. *Journalism Studies*, *25*(3), 299–318. https://doi.org/10.1080/1461670x.2023.2296034

Soto-Sanfiel, M. T., Ibiti, A., Machado, M., Marín Ochoa, B. E., Mendoza Michilot, M., Rosell Arce, C. G., & Angulo-Brunet, A. (2022). In search of the Global South: Assessing attitudes of Latin American journalists to artificial intelligence in journalism. *Journalism Studies*, *23*(10), 1197–1224. https://doi.org/10.1080/1461670x.2022.2075786

Spyridou, P., & Danezis, C. (2024). Do algorithms do it better? Analysing occupational ideology in the age of computational journalism. *Journalism Studies*, *25*(13), 1573–1597. https://doi.org/10.1080/1461670X.2024.2372429

Steensen, S. (2018). What is the matter with newsroom culture? A sociomaterial analysis of professional knowledge creation in the newsroom. *Journalism*, *19*(4), 464–480. https://doi.org/10.1177/1464884916657517

Stray, J. (2019). Making artificial intelligence work for investigative journalism. *Digital Journalism*, *7*(8), 1076–1097. https://doi.org/10.1080/21670811.2019.1630289

Svensson, J. (2021). Logics, tensions and negotiations in the everyday life of a news-ranking algorithm. *Journalism*, *24*(7), 1518–1535. https://doi.org/10.1177/14648849211063373

Tariq, M., Aslam, M. J., Shakoor, A., & Ilyas, S. (2024). Artificial Intelligence and the reshaping of journalism. *Qlantic Journal of Social Sciences*, *5*(1), 44–53. https://qjss.com.pk/index.php/qjss/article/view/222

Tejedor, S., & Vila, P. (2021). Exo journalism: A conceptual approach to a hybrid formula between journalism and artificial intelligence. *Journalism and Media*, *2*(4), 830–840. https://doi.org/10.3390/journalmedia2040048

Thurman, N., Dörr, K., & Kunert, J. (2017). When reporters get hands-on with robo-writing: Professionals consider automated journalism's capabilities and consequences. *Digital Journalism*, *5*(10), 1240–1259. https://doi.org/10.1080/21670811.2017.1289819

Trang, T. T., Chien Thang, P., Hai, L. D., Phuong, V. T., & Quy, T. Q. (2024). Understanding the adoption of artificial intelligence in journalism: An empirical study in Vietnam. *Sage Open*, *14*(2). https://doi.org/10.1177/21582440241255241

Túñez-López, J.-M., Fieiras-Ceide, C., & Vaz-Álvarez, M. (2021). Impact of artificial intelligence on journalism: Transformations in the company, products, contents, and professional profile. *Communication & Society*, *34*(1), 177–193. https://doi.org/10.15581/003.34.1.177-193

Usher, N. (2017). Venture-backed news startups and the field of journalism. *Digital Journalism*, *5*(9), 1116–1133. https://doi.org/10.1080/21670811.2016.1272064

van Dalen, A. (2012). The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists. *Journalism Practice*, *6*(5/6), 648–658. https://doi.org/10.1080/17512786.2012.667268

van Drunen, M. Z., & Fechner, D. (2022). Safeguarding editorial independence in an automated media system: The relationship between law and journalistic perspectives. *Digital Journalism*, *11*(9), 1723–1750. https://doi.org/10.1080/21670811.2022.2108868

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

Vos, T. P., Craft, S., & Ashley, S. (2012). New media, old criticism: Bloggers' press criticism and the journalistic field. *Journalism*, *13*(7), 850–868. https://doi.org/10.1177/1464884911421705

Ward, S. (2018). *Disrupting journalism ethics: Radical change on the frontier of digital media*. Routledge. https://doi.org/10.4324/9781315179377

Weiswasser, S. (1997). Role of technology in communication. *Fordham International Law Journal*, *21*(2), 439–444.

Wölker, A., & Powell, T. (2021). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, *22*(1), 86–103. https://doi.org/10.1177/1464884918757072

Yu, Y., & Huang, K. (2021). Friend or foe? Human journalists' perspectives on artificial intelligence in Chinese media outlets. *Chinese Journal of Communication*, *14*(4), 409–429. https://doi.org/10.1080/17544750.2021.1915832

Zaid, B., Ibahrine, M., & Fedtke, J. (2022). The impact of the platformization of Arab news websites on quality journalism. *Global Media and Communication*, *18*(2), 243–260. https://doi.org/10.1177/17427665221098022

Zaragoza Fuster, M. T., & García Avilés, J. A. (2022). Public service media laboratories as communities of practice: Implementing innovation at BBC News Labs and RTVE Lab. *Journalism Practice*, *18*(5), 1256–1274. https://doi.org/10.1080/17512786.2022.2088602

Zhou, Y. (2008). Voluntary adopters versus forced adopters: Integrating the diffusion of innovation theory and the technology acceptance model to study intra-organizational adoption. *New Media & Society*, *10*(3), 475–496. https://doi.org/10.1177/1461444807085382

About the Authors

Sangyon Oh is a PhD candidate at the Korea Advanced Institute of Science and Technology (KAIST). With nearly two decades of journalistic experience, including current work at MBC in Korea, she has been actively involved in all facets of news reporting, production, and broadcasting. Building upon her extensive professional background, her research delves into the intersection of artificial intelligence, journalism ethics, and the rapidly evolving landscape of news production. She focuses on how emerging technologies, particularly AI, are transforming the workflows of journalists, the production of news content, and the ways in which audiences consume and engage with media.

Jaemin Jung (PhD, University of Florida) is a professor at the Moon Soul Graduate School of Future Strategy at the Korea Advanced Institute of Science and Technology (KAIST). His research focuses on media management, media economics, and the impact of AI on journalism and media industries, with a keen interest in exploring how AI technologies are reshaping the landscape of news production and media consumption.

Media and Communication
2025 • Volume 13 • Article 9639
https://doi.org/10.17645/mac.9639

cogitatio

ARTICLE

Open Access Journal

Who Wants to Try AI? Profiling AI Adopters and AI-Trusting Publics in South Korea

Hyelim Lee [1], Chanyoung Jung [2], Nayoung Koo [3], Seongbum Seo [3], Sangbong Yoo [4], Hyein Hong [5], and Yun Jang [3]

[1] College of Media & Communication, Korea University, Republic of Korea
[2] Data Science Group, INTERX, Republic of Korea
[3] Sejong University, Republic of Korea
[4] AI, Information and Reasoning Laboratory, Korea Institute of Science and Technology, Republic of Korea
[5] Miridih, Republic of Korea

Correspondence: Hyelim Lee (hyelim_lee@korea.ac.kr)

Submitted: 16 November 2024 **Accepted:** 5 March 2025 **Published:** 14 May 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

The study introduces new approaches that integrate public relations concepts to identify adopters of AI technologies. It seeks to discover novel methods for identifying target groups likely to adopt AI technologies, even in contexts where these technologies are not yet familiar. The study achieves this by employing latent profile analysis, regression, and structural equation modeling using data from a South Korean online panel survey (N = 625). The results identify five distinct public profiles based on their relationship assessments, problem recognition, and general trust in AI: the Cautious, Balanced, Uninterested, Confident, and Enthusiastic. Notably, the Enthusiastic group—characterized by high trust and strong relationships with the service provider—showed the strongest interest in adopting AI, highlighting their potential as key publics for introducing new AI services in business contexts. Additionally, the article contributes to public segmentation theory through the use of the situational theory of problem-solving and enhances the applicability of established public relations frameworks to the evolving field of AI.

Keywords

AI; AI adoption; public relations; organization–public relationship; situational theory of problem-solving

1. Introduction

Throughout history, every civilization has experienced fear in response to new technologies. Such fear can deeply permeate society, slowing the adoption of technological innovations (Orben, 2020). This underlying panic stems from various factors with some scholars tracing it to psychological reactance (Contzen et al., 2021; Feng et al., 2019). People may fear that a single change could trigger a butterfly effect, disrupting various aspects of life. These sweeping changes can make individuals fear losing their stability and threaten their sense of normalcy. Moreover, those who have become accustomed to using a certain technology for a long time may feel burdened by the need to learn new systems, which often requires time, energy, and, sometimes, financial investment. From an economic standpoint, not all users readily embrace new technologies.

Despite these concerns, the structure of capitalism in the information age drives technological innovations. New technologies often directly correlate with business expansion opportunities. In today's globalized and interconnected world, developing technologies that surpass existing capabilities is ongoing. Companies or countries that gain early access to superior technology can dominate entire economic systems, as modern life increasingly depends on information and communication technologies. Therefore, while public resistance to new technologies persists, entities seeking to develop and introduce innovations consistently look for strategies to popularize their technologies and encourage active consumer adoption.

AI is no different from past innovations when it comes to stirring public anxiety. In 2023, OpenAI launched its generative AI platforms like ChatGPT, which quickly raised alarms about machines taking over human jobs. This emergence sparked widespread debate about job markets, industrial structures, and the broader implications for the value of human labor (Farina & Lavazza, 2023). We can attribute these unprecedented societal concerns about AI to various unresolved ethical issues. That is, as AI-based techniques often function as "black boxes," it is difficult to assess whether their algorithms have a critical bias or infringe on privacy. These built-in ethical issues make people more skeptical of AI today than they were with earlier technologies (Mbiazi et al., 2023).

People's subtle but growing anxieties now focus on the possibility of widespread layoffs, impacting administrators, clerks handling routine paperwork, and, ironically, even developers due to AI's fast and accurate programming capabilities (Constantz & Bloomberg, 2024). Given the current public perception of AI as a threat to humanity, businesses introducing new AI-based services must be cautious to avoid potential customer backlash. Businesses can prevent such a situation and enhance competitiveness by identifying which groups are willing to adopt these innovations.

The diffusion of innovations theory (Rogers, 2003) emphasizes the critical role of early adopters in promoting new technologies. This theory posits that early adopters share several key characteristics. For instance, in a meta-analytic review, Ortt et al. (2017) found that early adopters exhibit high levels of enthusiasm toward new technology, including innovativeness, economic motivation, opinion leadership, and a strong desire to communicate. They tend to engage more with technology and have access to useful resources such as prior experience, technical skills, social networks, and relevant knowledge. Businesses can leverage these factors to identify potential pioneers who can help attract more users to the market. While the role of early adopters in innovation diffusion is sometimes unclear (Bianchi et al., 2017) and can vary across different products or services (Frattini et al., 2014), they play a crucial part in spreading information to non-users, aiding in their understanding, awareness, and decision-making.

Scholars frequently turn to the diffusion of innovations theory (Rogers, 2003) when discussing how new technologies gain traction. The theory is particularly useful for tracking macro-level diffusion trends like the S-curve based on how quickly different groups adopt innovation. However, the theory can overemphasize the influence of innovators and early adopters. As previous literature suggests, individuals categorized as innovators or early adopters make up less than 20% of the population (Ortt et al., 2017) and often come from privileged or higher social status backgrounds. With innovative technologies now more accessible to the general population, it becomes crucial to understand who will adopt AI services when introduced by familiar companies.

In today's technological ecosystem, tech companies—such as those in e-commerce, social networking platforms, and telecommunications—actively promote their technological advancements to customers. Since most people rely on communication technologies, they frequently encounter marketing messages from brands they already use. This loyalty-based ecosystem helps encourage users to try new technologies (Prins & Verhoef, 2007).

Within this context, this study identifies segments of the public who are inclined to try new AI services offered by companies they already use. These individuals may not be early adopters or innovators with abundant resources, but they represent a crucial public. Businesses that focus solely on targeting early adopters may overlook the needs and preferences of more typical users. Therefore, expanding the target public to include potential early adopters who differ from the traditional early adopter profile allows companies to reach a broader market.

This study addresses this issue by employing a public relations perspective, focusing on the importance of relationship quality and situational factors. In public relations, researchers often use organization–public relationship (OPR) and problem recognition to classify and profile key publics. Applying these concepts offers new insights into identifying traits that make certain users more likely to adopt AI. In sum, this study applies a public relations perspective to identify the "AI-trusting public." By examining OPR and problem recognition, it explores how trust in the organization and situational motivation shape people's willingness to adopt AI-based services.

The article unfolds in five sections. First, the literature review discusses the current issues surrounding public trust in AI and examines and presents the study's theoretical concepts, including OPR, situational theory of problem-solving (STOPS), and the research questions and hypotheses. Next, the methods section details the study's data collection process and survey items. Then, the results section follows, presenting findings from regression and structural equation modeling (SEM) analyses. Finally, the article concludes with a discussion of the theoretical and practical implications before offering closing remarks.

2. Literature Review

2.1. Trust Issues about AI

A prevailing distrust toward technology is undeniable. Pieters (2011) outlines the rationale behind how individuals develop trust or a sense of reliability toward technology, observing that most new technologies, including AI, often feel like "black boxes." Most people find it difficult to understand how AI works because

its algorithms are so complex. Even experts struggle to keep up with the speed at which AI-driven algorithms process, compute, and analyze data. Scholars and experts have voiced growing concern about AI's invasive impact in the workplace (e.g., Hunt, 2023; Roose, 2023).

News media outlets further diffuse and amplify these sentiments through casual reporting (e.g., Kelly, 2024) and political commentary (Orben, 2020). Consequently, the current negative aura surrounding AI technology can appear "extraordinary" (Battista, 2024) or "magical" (Nagy & Neff, 2024). Widespread concerns about AI's impact on human labor resonate widely with experts and the general population, fueling the broader trust issues people feel toward these technologies (Omrani et al., 2022). For businesses promoting AI, this commonly shared apprehension poses significant challenges (Gerlich, 2023).

Researchers have not just observed this anxiety; they have actively explored the issues of transparency and trust in automated AI-based systems. For instance, Langer et al. (2023) found that users showed very little trust in low-level automated systems, and even when they introduced interventions, they made little difference. Similarly, in the UK, a series of recent studies on AI-based virtual humans and job displacement (Gerlich, 2024a, 2024b) revealed that participants felt anxious and deeply concerned about data exploitation and potential consequences professionally and personally. These findings underscore the importance of transparent AI governance. In response, the Ada Lovelace Institute, a UK-based think tank, proposed policy-based solutions to address society's fear of AI in a recent report (Davies & Birtwistle, 2023).

Trust in technology plays a crucial role in the adoption, use, and application of new technologies. When people hesitate to engage with unfamiliar technologies, that fear can grow stronger, even before any firsthand experience. This article does not aim to assess whether AI will ultimately benefit society, nor does it attempt a deep dive into the ethics of AI usage. Instead, we focus on the pressing need to address the public's trust issues with AI technologies in a more concrete and meaningful way.

However, trust is not a simple concept. Lankton et al. (2015) identified two main concepts: "human-like trusting beliefs" and "system-like trusting beliefs." They further categorized these into sub-concepts: Human-like trust beliefs entail integrity, ability, competence, and benevolence—qualities similar to those found in dyadic human relationships; conversely, system-like trust beliefs include reliability, functionality, and helpfulness, which focus more on operational effectiveness. Thus, when people distrust technology, they tend to question its "human-like" traits, especially whether the technology or those who create it actually have their best interests at heart. Mayer et al. (1995, pp. 718–719) describe this form of trust as "the belief that the trustee will want to do good to the trustor, aside from an egocentric profit motive," implying expectations of the technology's non-harmful intentions.

Following this logic, the negative tone surrounding AI today (Frank et al., 2023), amplified by media and experts, can further deepen public suspicion toward emerging AI services. Businesses find themselves up against this wave of skepticism, which can feel overwhelming. The difficulty in solving such a problem within the business may stem from focusing solely on the relationship between technology and users. However, businesses that offer AI-driven services can act as connectors or mediators, enhancing accessibility to AI technologies by drawing on the relationships they've already built with customers over time. For example, Ameen et al. (2021) found that a company's commitment to its customer relationships plays a significant role in how people experience AI-enabled services.

Given this background, we suggest a new direction: Instead of only analyzing general public trust in AI, we should identify specific user groups that are more open to adopting these services. Some people may show interest due to particular situations, like curiosity about a new technology or a broader concern with it. Others may trust the AI services simply because they trust the company offering them. By looking more closely at these groups, we explore new possibilities for encouraging meaningful and responsible adoption of AI technologies.

2.2. Identifying AI-Trusting 'Publics'

There are several useful theoretical frameworks to understand why people adopt novel ideas and technologies. Among them, diffusion of innovations by Rogers (2003) remains one of the most influential. Rogers introduced a framework for identifying early adopters, those who tend to embrace innovations before others. He defined innovation as "an idea, practice, or project that is perceived as new by an individual or other unit of adoption" (Rogers, 2003, p. 12). His theory treats innovation as inclusive, even when no one in a particular community has previously used it; for example, boiling water in a Peruvian village. Scholars have widely applied this framework to understand how new technologies spread, especially as information and communication technologies have rapidly evolved over recent decades.

Building on Rogers' legacy, many researchers have explored what drives individuals to adopt technology early. For example, Son and Han (2011) identified technology readiness—defined as people's propensity to embrace and use new technologies—as a key predictor of adoption and satisfaction. Balkrishan and Joshi (2013) proposed the use–usage model, which considers multiple environmental factors that shape adoption, including factors like prevalence, utility, and proactive attitudes toward technology. Mahardika et al. (2019) examined impediments to adopting new technologies, highlighting distinctions between behavioral intention and behavioral expectation, which vary according to individuals' experience levels with technology.

These studies primarily focus on an individual's propensity or disposition toward adoption. However, the adoption process for new technologies and innovative tech services is more dynamic; it evolves in response to public awareness and how trustworthy people perceive the sources promoting the innovation.

Besides the diffusion of innovations theory, researchers often turn to other well-established theoretical frameworks such as the technology acceptance model (Davis, 1989) and the unified theory of acceptance and use of technology (Venkatesh et al., 2003). Both models elucidate utility-based factors such as ease of use, perceived usefulness, performance expectancy, and effort expectancy. The premise of these two theories is that users are rational decision-makers who aim to maximize benefits while minimizing the psychological or actual costs of adoption behaviors.

2.2.1. OPR

Unlike previous efforts to identify the strategies or tactics of adopters through frameworks like the diffusion of innovation theory, the technology acceptance model, or the unified theory of acceptance and use of technology, this study emphasizes the significance of relational factors, a core element of public relations research. Public relations revolves around building, maintaining, and cultivating relationships with key publics or stakeholders to help organizations reach their goals. Reaching relationship-based goals requires a

long-term investment in creating and nurturing relationships with the public, which becomes central to the decision-making process (Grunig & Hunt, 1984).

Public relations strategies can facilitate the adoption of new technologies by leveraging existing relationships, making it easier for the public to accept and engage with these innovations. However, despite the potential for public relations to influence technology adoption, research has largely concentrated on how public relations might innovate through technology (e.g., Panopoulos et al., 2018) rather than how it can help organizations guide the public toward adopting new technologies.

Public relations concepts offer significant potential for helping organizations identify early adopters when launching new tech services. Among these, the OPR measures how the public evaluates an organization's efforts to build and maintain relationships (e.g., Hon & Grunig, 1999; Huang, 2001a), perceived as their assessment of their connection with the organization. In this context, OPR offers a practical advantage: it helps managers identify favorable customers—those more likely to try new products and possibly influence others to do the same.

Public relations research has extensively studied the antecedents and outcomes of strong OPRs. For example, Huang (2001b) identified several strategies that advance OPR, including mediated communication, social activities, two-way dialogue, ethical and symmetrical communication, and interpersonal interaction. Huang also found that OPR can mediate conflict effectively, using strategies like non-confrontational communication and third-party resolution.

Researchers have also looked at the outcomes of OPR in more detail (Cheng, 2018). For instance, Yang (2007) found that a good OPR can positively affect an organization's reputation. Similarly, Kazoleas and Wright (2001) reported that a strong organization–employee relationship can improve employee morale and job satisfaction. Hon (1997) identified 15 recurring themes from expert interviews showing how OPR can improve public relations effectiveness, such as changing public attitudes, perceptions, and behaviors.

This body of research suggests that OPR can significantly influence how people respond to organizations. Therefore, we can reasonably infer that individuals who feel a strong positive connection with an organization will be more open to trying new services from that organization, even when those services involve technologies that might face general skepticism. For example, when people trust an organization, they are more likely to see its offerings as reliable and credible, regardless of general hesitation toward adopting new technologies.

2.2.2. Problem Recognition of STOPS

Another public relations concept that can help identify key adopters who may become champions of new services is the STOPS (Kim & Grunig, 2011). STOPS views members of the public as active communicators and problem solvers in situations they perceive as problematic. When a member of the public begins to recognize a specific issue as a "problem," they become motivated to act. This motivation often drives communication behaviors known as communicative actions for problem-solving.

Three key factors influence this situational motivation: problem recognition, constraint recognition, and involvement recognition. Among these, problem recognition plays a critical role by raising individuals'

awareness of an issue and prompting them to conceptualize it as a problem. Grunig (1997) initially defined problem recognition as the moment when people realize that something needs attention and start thinking about how to respond. Later, Kim and Grunig (2011) refined this idea, defining problem recognition as "one's perception that something is missing and that there is no immediately applicable solution to it" (p. 128). Following this logic, without the public's recognition of an issue as a problem, an organization's decisions have minimal impact on the environment or its public. However, once people start framing an issue as a problem, they are more likely to take action and get involved in solving it.

In traditional technology adoption studies, we can draw parallels between problem recognition and how people perceive the prevalence, usefulness, or relevance of new technologies (e.g., Balkrishan & Joshi, 2013). However, problem recognition goes a step further. It reflects how compelled people feel to take an interest in new technology based on their recognition of a related problem. For example, if the public sees current challenges in information technology as urgent problems, that recognition can create a stronger drive to discover new solutions. In contrast, people who do not view tech-related issues as problems are less likely to see value in adopting new and innovative services, such as AI-powered tools. Essentially, how people recognize the issue can significantly shape their openness to new technologies.

3. Research Questions and Hypotheses

Given the research background and literature review, we propose two research questions and four hypotheses. The research questions explore the profiles of individuals more or less likely to adopt new AI services, considering their levels of general trust in AI, their relationship with the current service provider, and their recognition of technology-related problems. The hypotheses predict relationships between relationship metrics, levels of problem recognition, trust in AI services, and adoption intentions. The specific research questions and hypotheses are as follows.

RQ1: Which members of the public intend to adopt new AI services based on their general trust in AI, their relationship with the current service provider, and their recognition of technology-related problems?

RQ2: Which members of the public express trust in new AI services, considering their general trust in AI, their relationship with the current service provider, and their recognition of technology-related problems?

H1: Higher levels of (a) OPR and (b) technology problem recognition (TPR) will predict stronger intentions to adopt AI services.

H2: Higher levels of (a) OPR and (b) TPR will predict greater trust in AI services.

H3: OPR, TPR, and general trust in AI technologies will interact to influence (a) AI service adoption intention and (b) trust in AI services.

H4: Trust in AI services will mediate the relationship between (a) OPR and AI service adoption intention and (b) TPR and AI service adoption intention.

4. Methods

4.1. Data

One of the major Korean telecommunications operators provided us with a secondary dataset, which Kantar Korea, an independent online panel company, collected with informed consent from all survey participants. In February 2023, Kantar Korea surveyed 625 South Korean respondents using 40 questions focused on customer satisfaction with telecommunications services and AI products.

The demographics of the sample are as follows: The gender distribution was nearly equal, with 49.3% female (n = 308) and 50.7% male (n = 317). The average age of participants was 33.2 years. Over half of the respondents reported a monthly household income between 3 to 7 million Korean Won (approximately USD 2,200–5,200, with an exchange rate of 1 USD = 1,400 KRW; n = 316, 50.6%). Most respondents had a college-level education (n = 435, 73.1%) and lived in Seoul-si or Gyeonggi-do (n = 316, 49.6%).

4.2. Measures

This study measured variables based on previous research using established frameworks from OPR and STOPS (Kim & Grunig, 2011). Respondents rated all survey items on a five-point Likert scale. All variables showed acceptable internal validity, with Cronbach's alpha values exceeding 0.6 and average variance extracted (AVE) values above 0.5 (Ahmad et al., 2016). In addition, to reduce common method variance, we designed the survey to separate predictor and outcome variables clearly and applied techniques to minimize response bias (Eichhorn, 2014). Appendix 1 of the Supplementary File presents the full set of survey items.

We assessed OPR using four items that captured respondents' perceptions of their relationship with their telecommunications provider. An example item is: "I have confidence in the capabilities of Company A" (α = 0.86, AVE = 0.78), rated on a scale from 1 = *strongly disagree* to 5 = *strongly agree*.

We measured TPR with four items that examined respondents' awareness of technology-related issues. For example, "I often think about the lawsuit regarding Operator X's broadband network usage fees" (α = 0.86, AVE = 0.75). Responses ranged on a scale from 1 = *strongly disagree* to 5 = *strongly agree*.

To assess general trust toward AI technologies, we asked questions such as, "With the introduction of AI, society is gradually transitioning into a new era. How do you view the societal changes brought about by artificial intelligence?" The response ratings were from 1 = *very negative view* to 5 = *very positive view*.

We measured the adoption intention of AI services using six items, including: "Would you be willing to use a service that provides messages and gifts to your loved ones on birthdays or anniversaries even after your passing?" (α = 0.87, *AVE* = 0.72; 1 = *very negative*; 5 = *very positive*).

Finally, we evaluated trust toward AI services from telecommunication operator X using seven items, including: "If AI were to replace humans in sports (e.g., managing games and officiating), how much would you trust it?" (α = 0.83, *AVE* = 0.61), rated from 1 = *strongly distrust* to 5 = *strongly trust*.

Table 1 presents the zero-order correlations among the five key variables, reporting average-to-moderate mean values across variables, with the highest mean for OPR ($M = 3.42$) and the lowest for TPR ($M = 2.64$).

Table 1. Zero-order correlations.

	Variables	1	2	3	4	5
1	OPR	1.00				
2	TPR	0.04	1.00			
3	General Trust in AI	0.14***	0.13***	1.00		
4	AI Service Adoption Intention	0.14***	0.29***	0.29***	1.00	
5	Trust Towards AI Services	0.05	0.21***	0.38***	0.38***	1.00
	M	3.42	2.64	3.40	2.94	3.27
	SD	0.81	0.93	0.88	0.86	0.68

Note: *** $p < 0.01$.

5. Results

We used the Stata 18 MP version for latent profile analysis (LPA), regression, and SEM analysis.

5.1. LPA

LPA uses "a categorical latent variable approach that focuses on identifying latent subpopulations based on a certain set of variables" (Spurk et al., 2020, p. 1). This approach benefits research that seeks to identify previously unknown groups within the population. To that end, RQ1 and RQ2 aimed to identify potential adopters of AI services through a public relations lens. We used three latent variables to define public segments: OPR, technological problem recognition, and general trust in AI technologies.

As Table 2 shows, the five-class model (Class 5) emerged as the optimal model. It yielded the lowest values for key model fit indicators: Bayesian information criterion (4,078.4), sample-size adjusted Bayesian information criterion (4,064.1), and a statistically significant Lo–Mendell–Rubin test ($p = 0.00$). Although the six-class model (Class 6) had a slightly lower Akaike information criterion (3,977.1), other indicators supported Class 5 as the most suitable for creating distinctive profiles (Debus et al., 2024). Further, following Occam's razor principle, a long-standing scientific principle that favors simpler explanations over unnecessarily complex ones,

Table 2. LPA model comparison.

Class	N	Log-likelihood	Free parameters	Akaike information criterion	Bayesian information criterion	Sample-size adjusted-Bayesian information criterion	Lo–Mendell–Rubin
2	625	−2,386.367	10	4,792.734	4,837.112	4,822.83	0.000
3	625	−2,381.806	14	4,791.611	4,853.74	4,839.46	0.002
4	625	−2,370.384	18	4,776.767	4,856.647	4,842.36	0.000
5	625	−1,968.392	22	3,980.783	4,078.414	4,064.13	0.000
6	625	−1,962.529	26	3,977.058	4,092.44	4,078.16	0.001

adding more classes beyond five could overcomplicate the model and hinder meaningful interpretation of each group's characteristics.

We then assigned each respondent to one of the five classes based on their highest latent class probability. The results revealed five groups: (a) the Cautious ($n = 92$, 14.7%) had a moderate relationship with their service providers ($M = 2.47$) and moderate recognition of technology issues ($M = 3.26$), but they reported low trust in AI ($M = 1.86$); (b) the Balanced ($n = 218$, 34.8%) showed a good relationship with their service providers ($M = 3.43$), moderate recognition in technological problems ($M = 2.57$), and a moderate level of trust in AI systems ($M = 3$); (c) the Uninterested ($n = 6$, 1%) had low recognition of technology issues ($M = 1.88$), poor relationships with providers ($M = 1.12$), but moderate trust in AI ($M = 3$); (d) the Confident ($n = 259$, 41.4%) reported high trust in AI ($M = 4$), good relationships with providers ($M = 3.45$), and moderate recognition of technology issues ($M = 2.72$); and (e) the Enthusiastic ($n = 50$, 8%) showed very high trust in AI ($M = 5$), strong relationships with their providers ($M = 3.74$), and above-average interest in technology-related issues ($M = 2.89$).

We created these group names solely for interpretive convenience. They do not stereotype or oversimplify individuals in each class. For example, we consider someone "Cautious" only in the specific context of AI adoption or trust, not in broader terms.

See Figure 1 for a visual breakdown of these groups.

Among these groups, the Enthusiastic had the highest average scores for AI service adoption ($M = 3.71$) and trust ($M = 3.33$). In contrast, the Uninterested group showed the lowest scores for adoption ($M = 2.48$) and trust ($M = 1.94$). When comparing these two groups at opposite ends of the spectrum, the Uninterested group had a younger average age (24.5 years) than the Enthusiastic group (32.6 years). Additionally, the Enthusiastic group consisted mostly of men (70%) than women, while the Uninterested group had an equal

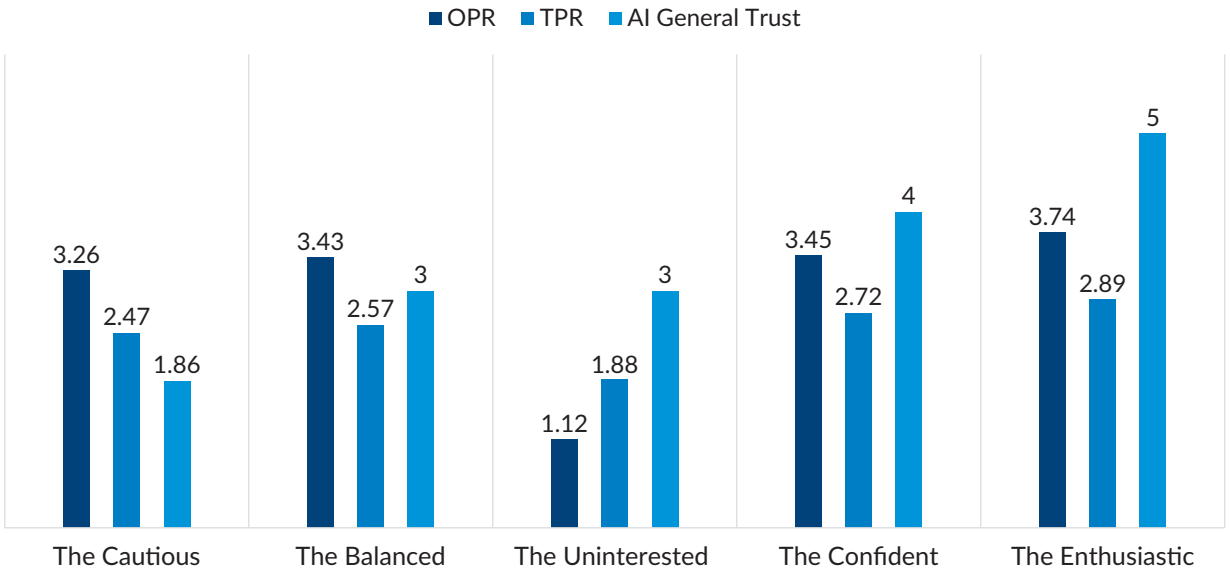


Figure 1. Profiles of OPR, TPR, and AI general trust (on a 5-Likert scale).

gender distribution. Aside from these demographic differences, we found no significant differences in other factors such as household income, education level, or place of residence (RQ1 and RQ2).

Table 3 presents average scores across the five public groups' profiles—the Cautious, Balanced, Uninterested, Confident, and Enthusiastic—based on key variables related to AI adoption.

Table 3. Mean values of variables by profiles.

	Cautious	Balanced	Uninterested	Confident	Enthusiastic
OPR	3.26	3.43	1.12	3.45	3.74
TPR	2.47	2.57	1.88	2.72	2.89
General Trust in AI	1.86	3	3	4	5
AI Service Adoption Intention	2.48	2.87	1.94	3.12	3.37
Trust toward AI Services	2.87	3.11	2.48	3.47	3.71
N	92	218	6	259	50
%	14.7	34.8	1	41.4	8

5.2. Regression Analysis

We tested H1 to H3 using regression analysis to examine the relationships among the key variables. The results show that OPR significantly increased AI service adoption intention ($\beta = 0.44$, $p < 0.01$), supporting H1a. However, OPR did not significantly influence trust in AI services ($\beta = 0.17$, not significant), rejecting H2a.

TPR had no significant effect on service adoption or trust (H1b and H2b). However, when we excluded interaction terms from the model, TPR showed a positive significant effect on AI service adoption ($\beta = 0.28$, $p < 0.001$) and trust in AI services ($\beta = 0.15$, $p < 0.001$).

In contrast, general trust in AI consistently showed stronger effects across both outcomes, positively influencing adoption ($\beta = 0.41$, $p < 0.05$) and trust in services ($\beta = 0.60$, $p < 0.01$).

We also found that those with higher household incomes and younger respondents were more willing to adopt AI services. These findings indicate that general trust in AI plays a more powerful role than situational factors like TPR or the quality of the user–provider relationship when it comes to trusting specific AI services. However, a strong, well-managed relationship with service providers can play a key role in motivating users to try out new technologies.

Table 4 examines the regression results of the predictors of AI service adoption and trust in AI services.

H3 expected interaction effects among the three predictors. We found a significant interaction effect only for AI service adoption ($\beta = -0.44$, $p < 0.05$). The AI adoption probability increased for individuals with lower general trust in AI as their relationship quality with the service provider improved (see Figure 2). This finding suggests that strong relationships with existing customers can play a crucial role in encouraging the adoption of new technologies, even among those who are generally skeptical of AI.

Table 4. Regression analysis results.

	AI Service Adoption		AI Service Trust	
OPR	0.097**	0.438**	−0.001	0.173
	(2.60)	(2.63)	(−0.02)	(1.04)
TPR	0.272***	0.183	0.148***	0.268
	(7.16)	(1)	(3.90)	(1.47)
General Trust in AI (Trust)	0.246***	0.411*	0.352***	0.604***
	(6.58)	(2.29)	(9.46)	(3.38)
Gender (Female = 1)	0.033	0.033	−0.115**	−0.115**
	(0.88)	(0.88)	(−3.07)	(−3.07)
Education Level (University = 1)	−0.043	−0.046	0.003	0.006
	(−1.14)	(−1.23)	(0.08)	(0.16)
Household income	0.095*	0.097**	−0.007	−0.008
	(2.55)	(2.60)	(−0.19)	(−0.21)
Age group	−0.080*	−0.081*	0.002	0.002
	(−2.12)	(−2.17)	(0.07)	(0.06)
Duration of membership	−0.06	−0.048	0.029	0.037
	(−1.62)	(−1.29)	(0.79)	(0.98)
OPR × TPR		−0.012		−0.019
		(−0.71)		(−0.11)
TPR × Trust		0.255		−0.138
		(1.41)		(−0.76)
OPR × Trust		−0.442*		−0.260
		(−1.97)		(−1.16)
N	625	625	625	625
Adjusted R^2	0.171	0.175	0.179	0.178

Notes: Standardized beta coefficients; *t* statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

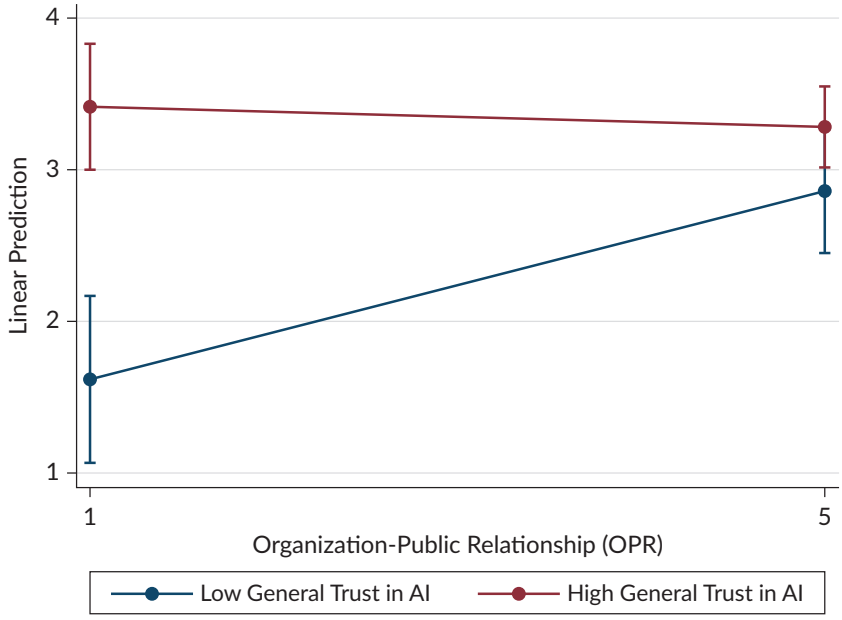


Figure 2. Interaction relationship plots of regression results: AI service adoption.

5.3. SEM

H4 proposed that trust in AI services mediates the relationship between two predictors—OPR and TPR—and AI service adoption. Prior literature emphasized the importance of established trust in AI, suggesting the value of testing its mediating role in this study. We included control variables such as service use duration, age, and income levels to account for their potential influence.

The second-order SEM model (see Figure 3) showed a good fit: $\chi^2(224) = 623.059$, $p < 0.001$, RMSEA = 0.047 [0.042, 0.051], CFI = 0.936, TLI = 0.928, SRMR = 0.066. The results support H4a, showing that TPR indirectly influenced AI service adoption through trust in AI services (TPR—Trust in AI Services: $\beta = 0.23$, $p < 0.001$; Trust in AI Services—AI Service Adoption: $\beta = 0.46$, $p < 0.001$). However, OPR did not significantly influence trust or adoption in this model.

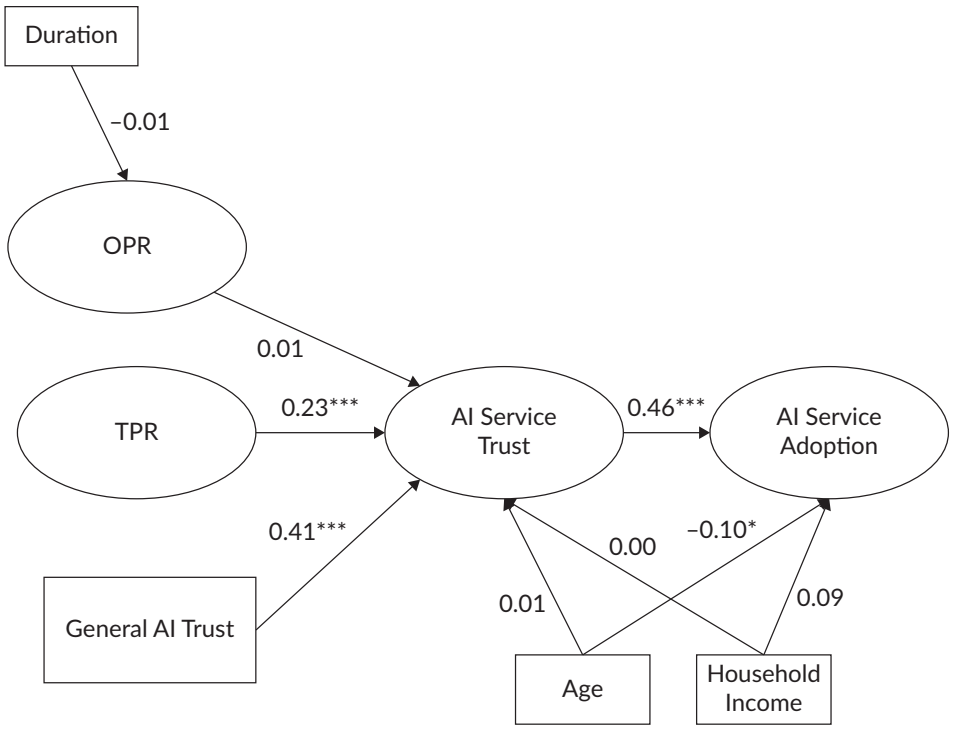


Figure 3. SEM results. Notes: *** $p < 0.001$, * $p < 0.05$; $\chi^2(224) = 623.059$, $p < 0.001$; RMSEA = 0.047 [0.042, 0.051]; CFI = 0.936; TLI = 0.928; SRMR = 0.066.

We also found support for H4b, as general trust in AI significantly influenced adoption intention through its effect on trust in AI services (General Trust—Trust in AI Services: $\beta = 0.41$, $p < 0.001$). All these findings suggest that trust in AI services plays a key mediating role, especially in translating problem recognition and general trust into actual adoption behavior.

6. Discussion

The study introduced a new approach to identifying AI technology adopters by integrating public relations concepts. The study achieved this by using LPA, regression analysis, and SEM. The results identified five public segments' profiles—the Cautious, Balanced, Uninterested, Confident, and Enthusiastic—based on their levels

of OPR, recognition of technology problems, and general trust in AI. Among them, the Enthusiastic group stood out, showing the highest levels of AI adoption intent, high trust in AI, and strong relationships with their service providers. This finding suggests that they could serve as the key public group for businesses launching new AI services.

The regression analysis, which tested three hypotheses, revealed that a strong relationship with the service provider positively influenced adoption intention. However, this relationship did not significantly impact trust in specific AI services. Demographic factors such as age and income affected adoption, while education level did not, raising questions about how cognitive factors like education influence AI adoption. We also observed interaction effects: Customers with low general trust in AI were more likely to adopt AI services when they had strong relationships with providers. This finding highlights how relationship-building can drive adoption, even among skeptical users. Strong relational ties can help organizations reduce fear around emerging technologies and shape effective business strategies.

The SEM analysis further revealed that trust in AI services plays a mediating role. Specifically, problem recognition and general trust in AI technologies significantly influenced AI service adoption intention through their effects on service trust. In other words, when people recognize technology-related issues and already trust AI in general, they are more likely to adopt AI services because they trust the service provider.

6.1. Implications for Theory and Practice

The study offers several contributions to theory and practice. First, it expands the application of public relations theory, particularly concepts from OPR and the STOPS, to the field of technology adoption. Unlike conventional approaches, such as the diffusion of innovations theory, which focus on innovators and early adopters with high knowledge or income (e.g., Tran et al., 2024; Uzumcu & Acilmis, 2024), this study highlights the importance of relational factors that apply to a broader range of customers. Communication managers can use these insights to identify strategic publics through relationship metrics, not just demographic or tech-savvy traits.

Second, the findings emphasize the power of long-term relationships. While new service development or market introduction often focuses on new market segments, public relations scholarship stresses the long-term benefits of sustained relationship management. Strong relationships can significantly increase the adoption rate of new services, especially in markets where negative public opinion or skepticism toward AI is prevalent. However, in challenging times such as crises, loyal customers who value their relationship with a company can act as brand advocates, defending the service and its benefits.

Third, focusing on relational factors allows organizations to develop specific adoption strategies. For example, public relations research on OPR has identified several ways to strengthen relationships, such as ethical communication and two-way symmetrical communication. By combining these relationship management strategies with effective marketing and sales efforts, businesses can enhance adoption and increase market share. These strategies are not just for technological companies. Industries like pharmaceuticals, dining services, and retail can also benefit. For instance, a pharmaceutical company launching a new cosmetics line or a restaurant introducing a tech-enabled service system can improve success by understanding and leveraging the relationships they have already built with customers. These

relationships are especially important when entering foreign markets, where connected relationships can facilitate the adoption of unfamiliar products or services.

6.2. Limitations and Future Directions

Despite its contributions, the study has some limitations. The most notable is its sample, which includes only Korean participants. Although we controlled for demographic variables, the results may not generalize to other cultures. For example, Koreans may be more open to rapid technological changes compared to consumers in more conservative or risk-averse societies. We recommend future cross-national studies to test external validity in diverse cultural contexts and identify early adopters.

Additionally, while the study employed methods such as LPA, ordinary least squares, and SEM, the one-time survey design limits causal interpretations. The online survey format may also introduce self-selection bias and representativeness issues, which could affect data quality. Therefore, future research should incorporate diverse methods, such as computational analysis, focus groups, in-depth interviews, and experiments, to gain a deeper understanding of how relational and situational recognition factors influence trust in services and adoption behavior. Qualitative methods, in particular, could provide deeper insights into the mechanisms driving these relationships.

Finally, the study relied on OPR and the STOPS. Future research could capture the full scope of relational and situational dynamics by incorporating additional public relations concepts, such as relationship cultivation strategies, communal vs. transactional relationships, and other STOPS variables like involvement recognition, constraint recognition, referent criterion, and situational motivation. Understanding how people perceive tech-related issues could further clarify how they adopt new services.

Beyond academic implications, we must acknowledge the broader social and political uncertainty surrounding AI. Businesses and governments alike face immense pressure as they attempt to regulate and manage AI technologies. Experts continue to debate how best to govern AI at national and international global levels (e.g., UN Global Compact, 2024). As AI begins to reshape market economies, geopolitical competition will increase. For instance, China's DeepSeek has faced allegations of copying AI technologies developed by OpenAI in the US (e.g., Criddle & Olcott, 2025). Meanwhile, the Trump administration 2.0 has reinforced its US-centered and market-driven strategies to maintain American dominance in the AI industry (The White House, 2025).

In this unstable and rapidly evolving context, it is more important than ever to understand how users perceive and adopt AI services. The topic needs more active discussions about how AI will affect users and the relevant strategies for engaging, informing, and building trust with the people who use it.

7. Conclusion

The contemporary era has entered a new phase where society and AI are co-evolving, with technological development expanding at an unprecedented pace. For business leaders and policymakers, this transformation presents challenges, especially in understanding how the public and stakeholders might respond to the ongoing shifts that AI has created. Therefore, it is essential to explore various approaches to understand how the general population adopts AI services.

This study offers novel methods by applying public relations concepts to identify and understand potential adopters of AI technologies. However, while the findings provide meaningful insights, the study also acknowledges its limitations, particularly the limited scope of business sectors and nationalities, which may affect the generalizability of the results. Nonetheless, the study highlights the research potential of various methods for distinguishing potential adopters, contributing to more effective strategies for promoting the adoption of new AI services.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material (appendix) for this article is available online in the format provided by the author (unedited).

References

Ahmad, S., Zulkurnain, N. N. A., & Khairushalimi, F. I. (2016). Assessing the validity and reliability of a measurement model in structural equation modeling (SEM). *British Journal of Mathematics & Computer Science*, *15*(3), 1–8. https://doi.org/10.9734/BJMCS/2016/25183

Ameen, N., Tarhini, A., Reppel, A., & Anand, A. (2021). Customer experiences in the age of artificial intelligence. *Computers in Human Behavior*, *114*, Article 106548. https://doi.org/10.1016/j.chb.2020.106548

Balkrishan, D. K., & Joshi, A. (2013). Technology adoption by 'emergent' users: The user-usage model. In S. Tripathi & A. Joshi (Eds.), *APCHI '13: Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction* (pp. 28–38). ACM. https://doi.org/10.1145/2525194.2525209

Battista, D. (2024). The road to AI: Pathways and obstacles. *Societes*, *163*(1), 55–72. https://doi.org/10.3917/soc.163.0055

Bianchi, M., Di Benedetto, A., Franzò, S., & Frattini, F. (2017). Selecting early adopters to foster the diffusion of innovations in industrial markets: Evidence from a multiple case study. *European Journal of Innovation Management*, *20*(4), 620–644. https://doi.org/10.1108/EJIM-07-2016-0068

Cheng, Y. (2018). Looking back, moving forward: A review and reflection of the organization-public relationship (OPR) research. *Public Relations Review*, *44*(1), 120–130. https://doi.org/10.1016/j.pubrev.2017.10.003

Constantz, J., & Bloomberg. (2024, February 9). Over 4,000 workers have lost their jobs to AI since May, outplacement firm estimates—and that's 'certainly undercounting.' *Fortune*. https://fortune.com/2024/02/08/how-many-workers-laid-off-because-of-ai

Contzen, N., Handreke, A. V., Perlaviciute, G., & Steg, L. (2021). Emotions towards a mandatory adoption of renewable energy innovations: The role of psychological reactance and egoistic and biospheric values. *Energy Research & Social Science*, *80*, Article 102232. https://doi.org/10.1016/j.erss.2021.102232

Criddle, C., & Olcott, E. (2025, January 29). OpenAI says it has evidence China's DeepSeek used its model to train competitors. *Financial Times*. https://www.ft.com/content/a0dfedd1-5255-4fa9-8ccc-1fe01de87ea6

Davies, M., & Birtwistle, M. (2023). *Regulating AI in the UK*. The Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Debus, M. E., Ingold, P. V., Gross, C., & Bolino, M. C. (2024). Reaching the top? Profiles of impression

management and career success. *Journal of Business and Psychology*, *39*, 1283–1301. https://doi.org/10.1007/s10869-024-09954-7

Eichhorn, B. R. (2014). *Common method variance techniques*. SAS Institute.

Farina, M., & Lavazza, A. (2023). ChatGPT in society: Emerging issues. *Frontiers in Artificial Intelligence*, *6*, Article 1130913. https://doi.org/10.3389/frai.2023.1130913

Feng, W., Tu, R., Lu, T., & Zhou, Z. (2019). Understanding forced adoption of self-service technology: The impacts of users' psychological reactance. *Behaviour & Information Technology*, *38*(8), 820–832. https://doi.org/10.1080/0144929X.2018.1557745

Frank, D. A., Chrysochou, P., & Mitkidis, P. (2023). The paradox of technology: Negativity bias in consumer adoption of innovative technologies. *Psychology & Marketing*, *40*(3), 554–566. https://doi.org/10.1002/mar.21740

Frattini, F., Bianchi, M., De Massis, A., & Sikimic, U. (2014). The role of early adopters in the diffusion of new products: Differences between platform and nonplatform innovations. *Journal of Product Innovation Management*, *31*(3), 466–488. https://doi.org/10.1111/jpim.12108

Gerlich, M. (2023). Perceptions and acceptance of artificial intelligence: A multi-dimensional study. *Social Sciences*, *12*(9), Article 502. https://doi.org/10.3390/socsci12090502

Gerlich, M. (2024a). Public anxieties about AI: Implications for corporate strategy and societal impact. *Administrative Sciences*, *14*(11), Article 288. https://doi.org/10.3390/admsci14110288

Gerlich, M. (2024b). Societal perceptions and acceptance of virtual humans: Trust and ethics across different contexts. *Social Sciences*, *13*(10), Article 516. https://doi.org/10.3390/socsci13100516

Grunig, J. E. (1997). A situational theory of publics: Conceptual history, recent challenges and new research. In D. Moss, T. MacManus, & D. Verčić (Eds.), *Public relations research: An international perspective* (pp. 3–46). ITB Press.

Grunig, J. E., & Hunt, T. (1984). *Managing public relations*. Hilt, Rinehart and Winston.

Hon, L. C. (1997). What have you done for me lately? Exploring effectiveness in public relations. *Journal of Public Relations Research*, *9*(1), 1–30.

Hon, L. C., & Grunig, J. E. (1999). *Guidelines for measuring relationships in public relations*. Institution of Public Relations.

Huang, Y. H. (2001a). OPRA: A cross-cultural, multiple-item scale for measuring organization-public relationships. *Journal of Public Relations Research*, *13*(1), 61–90. https://doi.org/10.1207/S1532754XJPRR1301_4

Huang, Y. H. (2001b). Values of public relations: Effects on organization–public relationships mediating conflict resolution. *Journal of Public Relations Research*, *13*(4), 265–301. https://doi.org/10.1207/S1532754XJPRR1304_01

Hunt, T. (2023, May 25). Here's why AI may be extremely dangerous—Whether it's conscious or not. *Scientific American*. https://www.scientificamerican.com/article/heres-why-ai-may-be-extremely-dangerous-whether-its-conscious-or-not

Kazoleas, D., & Wright, A. (2001). Improving corporate and organizational communication: A new look at developing and implementing the communication audit. In R. L. Heath (Ed.), *Handbook of public relations* (pp. 471–478). Sage.

Kelly, M. S. (2024, May 23). Elon Musk says AI will take all our jobs. *CNN*. https://edition.cnn.com/2024/05/23/tech/elon-musk-ai-your-job/index.html

Kim, J.-N., & Grunig, J. E. (2011). Problem solving and communicative action: A situational theory of problem solving. *Journal of Communication*, *61*(1), 120–149. https://doi.org/10.1111/j.1460-2466.2010.01529.x

Langer, M., König, C. J., Back, C., & Hemsing, V. (2023). Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, *38*(3), 493–508. https://doi.org/10.1007/s10869-022-09829-9

Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 880–918. https://doi.org/10.17705/1jais.00411

Mahardika, H., Thomas, D., Ewing, M. T., & Japutra, A. (2019). Experience and facilitating conditions as impediments to consumers' new technology adoption. *The International Review of Retail, Distribution and Consumer Research*, *29*(1), 79–98. https://doi.org/10.1080/09593969.2018.1556181

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/10.5465/amr.1995.9508080335

Mbiazi, D., Bhange, M., Babaei, M., Sheth, I., & Kenfack, P. J. (2023). *Survey on AI ethics: A socio-technical perspective*. arXiv. https://doi.org/10.48550/arXiv.2311.17228

Nagy, P., & Neff, G. (2024). Conjuring algorithms: Understanding the tech industry as stage magicians. *New Media & Society*, *26*(9), 4938–4954. https://doi.org/10.1177/14614448241251789

Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., & Agreda, S. G. (2022). To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change*, *181*, Article 121763. https://doi.org/10.1016/j.techfore.2022.121763

Orben, A. (2020). The Sisyphean cycle of technology panics. *Perspectives on Psychological Science*, *15*(5), 1143–1157. https://doi.org/10.1177/1745691620919372

Ortt, R., Dedehayir, O., Miralles, F., & Riverola, C. (2017). Innovators and early adopters in the diffusion of innovations: A literature review. In E. Huizingh (Ed.), *ISPIM Conference Proceedings* (pp. 1–16). The International Society for Professional Innovation Management (ISPIM).

Panopoulos, A., Theodoridis, P., & Poulis, A. (2018). Revisiting innovation adoption theory through electronic public relations. *Information Technology & People*, *31*(1), 21–40. https://doi.org/10.1108/ITP-05-2016-0101

Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology*, *13*, 53–64. https://doi.org/10.1007/s10676-010-9253-3

Prins, R., & Verhoef, P. C. (2007). Marketing communication drivers of adoption timing of a new e-service among existing customers. *Journal of Marketing*, *71*(2), 169–183. https://doi.org/10.1509/jmkg.71.2.169

Rogers, M. E. (2003). *Diffusion of innovations* (5th ed.). Free Press.

Roose, K. (2023, May 30). A.I. poses 'risk of extinction,' industry leaders warn. *The New York Times*. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

Son, M., & Han, K. (2011). Beyond the technology adoption: Technology readiness effects on post-adoption behavior. *Journal of Business Research*, *64*(11), 1178–1182. https://doi.org/10.1016/j.jbusres.2011.06.019

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, Article 103445. https://doi.org/10.1016/j.jvb.2020.103445

The White House. (2025). Removing barriers To American leadership in artificial intelligence. *The White House*. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence

Tran, T. H. U., Lau, K. H., & Ong, C. E. (2024). Social sustainability practice innovation diffusion and its relationship to organizational improvement: A mechanism for Vietnamese handicraft companies. *Asia Pacific Journal of Management*. Advance online publication. https://doi.org/10.1007/s10490-024-09953-5

UN Global Compact. (2024). *Gen AI for the global goals report*. https://info.unglobalcompact.org/gen-ai-for-global-goals#form_001

Uzumcu, O., & Acilmis, H. (2024). Do innovative teachers use AI-powered tools more interactively? A study in the context of diffusion of innovation theory. *Technology, Knowledge and Learning*, *29*(2), 1109–1128. https://doi.org/10.1007/s10758-023-09687-1

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Yang, S.-U. (2007). An integrated model for organization—Public relational outcomes, organizational reputation, and their antecedents. *Journal of Public Relations Research*, *19*(2), 91–121. https://doi.org/10.1080/10627260701290612

About the Authors

Hyelim Lee is an assistant professor of strategic communication at Korea University's College of Media and Communication. She earned her PhD from the University of Oklahoma, specializing in computational public relations, integrating public relations theories with data analytics. Her research explores data-driven approaches and AI methods in communication strategies.

Chanyoung Jung earned a BS in mathematics in 2022 and an MS in computer engineering and convergence engineering for intelligent drones, both at Sejong University, South Korea. Since 2024, he has been working as a researcher employed by a data science group called INTERX.

Nayoung Koo is currently studying computer engineering at Sejong University in Seoul, South Korea. Her main research focus is data visualization.

Seongbum Seo is a researcher at the Data Visualization Lab at Sejong University in Seoul, South Korea. He received his BS (2023) and MS (2025) in computer engineering from Sejong University. His research focuses on natural language processing and data visualization.

Sangbong Yoo received a BS and PhD in computer engineering from Sejong University, Seoul, South Korea, in 2015 and 2022, respectively. He was a postdoctoral researcher at Sejong University and is now a postdoctoral researcher at the Korea Institute of Science and Technology (KIST). His research interests include information visualization, visual analytics, eye-gaze analysis, data quality, and dimensionality reduction.

Hyein Hong received a BS and an MS in computer engineering from Sejong University, Seoul, South Korea, in 2021 and 2023, respectively. She is currently a data analyst at Miridih, South Korea. Her research interests include recommendation models, visual analytics, and data quality.

Yun Jang received a BA from Seoul National University (2000), and a MS and PhD from Purdue University (2002, 2007). After postdoctoral research at Centro Svizzero di Calcolo Scientifico (CSCS) and ETH Zürich (2007–2011), he became a professor of computer engineering at Sejong University. His research focuses on interactive visualization and volume rendering.

Media and Communication
2025 • Volume 13 • Article 9508
https://doi.org/10.17645/mac.9508

ARTICLE

Open Access Journal

Motivations and Affordances of ChatGPT Usage for College Students' Learning

Sun Kyong Lee, Jongsang Ryu, Yeowon Jie, and Dong Hoon Ma

College of Media and Communication, Korea University, South Korea

Correspondence: Sun Kyong Lee (sunnylee@korea.ac.kr)

Submitted: 28 October 2024 **Accepted:** 5 February 2025 **Published:** 3 April 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

This study explored college students' experiences and evaluations of using ChatGPT for class-related activities including essay writing, exam preparation, and homework. Students from two classes on the same subject were surveyed, and quantitative data on their motivations and usage of ChatGPT were collected (Class 1, $n = 48$; Class 2, $n = 106$; $N = 154$). Hierarchical regression analysis revealed that using ChatGPT as a study guide and for active interaction were significant predictors of actual usage level, while its usage for entertainment and study guide was associated with higher trust in the tool. We further collected qualitative data through open-ended surveys (Class 1, $n = 154$; Class 2, $n = 106$). Responses were manually coded and thematically analyzed, with comparisons drawn between the two classes. Students' perceptions varied, with many acknowledging the affordances of ChatGPT, such as helping to organize thoughts, clarifying concepts, and structuring essays. However, some participants raised concerns about the tool's limitations—particularly its potential to inhibit critical and creative thinking—as well as issues related to the reliability, accuracy, and quality of information provided. The implications of these findings are discussed in relation to the uses and gratifications theory, the technology acceptance model, and the concept of media affordances.

Keywords

affordances; ChatGPT; ChatGPT usage; generative AI; higher education; motivation

1. Introduction

Concerns have emerged regarding the role of generative AI, such as ChatGPT, with the potential to replace human labor, including educators in higher education (Jensen et al., 2024). Predictions of singularity—the idea that machines may surpass humans in cognition—raise questions about the future roles of teachers

(Bostrom, 2014). However, technological shifts in education tend to evolve incrementally. Historically, oral communication dominated knowledge transmission until the 16th century, although writing existed before the printing press (McLuhan, 1962; Ong, 1982). The Gutenberg press catalyzed a shift toward literacy, but oral traditions remained integral to learning (Eisenstein, 1979). In the 19th and 20th centuries, films and television were introduced, adding visuality as the primary mode of communication (Rheingold, 2000). Yet, classrooms continued to rely on oral and written methods, illustrating the persistence of older forms alongside newer ones.

The late 20th century introduced interactive technologies through personal computers; however, oral, and literacy-based learning remained significant (Turkle, 2015). In the 21st century, despite the proliferation of AI and virtual platforms, older forms of communication remain essential (Castells, 2000). This period, often referred to as the "age of real virtuality," highlights the interplay of various media, particularly as society seeks authentic communication during times of isolation (Turkle, 2011). Amid these digital advances, Benjamin's (1935/1968) notion of "aura" has regained relevance, underscoring the value of in-person interaction.

Generative AI refers to a category of AI that can create new content such as text, images, or music by learning patterns from existing datasets. These models leverage deep learning techniques—particularly large language models—to generate human-like outputs based on user prompts (Kar et al., 2023). Generative AI tools such as Microsoft's new Bing, Google's Gemini, and OpenAI's ChatGPT aim to integrate human-like communication through orality, literacy, and visuality (Abdul-Kader & John, 2015). These technologies show significant potential in education through adaptive feedback and personalized interaction, aligning with constructivist and social constructivist pedagogies (Piaget, 1971; Vygotsky, 1978). By automating administrative tasks, AI also supports teachers, enabling them to focus on more meaningful engagement with their students (Gamage et al., 2022). However, the rise of AI presents ethical challenges, including concerns about data privacy, algorithmic bias, and the risk of depersonalizing learning experiences (Floridi, 2021).

The thoughtful implementation of AI tools such as ChatGPT and research examining its impact on learning outcomes, user engagement, as well as its social and ethical impact on education, can enable educators to harness their potential to create more inclusive, adaptive, and engaging educational experiences while preserving the rich traditions of human interaction and communication fundamental to learning.

South Korea is an exemplary site for studying technology adoption, including AI tools such as ChatGPT, owing to its advanced digital infrastructure, proactive government strategies, and tech-savvy population. The country boasts some of the world's fastest internet speeds and extensive 5G coverage, facilitating a seamless integration of digital tools into daily life (OECD, 2023). The South Korean government has implemented comprehensive AI strategies such as the National Strategy for AI, aiming to position the nation as a global leader in AI by 2030 (Ministry of Science and ICT, 2019). Additionally, South Korea's emphasis on digital education and innovation fosters a culture that readily embraces new technologies, rendering it an ideal environment for observing and analyzing the dynamics of technology adoption and usage.

Against this background, this study explored Korean college students' experiences and evaluations of using ChatGPT for class-related activities. The research aimed to understand their motivations for its usage, actual usage levels, and how they are related to students' trust in generative AI technology. This study adopted

a mixed-method approach. Section 2 explains the theoretical background along with a review of existing literature and suggests research questions.

2. Literature Review and Theoretical Background

2.1. ChatGPT as a Learning Tool

The integration of ChatGPT—a generative AI product developed by OpenAI—into learning environments presents both advantages and challenges (Rasul et al., 2023). A comprehensive review of prior studies has revealed the following advantages: First, adaptive learning provides personalized learning experiences by tailoring feedback and resources based on student progress (Kerr, 2016). Through its interactive capabilities, ChatGPT offers adaptive learning experiences by personalizing content delivery and feedback (Rudolph et al., 2023). Individualized feedback is considered an important factor in learning as it supports students' specific needs, enhancing comprehension and performance (Nicol & Macfarlane-Dick, 2006). ChatGPT's capacity to deliver such feedback supports constructivist learning by enabling personalized guidance (Rudolph et al., 2023). This product of OpenAI also assists in literature reviews, summarizing research, and initial drafting of papers (Dwivedi et al., 2023), thus enabling students to manage information more efficiently within a small timeframe and enhancing productivity.

Furthermore, automated tools such as ChatGPT streamline tasks such as progress tracking, reminders, and academic feedback, enhancing learning efficiency and reducing administrative burdens (June et al., 2014; Zhao et al., 2022). Finally, ChatGPT facilitates innovative assessments by generating unique questions and case studies, and fostering critical thinking, creativity, collaboration, and real-time feedback, ultimately improving knowledge acquisition (Boud & Soler, 2016; Kumar, 2021).

The key challenges in using ChatGPT in higher education include issues related to academic integrity, reliability, skill assessment, learning outcomes, and misinformation. Concerns about academic integrity arise from the potential for plagiarism and contract cheating, as the ease of generating content through ChatGPT conflicts with the constructivist emphasis on active learning (Cotton et al., 2024). Reliability is another significant issue, as large language models such as ChatGPT can produce biased or inaccurate information owing to limitations in their training data, which may impede the development of critical thinking skills (Chen et al., 2023). Moreover, ChatGPT's inability to evaluate essential skills—such as leadership and problem-solving—presents a challenge, as these competencies are typically developed through experiential learning rather than automated processes (Atlas, 2023). The passive nature of AI-driven assessments further limits their effectiveness in measuring learning outcomes as they often fail to foster deep learning and meaningful engagement (Biggs, 2014). Finally, misinformation remains a critical concern, as ChatGPT can generate misleading outputs based on skewed datasets or fabricate references, exacerbating this issue (Hsu & Thompson, 2023).

2.2. Motivations for Technology Adoption in Learning

The uses and gratifications theory provides a useful framework for examining technology users' diverse motivations and satisfaction with their usage of new and emerging technologies. While Katz et al. (1973) laid the groundwork for identifying basic social and psychological needs, and with media being one of the major

sources for gratifying such needs, many subsequent scholars have discovered similar and distinct motivations for emerging media of the time, such as telephones, mobile phones, the internet, social media, and now chatbots. Some may view these technologies as eliciting new and different motivations that did not exist in previous times (or media), emerging mainly as responses to the unique characteristics of the latter media, while others may view them from the perspective of media affordances, defined as action possibilities and constraints perceived and/or actualized by media users based on their interactions with the media (Gibson, 1986; Norman, 1999).

Research on the motivations for using technology in learning has revealed key findings in the pre-AI era. Studies have shown that social influence, trust, and hedonic motivation significantly affect students' behavioral intentions to adopt technology (Holmes et al., 2021). Social influence pertains to the effects of peers, instructors, and the community on an individual's decision to use e-learning platforms. Trust involves confidence in the reliability and security of e-learning systems, which is crucial for user acceptance. Hedonic motivation refers to the enjoyment and pleasure derived from using a technology, which further drives its adoption (Holmes et al., 2021). These factors collectively contribute to a comprehensive understanding of the determinants influencing students' intentions to engage with e-learning technologies (Tarhini et al., 2017).

Investigating the motivations behind the adoption of ChatGPT in higher education is crucial for understanding its integration into academic settings. Studies have identified factors such as performance expectancy, effort expectancy, and hedonic motivation as significant predictors of educators' intentions to use ChatGPT. Additionally, research incorporating the technology acceptance model (TAM; Davis, 1989) and self-determination theory (Deci & Ryan, 1985) has highlighted the roles of trust, social influence, and personal innovativeness in shaping students' adoption of ChatGPT. Understanding these motivations can inform the development of effective integration strategies and policies in higher education.

Therefore, based on Park's (2010) integrated model of uses and gratifications theory (Katz et al., 1973) and the TAM (Davis, 1989), this study also examines whether students use ChatGPT with similar motivations as those for using other types of chatbots or media, or with differing motivations because the technology itself has new and different aspects compared to preceding ones. Park (2010) studied the adoption and usage of voice over internet protocol phone service and identified three dimensions of motivation (namely communication, instrumental, and entertainment) from his online survey data. His findings showed that motives for communication were positively associated with perceived ease of use (PEOU) and perceived usefulness (PU), entertainment motives were negatively associated with PEOU, and instrumental motives were positively associated with PU and actual usage of voice over internet protocol services. In Park's (2010) study, only instrumental motives were directly related to actual usage, whereas the other motives had indirect effects.

As the adoption of various forms of generative AI including ChatGPT is still in its early stages, it would be fruitful to examine motivations for college students' use of ChatGPT and how their motivations and PEOU and PU of ChatGPT are related to its actual usage level. Given these observations, we developed the following research questions (RQs):

> RQ1: (a) What are college students' motivations for using ChatGPT and (b) how do such motivations relate to its actual usage levels?

RQ2: How do college students' perceived (a) ease of use and (b) usefulness of ChatGPT relate to its actual usage levels?

2.3. Trust in AI

User experiences with generative AI are critical for determining how they build and sustain trust in these systems (S. K. Lee & Sun, 2023). Trust is not merely an outcome of initial interactions but develops over time through continuous use and is highly influenced by how well the AI system meets users' diverse motivations and expectations. If users encounter positive experiences, such as satisfaction with using the AI system to fulfill educational or personal goals, they are likely to view the system as reliable. This perceived trust can foster long-term engagement, encouraging users to rely on the system in various contexts beyond their initial use. By contrast, if users find their interactions with the model frustrating, or if it fails to meet their expectations, they may not only discontinue its use but also develop skepticism towards future AI innovations, affecting overall trust in AI technologies (S. K. Lee et al., 2021).

Trust in AI systems—including generative AI—plays a key role in how users integrate such technologies into their daily routines and decision-making processes. Numerous studies have demonstrated that human–machine trust is crucial in establishing effective interactions with virtual AI assistants or robots, allowing users to engage with these systems more confidently and comfortably (S. K. Lee et al., 2021; S. K. Lee & Sun, 2023). Trust in AI thus acts as a mediator between the initial motivation to use the technology and the sustained engagement necessary for educational success, problem-solving, and the use of other types of applications. When trust is established, users are more likely to expand their usage to more complex tasks, whereas a lack of trust may limit their engagement to surface-level interactions.

Moreover, in the context of higher education, where critical thinking and informed decision-making are central, trust in AI is particularly important. Students may rely on generative AI tools such as ChatGPT not only for basic information retrieval but also for developing ideas, structuring research, or enhancing creativity (Abramson, 2023; Baidoo-Anu & Ansah, 2023). A high level of trust in these systems can lead to more meaningful educational outcomes, whereas distrust may hinder the potential benefits of AI-assisted learning. Therefore, understanding how students' motivations for using ChatGPT align with their level of trust in the system is essential to improving AI integration in educational environments. Through this study, we aimed to explore the relationship between students' motivations, their actual use of ChatGPT in higher education, and how these factors collectively influence their trust in the generative AI system. Figure 1 presents the conceptual model used in this study from a quantitative perspective. With this, we propose:

RQ3: How do college students (a) motivations and (b) ChatGPT usage levels relate to their trust in the generative AI system?

2.4. Uses of AI-Chatbots in Education and Their Affordances

Kuhail et al. (2023) analyzed 36 educational chatbots by evaluating them within seven dimensions: educational field, platform, educational role, interaction style, design principles, empirical principles, and challenges/limitations. The results showed that chatbots were proposed mainly in computer science, language, general education, and a few other fields and were accessible mostly via web platforms.

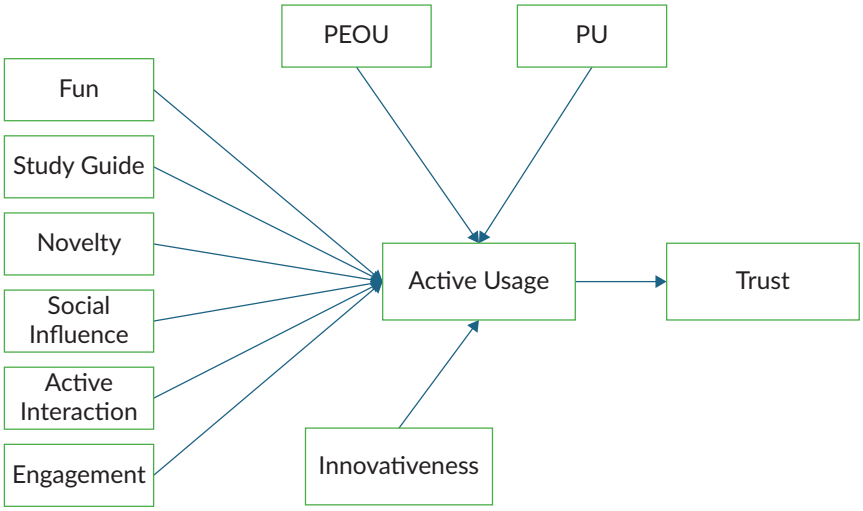


Figure 1. Conceptual model of the quantitative research.

The chatbots reflected differing educational roles, interaction styles, and design principles. J. Y. Lee and Hwang (2022) conducted a meta-analysis of 16 experimental studies that used AI chatbots for English language instruction in South Korea. Their results showed that chatbot technology had a significant effect on student learning. Lower school levels and shorter usage periods were more effective and using a purpose-built chatbot showed greater effects than using a general-purpose chatbot. The effects of linguistic competence and the affective component were particularly notable (J. Y. Lee & Hwang, 2022).

The affordances of ChatGPT in an academic context vary according to each student's individual learning skills, digital literacy, and innovativeness. During the early stages of technology adoption, the affordances of a specific medium are not well known. Thus, while inquiring about students' experiences of using ChatGPT for their classroom-related activities including writing essays, taking exams, and finding relevant information, we also explored how they perceived ChatGPT's action possibilities (what they can do with it) and constraints (challenges and/or limitations).

Huang et al. (2022) conducted a systematic review of chatbot-supported learning using a grounded theory approach to identify the pedagogical and technological affordances of AI use. The results showed three technological affordances—timeliness, personalization, and ease of use—and five pedagogical uses—for recommendations, for transmissions, as interlocutors, as simulations, and as helplines. The chatbots also encouraged students' social presence through effective and open communication. Okonkwo and Ade-Ibijola (2021) identified the integration of content, quick access, motivation and engagement, and immediate assistance as advantages of AI chatbots, while Pérez et al. (2020) found that chatbots have been successfully applied in the pedagogical domain. Overall, the use of chatbots in education has shown potential, both as administrative and teaching tools.

On the other hand, Deng and Yu (2023) examined the use of chatbot technology in learning and showed that it could not significantly improve critical thinking, learning engagement, and learning motivation. Moreover, the intervention duration did not influence chatbot-assisted learning, different from aforementioned J. Y. Lee and Hwang's (2022) findings. However, chatbot technology significantly improved explicit reasoning, learning achievement, knowledge retention, and learning interest.

Pillai et al. (2023) investigated the adoption intention and actual usage of AI-based teacher bots for learning, using a mixed-method design to explore the adoption of teacher-bots for learning. The study found direct influences on adoption intention, including PEOU, PU, personalization, interactivity, perceived trust, anthropomorphism, and perceived intelligence. Essel et al. (2022) conducted an experiment to investigate the impact of a virtual teaching assistant (chatbot) on students in Ghanaian higher education. They tracked the academic achievement of 68 university students for four months and the results showed that students who interacted with the chatbot demonstrated better academic performance than students who interacted only with their professors. Therefore, the overall results of using chatbots in education indicate that chatbots in education have promising effects.

Despite various efforts, empirical research on the evaluation or satisfaction of students when ChatGPT is applied in classes with human teachers in higher education is still lacking. Because generative AI adoption and usage is still in infancy, many users including college students are in the process of exploring this new technology to determine its utility and limitations. Therefore, we asked the following RQ:

RQ4: How do college students perceive various affordances of ChatGPT for their classroom activities?

3. Methods

The goal of the mixed-method approach was to provide a richer interpretation of the study results by explaining the relationships between variables through statistical analyses identified as relevant and significant based on existing research, while also incorporating structured questionnaire analyses that captured the lively experiences and voices of the participating students, going beyond a simple analysis of variable correlations.

3.1. Data Collection and Participants

Upon acquiring an institutional review board's approval, we first conducted a survey with five open-ended questions asking students taking a class in which a professor (one of the authors) encouraged them to try using ChatGPT for various class-related activities. The course was designed to explore the historical and thematic aspects of "media technologies and culture" by examining three modules: (a) media ontology, (b) media epistemology, and (c) media axiology. The professor provided a specific guideline for generative AI usage in the syllabus: the students must use it to develop ideas, revise, and improve final drafts of their essays, but when they do, its usage must be acknowledged. The first survey was distributed one day after the class' midterm exam (late April 2023), and participation in the survey was voluntary. More than 90% of the students ($n = 104$) participated in the survey. Table 1 summarizes the research steps.

Table 1. A summary of the research steps.

Period	Methods	Sample
April 2023	Open-ended survey with five questions (thematic analysis #1)	$n_1 = 154$
June 2023	Quantitative survey	$n_2 = 48$
June 2024	A survey including both open-ended questions (thematic analysis #2) and quantitative measures	$n_3 = 106$
August 2024	Statistical modeling of the combined quantitative data and comparative analysis of the two thematic analyses results	$n_4 = 48 + 106 = 154$

The professor explained that the purpose of the survey was to explore students' experiences and evaluations of using ChatGPT for their education and asked them to provide honest feedback. Students' names and ID numbers were collected only for assigning extra credits, and no other personally identifiable information was collected. We therefore do not have information on their age or gender distribution. The following five questions were used in the open-ended survey:

1. How useful was ChatGPT in aiding your understanding and mastery of the exam questions?
2. How did you approach using ChatGPT to assist in answering the examination?
3. Did you encounter any challenges or limitations in using ChatGPT during the examination? If so, please describe them.
4. Do you think the "Open-ChatGPT" exam enhances or detracts from your learning experiences? Please describe your experiences in detail.
5. Do you have any suggestions for the use of ChatGPT for helping with effective learning in the classroom, homework, and examinations and to ensure more responsible and ethical use of this technology?

Some students were predicted to have started using ChatGPT in November 2022, but many were introduced to the chatbot upon joining this class (i.e., Media, Culture, & Technology) in March 2023, which was two months before this survey.

Shortly before the semester ended in mid-June 2023, we conducted a closed-ended survey with quantitative measures. These measures were adopted from the previous literature on chatbot usage and interactions with an AI assistant (S. K. Lee et al., 2021; see Section 3.2 for more details). As the survey was anonymous and no extra credit was offered for the second time, participation was much lower, and only 48 students took the survey. Toward the end of the spring semester of 2024, another survey including both quantitative measures and open-ended questions was distributed in a class on the same subject with the same instructor but a different group of students ($n = 106$). The quantitative measures and open-ended questions used in the survey were the same as those used in the 2023 class.

We combined the two samples for statistical modeling (154 students). Participants' ages ranged from 19 to 30 years ($M = 21.92$, $SD = 1.85$), with 27 male (17.5%) and 115 female (74.7%) students. Twelve students did not disclose their gender. Their monthly income ranged from \$0 and \$9,000 ($M = \$750.21$, $SD = \$985.36$).

3.2. Measures

3.2.1. Motivations

Various types of motivations for using ChatGPT were measured with a total of 21 items covering six dimensions (Menon, 2022): entertainment, novelty, study guide, social influence, active interaction, and engagement. Each item was measured on a 5-point Likert-type scale (1 = *strongly disagree*, 5 = *strongly agree*), as were the rest of the major variables in this study. An example item for the entertainment

dimension was “because it is entertaining”; for the novelty dimension was “the technology is innovative”; for the study guide dimension was “to get guidance on the study course”; for the social influence dimension was “because my friends and peers are using it”; for the active interaction dimension was “I feel active when I use it,” and for the engagement dimension was “it is very engaging.” Three additional items were used for the study guide dimension because the main context of this study was a higher education setting.

3.2.2. PU, PEOU, and Actual Usage

Two major variables of the TAM were measured with three items each (Davis, 1989) and the statements were adapted to the uses of ChatGPT (Pillai et al., 2023; Sabah, 2016). An example item for measuring PU was “I feel with ChatGPT that I learn better” and one for PEOU was “ChatGPT is easy to use.”

Three items inquiring how often participants used ChatGPT measured actual usage levels (Mohammadi, 2015; Pillai et al., 2023). An example item for measuring the level of actual usage was “I use ChatGPT every class.”

3.2.3. Trust and Innovativeness

Three items inquiring how much participants trusted ChatGPT measured the level of trust in the AI system (Roca et al., 2009). An example item of perceived trust was “I feel interaction with ChatGPT is secure enough.”

Each participant’s innovativeness was measured with four items to control for their effects on the usage and trust of ChatGPT (KISDI, 2023). Innovative individuals are more likely to adopt and use new technologies (Welch et al., 2020). An example of innovativeness was “I tend to purchase a new product with added features that are not present in my current product.”

3.3. Data Analysis

For quantitative data, after checking for normal distribution and missing data, factor analyses were conducted to check the measurement structure and reliability of the motivation scale. Five factors (entertainment, novelty, study guide, social influence, and active interaction) were extracted after excluding six items with low factor loadings. The Cronbach’s alpha value for each factor was 0.74 or higher, and the dimension of active interaction included two items for which a Spearman’s rho correlation (0.59, $p < 0.001$) was calculated. Additionally, a bivariate correlation analysis was conducted among the major variables (see Table 2 for descriptive statistics).

For the qualitative data (i.e., open-ended survey answers to five questions), one author manually coded the entire dataset and conducted a thematic analysis separately for each class’ data. After categorizing the data based on their similarities and differences, the frequency of such answers in the sample was counted. Results from each analysis (one for 2023 and another for 2024), and the main themes found and labeled from each dataset, were compared for further analysis and integrated for the presentation of this article.

Table 2. Correlations between the major variables.

Variables	*M* (*SD*)	1	2	3	4	5	6	7	8	9
1. Entertainment	3.78 (0.81)	—								
2. Novelty	3.99 (0.80)	0.73**	—							
3. Study guide	3.37 (0.90)	0.28**	0.28**	—						
4. Social influence	2.85 (1.02)	0.22**	0.32**	0.12	—					
5. Active interaction	3.09 (1.05)	0.37**	0.25**	0.28**	0.34**	—				
6. PEOU	3.86 (0.78)	0.34**	0.31**	0.33**	–0.01	0.20*	—			
7. PU	3.41 (0.93)	0.37**	0.32**	0.57**	0.18*	0.42**	0.51**	—		
8. Trust	2.81 (0.98)	0.37**	0.27**	0.32**	0.15	0.31**	0.45**	0.53**	—	
9. Actual usage	2.81 (1.13)	0.25**	0.16*	0.51**	0.16*	0.40**	0.44**	0.47**	0.42**	—
10. Innovativeness	3.08 (0.96)	0.38**	0.40**	0.26**	0.38**	0.36**	0.12	0.18*	0.22**	0.38**

Note: * $p < 0.05$; ** $p < 0.01$.

4. Results

4.1. Results of Hierarchical Regressions for RQ1–RQ3

4.1.1. Motives of ChatGPT Usage

To explore the potentially diverse motivations (RQ1a) of using ChatGPT in an educational setting, we asked the students how strongly they agreed with each of the following motives: study guide, engagement, novelty, social influence, active interaction, and entertainment. The results of the factor analysis revealed five dimensions of motives in our student sample, and the dimension of engagement did not form a meaningful factor. The descriptive statistics revealed that participants of this study showed the strongest motive in the novelty dimension ($M = 3.99$, $SD = 0.80$, Cronbach's $\alpha = 0.82$), followed by the entertainment motive ($M = 3.78$, $SD = 0.81$, Cronbach's $\alpha = 0.92$). The motive of using ChatGPT as a study guide ranked third ($M = 3.37$, $SD = 0.90$, Cronbach's $\alpha = 0.81$), and the motives of active interaction ($M = 3.09$, $SD = 1.05$, Spearman's rho = 0.59) and social influence ($M = 2.85$, $SD = 1.02$, Cronbach's $\alpha = 0.74$) were weaker than the other motives.

4.1.2. Correlations Between Motives and Actual Usage

To answer RQ1b, we performed a hierarchical linear regression with the actual usage of ChatGPT as a criterion variable, the five dimensions of motives as predictors, and the demographic variables and innovativeness as controls. Controlling for the effect of individual innovativeness ($\beta = 0.20$, $t = 2.34$, $p < 0.05$), the analysis revealed that two out of five motives were statistically significant predictors of the actual usage level of ChatGPT. The motive of study guide was a strong predictor ($\beta = 0.41$, $t = 5.29$, $p < 0.001$), and that of active interaction ($\beta = 0.24$, $t = 2.81$, $p < 0.01$) was also significantly correlated with the level of actual usage (see Table 3).

Table 3. Hierarchical regression predicting actual usage levels.

	Model I		Model II		Model III		Model IV	
	β (*t*)		*β* (*t*)		*β* (*t*)		*β* (*t*)	
Gender (Female = 0)	0.18	(1.91)	0.12	(1.34)	0.09	(1.14)	0.07	(0.95)
Income	0.04	(0.42)	0.00	(0.03)	0.00	(−0.01)	0.00	(0.03)
Age	0.01	(0.15)	−0.01	(−0.09)	0.04	(0.51)	0.05	(0.75)
Innovativeness			0.38***	(4.41)	0.21*	(2.34)	0.26**	(3.11)
Entertainment					0.03	(0.29)	−0.07	(−0.67)
Novelty					−0.08	(−0.74)	−0.11	(−1.05)
Study guide					0.41***	(5.29)	0.28**	(3.27)
Social influence					0.04	(0.46)	0.05	(0.68)
Active interaction					0.24**	(2.81)	0.20*	(2.30)
PEOU							0.23*	(2.63)
PU							0.15	(1.48)
R^2 change	0.04		0.14***		0.24***		0.07**	
R^2	0.04		0.17***		0.41***		0.48***	
Adjusted R^2	0.01		0.14***		0.36***		0.42***	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Thus, participants who had strong motivations for using ChatGPT as a study guide or wanted to actively interact with it seemed to have used it more frequently than others who did not have such motives strongly. The other three motive dimensions were not significantly correlated with the actual usage of ChatGPT and no demographic variables significantly predicted the level of usage. The motivation variables in the regression increased the amount of explained variance in the actual usage level of ChatGPT by 23.7%.

4.1.3. Correlations Between PEOU, PU, and Actual Usage

To address RQ2, in the final block of the hierarchical regression, we entered PEOU and PU as predictors of actual usage level. The results showed that PEOU, not PU, was significantly associated with the level of ChatGPT usage ($\beta = 0.23$, $t = 2.63$, $p < 0.05$). Participants' innovativeness and the two motives (i.e., study guide and active interaction) remained significant predictors of actual usage after considering PEOU and PU. Thus, all things being equal, when ChatGPT usage was perceived as easy, the participants seemed to have used it more frequently. Including PEOU as a predictor of usage level increased the amount of explained variance by 6.7%, and all predictors together explained approximately 42.4% of the variance in the actual usage level of ChatGPT.

4.1.4. Relationships Between Motives, Actual Usage, and Trust

To address RQ3, we ran another hierarchical regression with trust in ChatGPT as the criterion variable (see Table 4). Participants' (a) motivations for using ChatGPT were entered as predictors after controlling for the effects of demographics and individual innovativeness, and (b) the actual usage level of ChatGPT was considered after controlling for the effects of PEOU and PU. We found that gender ($\beta = 0.24$, $t = 2.86$,

$p < 0.01$) was a significant predictor, and two out of five motives were significantly related to the level of trust in ChatGPT. Male participants seemed to trust ChatGPT more than female participants. Entertainment ($\beta = 0.35$, $t = 2.93$, $p < 0.01$) and study guide ($\beta = 0.18$, $t = 2.10$, $p < 0.05$) motives significantly predicted the level of trust, and the amount of explained variance increased by 17.4%.

Table 4. Hierarchical regression predicting trust in ChatGPT.

	Model I		Model II		Model III		Model IV	
	β (t)		β (t)		β (t)		β (t)	
Gender (Female = 0)	0.27**	3.06	0.24**	2.86	0.24**	3.27	0.24**	3.21
Income	0.04	0.49	0.10	1.10	0.12	1.55	0.12	1.54
Age	−0.09	−1.00	−0.06	−0.71	0.00	−0.03	−0.01	−0.06
Innovativeness	0.17	1.94	−0.05	−0.56	0.02	0.28	0.01	0.14
Entertainment			0.35**	2.93	0.18	1.65	0.18	1.66
Novelty			−0.06	−0.47	−0.07	−0.66	−0.07	−0.62
Study guide			0.18*	2.10	−0.07	−0.81	−0.08	−0.90
Social influence			0.07	0.77	0.07	0.81	0.06	0.78
Active interaction			0.14	1.47	0.03	0.39	0.03	0.29
PEOU					0.17	1.88	0.16	1.71
PU					0.44***	4.27	0.44***	4.16
Actual usage							0.04	0.42
R^2 change	0.12**		0.17***		0.16***		0.00	
R^2	0.12**		0.30***		0.46***		0.46***	
Adjusted R^2	0.10**		0.24***		0.40***		0.40***	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Contrary to the predictions of the actual usage level of ChatGPT, PU, not PEOU, was significantly correlated with the level of trust in ChatGPT ($\beta = 0.44$, $t = 4.16$, $p < 0.001$). Thus, when participants perceived the usefulness of ChatGPT to be higher, they seemed to trust its safety and information security more. However, the actual usage level of ChatGPT did not significantly predict the level of trust in ChatGPT after controlling for the effects of PU. Together, all predictors explained approximately 40% of the trust level in ChatGPT.

4.2. Thematic Analysis of Open-Ended Answers About ChatGPT Usage for Student Learning

To address RQ4, thematic analyses were conducted for open-ended answers from participants on their views and experiences of using ChatGPT for their class-related activities. The results of the two thematic analyses were compared and integrated to create a final report. Table 5 summarizes the thematic analysis results with exemplary themes, codes, and direct quotes.

Table 5. A summary of the thematic analysis results with examples.

Themes		Initial Codes	Representative Quotations
Positive assessments of generative AI's usefulness	Facilitating essay writing	Providing ideas for essay writing	"I only used it when I ran out of ideas and did not have any directions to go."
		Organizing ideas	"As in most subjects, it was helpful in providing broad guidance in gathering my thoughts."
	Enhancing efficiency in learning and research	Aiding comprehension of the subject	"It can explain key terms for me in very easy words compared to the ones I searched online. I use this sentence a lot: 'Please explain xxx in easy words that primary students can understand.'"
		Functioning as a search engine	"Generative AI is useful when it comes to looking for articles about specific information."
Negative assessments of generative AI's usefulness	Concerns regarding the credibility of generative AI	Lacking depth in information	"Useful to a certain extent, it often provides very generic answers or suggestions."
		Lacking accurate information	"Sometimes it is accurate but other times it's not."
Approach to using generative AI to assist essay writing	AI as a co-writer	Asking for additional topics or ideas	"I asked what kind of topics related to my essay should I write about."
	AI as a convenient research and learning tool	Asking for general, background information	"Look up the background or more general explanation of some of the points to understand easily."
Challenges/limitations of using generative AI during essay writing	Challenges in gaining relevant and useful information	Lack of novel and/or detailed information	"Generative AI gives very vague information (for example, it doesn't give any meaningful insight and just repeats the things that I already know just in an organized way)."
	Concerns regarding accuracy and credibility	Incorrect and illogical information	"The description of a concept has not always been accurate, and sometimes the description is not logical."
Positive assessments regarding open-generative AI exam	Enhanced convenience and engagement in the learning environment	Makes studying for exams less burdensome	"The parts that are unnecessary or close to pure labor are taken care of instead, and humans can focus on the more creative and thinking aspects. It also helps with understanding and translating difficult materials."
	Academic support for mastery of exam material	Facilitates comprehension of exam material	"It enhances my learning experiences because it helped me understand concepts that were unclear during class or my readings."

Table 5. (Cont.) A summary of the thematic analysis results with examples.

Themes		Initial Codes	Representative Quotations
Negative assessments regarding the open-generative AI exam	Deterrence of independent learning	Promotes superficial learning	"Detracts because many would not study the content of the class, confident that they can pass the exam thanks to generative AI."
	Interference with students' comprehension of material	Confuses students with inadequate and inaccurate information	"Detracts from my learning experience as it provides skewed responses especially for controversial issues; it also is not very creative as it is based on previous data compiled online."
Suggestions for the use of generative AI for helping effective learning	Establishing clear guidelines and boundaries for AI use	Limit the use of AI for language and writing	"I think ChatGPT is useful to use, but rather for checking grammar or other mistakes, than for really creative tasks."
	Promote critical thinking and autonomous learning	Self-regulation on behalf of students	"Users must self-regulate their use of ChatGPT as a tool or assistant. It should be used for task-based activities, organizing, and helping to jump-start ideas rather than replicating ideas."

4.2.1. Usefulness of ChatGPT for Student Learning

In both analyses, students generally acknowledged the usefulness of ChatGPT in their learning experiences, particularly in aiding comprehension and organizing ideas. In the first analysis, approximately 68% of the students found ChatGPT useful, whereas approximately 52% in the second analysis reported its utility in helping with essay writing. Common benefits included facilitating comprehension, providing outlines, and summarizing key points, which saved students' time and effort.

One point of distinction is that the first analysis highlighted students' use of ChatGPT for diverse purposes, including as a dictionary or translation tool, whereas the second emphasized ChatGPT's ability to function as a cowriter or editor. For instance, in the second analysis, students used ChatGPT to overcome writer's block and generate thesis statements. Both analyses reflected the students' appreciation of ChatGPT's role in summarizing materials and offering alternative perspectives.

4.2.2. Limitations and Concerns

Both analyses noted concerns regarding the depth and accuracy of the ChatGPT responses. The first analysis reported that 33% of the students mentioned that ChatGPT provided only surface-level knowledge and 43% experienced difficulties in obtaining the right information because of a lack of critical insight. Similarly, the second analysis revealed that students found generative AI responses generic and often repetitive, failing to offer meaningful insights or detailed information.

A unique concern in the second analysis was skepticism over the credibility of ChatGPT's outputs, with students expressing doubts about its accuracy and potential to generate incorrect or illogical information.

Additionally, concerns regarding time-consuming processes when using ChatGPT were mentioned in both analyses.

4.2.3. Approach to Using ChatGPT

In both analyses, students used ChatGPT either during or after the exam or the essay-writing process. The first analysis reported that nearly 97% of the students interacted with ChatGPT while formulating their responses, with some rephrasing questions or copying them directly into ChatGPT for better organization. Similarly, the second analysis discussed how students used ChatGPT as a cowriter to assist with structuring essays, generating thesis statements, and offering outlines.

The second analysis expanded on ChatGPT's role as an editor, with students using it to check grammar and make their writing more professional. This element was less pronounced in the first analysis, in which the focus was on students using ChatGPT for idea generation rather than refining their own work. It is not entirely clear whether this difference between the first and second datasets originated from students' cumulative experiences of using ChatGPT in the second dataset, collected more than a year after the technology became available, or whether ChatGPT itself has advanced in its functions with multiple upgrades since its launch. In either case, it seems that students used ChatGPT more proactively to aid and improve their writing, and the role of generative AI changed from an idea generator helper to a co-editor/writer.

4.2.4. Challenges Encountered

Both analyses discussed the challenges students faced when using ChatGPT. In the first analysis, approximately 13% of the students experienced technical issues, such as network problems or limitations on the questions asked per hour. The second analysis emphasized the challenges in obtaining relevant and accurate information, especially on controversial topics and issues with incomplete responses in non-English languages.

The second analysis introduced the theme of students encountering fabricated sources or incorrect information generated by ChatGPT, leading to further distrust in the system. Both reports mentioned that the lack of novel information forced students to rely on their own notes or alternative sources for exam preparation and essay writing.

4.2.5. Impact on Learning Experience

The first analysis revealed that 63.5% of the students believed that the open-ChatGPT exams enhanced their learning experiences by encouraging deeper engagement with the material. Students noted that ChatGPT helped them question the course content and provided structured insights for their essays. Similarly, the second analysis noted how students felt that ChatGPT made the learning process more efficient and accessible by simplifying complex concepts and acting as a "private tutor."

However, both reports highlighted concerns regarding the detraction of ChatGPT from learning. In the first analysis, some students felt that relying on ChatGPT disengaged them from the course material, leading to a lack of critical thinking and perceived cheating. This aligns with the findings of the second analysis, where students argued that ChatGPT prompted superficial learning and led to an over-reliance on AI.

4.2.6. Recommendations for Ethical Use

Both analyses emphasized the importance of the responsible and ethical use of ChatGPT. In the first analysis, students called for individual responsibility in critically assessing the ChatGPT output and cautioned against over-dependence on AI. The second analysis reinforced this finding by suggesting the need for clearer guidelines and AI literacy to promote critical thinking and ethical engagement. It also advocates fostering classroom discussions on the use of AI and encouraging self-regulation.

Both analyses provided a balanced view of the benefits and limitations of ChatGPT in aiding student learning. While ChatGPT proved to be a valuable tool for organizing ideas, summarizing information, and providing insights, concerns about its depth, accuracy, and potential to foster an overreliance on AI remained consistent across both reports. For ChatGPT to be effectively integrated into learning environments, clear guidelines, critical evaluation, and ethical use must be promoted, with a focus on supplementing, not replacing, student-driven inquiry and critical thinking.

5. Discussion

This study aimed to elucidate how students adopt new generative AI technology in college classrooms. While technology has advanced rapidly, human adaptation to it lags, imposing limitations on the swift and profound transformation of the classroom. Due to this delayed adaptation, social changes occur gradually and stabilize, presenting a paradox. Therefore, there is no need for excessive and hypersensitive reactions to the development of new technologies. The internalization and utilization of these technologies remain within the purview of humanity. We should calmly contemplate how to employ them in our livelihoods, industries, education, and everyday lives, and approach them astutely.

Furthermore, this study sheds light on the efficient use of generative AI in thematic essay writing for future higher education. The upcoming generation may not necessarily require proficiency solely in technology but rather an acute understanding of universal human orality, literacy, visuality, and interactivity (Abdul-Kader & John, 2015; Ong, 1982). The timeless principles of human culture, which have endured for thousands of years, provide the most essential and efficient blueprint for preparing for an unpredictable future in education. Thus, schools and universities across all levels in this era of “real virtuality” (Castells, 2000) must devise and implement a pedagogy that actively engages with the universal principles of human culture for the next generation. To contribute to this mission of higher education, this study, by utilizing both quantitative and qualitative methods, examined how college students used ChatGPT and perceived its affordances.

5.1. Key Motivations for ChatGPT Adoption in Education

This study identified five primary motivational dimensions for using ChatGPT in education: novelty, entertainment, study guidance, active interaction, and social influence. The quantitative findings revealed that novelty and entertainment were the strongest motives, suggesting that students viewed ChatGPT more as an exploratory or recreational tool than as a means for sustained engagement. This contrasts with earlier studies such as Kuhail et al. (2023) and Huang et al. (2022), which emphasized motivation and engagement as key affordances of chatbots. The lack of structured interactions and long-term integration into

educational practices may explain this discrepancy. The qualitative findings further highlighted students' mixed perceptions of ChatGPT's utility, with many describing it as providing surface-level knowledge and generic responses. For example, many students noted a lack of depth in their responses and reported challenges in obtaining critical insights. These findings suggest that, while novelty and entertainment may drive adoption, limitations in ChatGPT's ability to provide meaningful, detailed information temper its perceived value as a learning tool.

5.2. Predictors of ChatGPT Usage

The regression analysis showed that using ChatGPT as a study guide and for active interaction significantly predicted actual usage levels. This aligns with Pérez's et al.'s (2020) findings that highlighted the pedagogical value of chatbots as study aids. The qualitative data supported these insights, revealing that students frequently used ChatGPT to generate ideas, organize thoughts, and offer outlines. In the analyses of two datasets, students reported using ChatGPT as a cowriter or editor, particularly for checking grammar and professionalizing their writing. The qualitative data also revealed a shift in usage patterns over time, with students increasingly employing ChatGPT as a coeditor rather than solely as an idea generator. This evolution may reflect either students' growing familiarity with the tool or advancements in ChatGPT's capabilities through successive updates.

5.3. Trust and Its Drivers

Entertainment and study-guide motivations were positively linked to trust in ChatGPT, with male participants showing higher trust levels. Although PU emerged as a significant predictor of trust, PEOU did not, contrasting with its influence on usage level. The qualitative findings offered additional context, highlighting skepticism about the accuracy of ChatGPT. Notably, 43% of the students reported doubts about its credibility, citing concerns over fabricated sources and incorrect information. These issues were particularly pronounced when students sought information on complex and controversial topics. Despite these challenges, many students appreciated ChatGPT's role in simplifying complex concepts and enhancing accessibility, likening it to a "private tutor." This duality—trust in its utility but caution about its reliability—underscores the nuanced relationship between student perceptions and ChatGPT's evolving role in education.

5.4. Challenges and Impacts on Learning

The findings highlighted several challenges in using ChatGPT, including technical issues, time-consuming processes, and concerns regarding superficial learning. Qualitative findings showed that some students experienced technical limitations, such as network problems or usage restrictions, while others reported difficulties obtaining relevant and accurate information. Additionally, the lack of novel insights often forced students to rely on alternative sources, diminishing the perceived value of ChatGPT in academic contexts.

Despite these challenges, ChatGPT was seen as enhancing the learning experience for majority of the students in the first analysis, with many noting its ability to encourage deeper engagement with course content. However, both analyses also revealed concerns about overreliance on AI, which some students felt detracted from critical thinking and meaningful engagement. This aligns with the quantitative finding that

PU did not significantly predict usage, suggesting that ChatGPT's practical utility may be overshadowed by its limitations and students' cautious approach toward integrating it into their learning routines.

5.5. Theoretical Implications

The findings of this study have theoretical implications for various communication and media theories, including uses and gratifications (Katz et al., 1973), the TAM (Davis, 1989), and media affordances (Gibson, 1986; Norman, 1999). First, this study drew on Park's (2010) integrative model of uses and gratifications and the TAM to examine students' varied motivations for using ChatGPT and their perceptions of its usefulness (PU) and ease of use (PEOU). While Park's study considered three motives (entertainment, communication, and instrumental) and how they connected to PU, PEOU, and voice over internet protocol usage, this study found that three motives (entertainment, novelty, and study guide) affected ChatGPT usage in a higher education context, even after controlling for PU and PEOU. In Park's (2010) study, PEOU was not directly related to actual usage but was linked to entertainment and communication motives, with PU mediating the impact of PEOU on usage. By contrast, our study showed that PEOU, not PU, influenced actual usage, while the study guide and active interaction motives remained significant even after controlling for the effect of PEOU. These differences may reflect the distinct technological contexts between the two studies. Generative AI tools such as ChatGPT may benefit more from PEOU because of their novel and assistive roles in academic tasks. Nevertheless, the uses and gratifications theory and the TAM retain their theoretical relevance by demonstrating flexible applicability across different technological settings.

Second, this study highlights how ChatGPT's affordances, such as providing quick responses and assisting with concept clarification, shape students' learning experiences. Media affordance theory suggests that technologies offer specific capabilities that influence their use (Gibson, 1986; Norman, 1999). In this case, ChatGPT's affordances make it a useful tool for tasks such as organizing thoughts and accessing information efficiently. However, the study also highlights the constraints of ChatGPT, particularly its potential to limit creativity and critical thinking, which aligns with the findings of Deng and Yu (2023) but contrasts with Abramson's (2023) view that AI can enhance creativity by offloading routine tasks. Media affordance theory posits that, while technologies offer affordances, they also impose constraints. In ChatGPT's case, its ability to provide quick answers may inhibit deeper exploration and creative problem-solving if students rely too heavily on it.

5.6. Practical Implications

The findings of this study can inform the integration of generative AI tools into college curricula and support student learning. Our research suggests that college students are more likely to use ChatGPT as a "study guide" when instructed and guided by their professors. Rather than leaving students to determine how to use the tool ethically and responsibly, encouraging its use for class-related activities—such as writing essays, searching for information, and taking exams—can foster greater awareness and reflection on ChatGPT's benefits and limitations.

These insights are also valuable for industry stakeholders and developers of generative AI tools. The challenges highlighted by students, such as difficulties in locating accurate information when using non-English languages (e.g., Korean in this study), underscore the need to improve the tool's ability to identify and provide reliable

information, along with credible sources. Without addressing these issues, discussions on the ease of use or utility of generative AI tools in higher education risk becoming irrelevant.

5.7. Limitations and Future Directions

This study has several limitations that should be addressed in future research. First, the sample size for the statistical modeling was relatively small, as data collection was limited to two classes of college students enrolled in the same subject. As this was a field experiment, we aimed to ensure that all students were exposed to the same instructor's guidance on using ChatGPT for their class-related activities. A larger sample size would have allowed for more statistical power and advanced analyses, such as structural equation modeling, which could test both direct and indirect relationships among variables simultaneously. Therefore, future research should consider collecting data from multiple classes or from a larger single class.

Second, because our study examined ChatGPT usage and perceived affordances among Korean college students, the findings have limited generalizability. One potential reason why the "study guide" motive emerged as a significant predictor of usage level could be the unique context of our research, situated in higher education. Because students were encouraged by the professor, an authority figure, to use ChatGPT for all kinds of class activities, they might have believed that it could be used as a study guide and been artificially motivated. Although other motives—such as active interaction with ChatGPT or entertainment—were also related to the usage of or trust in the AI system, the findings may differ in other settings, such as workplaces. If full-time professionals were participants in the study, their motives for using ChatGPT as a work tool might have been highlighted. Therefore, we cannot claim that our findings can be readily generalizable to other contexts or populations of ChatGPT users.

Third, we cannot claim causal relationships among the variables examined in this study. Based on the uses and gratifications theory and TAM, we assumed causality from motives to usage, and from PU and PEOU to usage. However, it is possible that the participants' prior usage levels influenced their motives or perceptions of ChatGPT's usefulness and ease of use. Similarly, while our research model posited a causal influence of motives and usage on trust, the participants' existing trust in the AI system could also affect their motives and usage levels. Therefore, until we collect and analyze longitudinal data, these causal relationships remain speculative.

Finally, we suggest that future research include more measurement items to capture various types of motives for using ChatGPT. This study primarily used three items per dimension, except for the "study guide" dimension, which may have affected the factor analysis results, leading to the exclusion of the "engagement" dimension. To reduce measurement errors, future studies should include additional items per dimension, and prepare for the possibility of low factor loadings or omission of important dimensions.

Acknowledgments

The author(s) acknowledge the use of ChatGPT 4.0 to improve the language and readability of this article. After using this tool/service, the authors reviewed and edited the content as required and take full responsibility for the content.

Funding

Data collection was supported by BK21 funds from the School of Media and Communication at Korea University.

Conflict of Interests

The authors declare no conflict of interests.

Data Availability

Data from this study will become available upon reasonable request to the first author.

References

Abdul-Kader, S. A., & John, D. (2015). Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, *6*(7), 72–80. https://doi.org/10.14569/IJACSA.2015.060712

Abramson, A. (2023). How to use ChatGPT as a learning tool. *American Psychological Association Monitor*, *54*(3), 67. https://www.apa.org/monitor/2023/06/chatgpt-learning-tool

Atlas, S. (2023). *ChatGPT for higher education and professional development: A guide to conversational AI*. University of Rhode Island.

Baidoo-Anu, D., & Ansah, L. O. (2023). *Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning*. SSRN. https://doi.org/10.2139/ssrn.4337484

Benjamin, W. (1968). The work of art in the age of mechanical reproduction. In H. Arendt (Ed.), *Illuminations* (pp. 217–251). Schocken Books (Original work published 1935)

Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of Higher Education*, *1*, 5–22.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment and Evaluation in Higher Education*, *41*(3), 400–413. https://doi.org/10.1080/02602938.2015.1018133

Castells, M. (2000). *The rise of the network society* (2nd ed., Vol. 1). Blackwell.

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers*, *25*(1), 161–182. https://doi.org/10.1007/s10796-022-10291-4

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer.

Deng, X., & Yu, Z. (2023). A meta-analysis and systematic review of the effect of chatbot technology use in sustainable education. *Sustainability*, *15*(4), Article 2940. https://doi.org/10.3390/su15042940

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., & Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eisenstein, E. L. (1979). *The printing press as an agent of change: Communications and cultural transformations in early-modern Europe*. Cambridge University Press.

Essel, H. B., Vlachopoulos, D., Tachie-Menson, A., Johnson, E. E., & Baah, P. K. (2022). The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *International Journal of Educational Technology in Higher Education*, *19*(1), Article 7. https://doi.org/10.1186/s41239-022-00362-6

Floridi, L. (2021). *Ethics, governance, and policies in AI*. Springer.

Gamage, S. H. P. W., Ayres, J. R., & Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. *International Journal of STEM Education*, *9*(1), Article 9. https://doi.org/10.1186/s40594-021-00323-x

Gibson, J. J. (1986). *The ecological approach to visual perception*. Erlbaum.

Holmes, W., Bialik, M., & Fadel, C. (2021). *Artificial intelligence in education: Promises and implications for teaching and learning*. Routledge.

Hsu, T., & Thompson, S. A. (2023, February 13). Disinformation researchers raise alarms about AI. Chatbots. *The New York Times*. https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257. https://doi.org/10.1111/jcal.12610

Jensen, L. X., Buhl, A., Sharma, A., & Bearman, M. (2024). Generative AI and higher education: A review of claims from the first months of ChatGPT. *Higher Education*. Advance online publication. https://doi.org/10.1007/s10734-024-01265-3

June, S., Yaacob, A., & Kheng, Y. K. (2014). Assessing the use of YouTube videos and interactive activities as a critical thinking stimulator for tertiary students: An action research. *International Education Studies*, *7*(8), 56–67. https://doi.org/10.5539/ies.v7n8p56

Kar, S., Roy, C., Das, M., Mullick, S., & Saha, R. (2023). AI horizons: Unveiling the future of generative intelligence. *International Journal of Advanced Research in Science, Communication and Technology*, *387*, 387–391.

Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and gratifications research. *Public Opinion Quarterly*, *37*(4), 509–523. https://doi.org/10.1086/268109

Kerr, P. (2016). Adaptive learning. *ELT Journal*, *70*(1), 88–93. https://doi.org/10.1093/elt/ccv055

KISDI. (2023). *Media panel chosa*. https://stat.kisdi.re.kr/kor/contents/ContentsList.html?subject=SURV&sub_div=S

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Kumar, J. A. (2021). Educational chatbots for project-based learning: Investigating learning outcomes for a team-based design course. *International Journal of Educational Technology in Higher Education*, *18*(1), Article 65. https://doi.org/10.1186/s41239-021-00302-w

Lee, J. Y., & Hwang, Y. (2022). A meta-analysis of the effects of using AI chatbot in Korean EFL education. *Studies in English Language and Literature*, *48*(1), 213–243.

Lee, S. K., Kavya, P., & Lasser, S. C. (2021). Social interactions and relationships with an intelligent virtual agent. *International Journal of Human-Computer Studies*, *150*, Article 102608. https://doi.org/10.1016/j.ijhcs.2021.102608

Lee, S. K., & Sun, J. (2023). Testing a theoretical model of trust in human–machine communication: Emotional experiences and social presence. *Behaviour and Information Technology*, *42*(16), 2754–2767. https://doi.org/10.1080/0144929X.2022.2145998

McLuhan, M. (1962). *The Gutenberg galaxy: The making of typographic man*. University of Toronto Press.

Menon, D. (2022). Uses and gratifications of educational apps: A study during COVID-19 pandemic. *Computers and Education Open*, *3*, Article 100076. https://doi.org/10.1016/j.caeo.2022.100076

Ministry of Science and ICT. (2019). *National strategy for artificial intelligence*. https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=46&mId=10&mPid=9&nttSeqNo=9&sCode=eng

Mohammadi, H. (2015). Investigating users' perspectives on e-learning: An integration of TAM and IS success model. *Computers in Human Behavior*, *45*, 359–374. https://doi.org/10.1016/j.chb.2014.07.044

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Norman, D. A. (1999). Affordance, conventions, and design. *Interactions*, *6*(3), 38–43. https://doi.org/10.1145/301153.301168

OECD. (2023). *OECD digital economy outlook 2024 (volume 1)*. https://www.oecd.org/en/publications/oecd-digital-economy-outlook-2024-volume-1_a1689dc5-en.html

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education*, *2*, Article 100033. https://doi.org/10.1016/j.caeai.2021.100033

Ong, W. J. (1982). *Orality and literacy: The technologizing of the word*. Methuen.

Park, N. (2010). Adoption and use of computer-based voice over internet protocol phone service: Toward an integrated model. *Journal of Communication*, *60*(1), 40–72. https://doi.org/10.1111/j.1460-2466.2009.01440.x

Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549–1565. https://doi.org/10.1002/cae.22326

Piaget, J. (1971). *The theory of stages in cognitive development*. McGraw-Hill.

Pillai, R., Sivathanu, B., Metri, B., & Kaushik, N. (2023). Students' adoption of AI-based teacher-bots (T-bots) for learning in higher education. *Information Technology & People*, *37*(1), 328–355. https://doi.org/10.1108/ITP-02-2021-0152

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching*, *6*(1), 41–56. https://doi.org/10.37074/jalt.2023.6.1.29

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier*. MIT Press. https://doi.org/10.7551/mitpress/7105.001.0001

Roca, J. C., García, J. J., & de la Vega, J. J. (2009). The importance of perceived trust, security, and privacy in online trading systems. *Information Management and Computer Security*, *17*(2), 96–113. https://doi.org/10.1108/09685220910963983

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

Sabah, N. M. (2016). Exploring students' awareness and perceptions: Influencing factors and individual differences driving m-learning adoption. *Computers in Human Behavior*, *65*, 522–533. https://doi.org/10.1016/j.chb.2016.09.009

Tarhini, A., Masa'deh, R. E., Al-Busaidi, K. A., Mohammed, A. B., & Maqableh, M. (2017). Factors influencing students' adoption of e-learning: A structural equation modeling approach. *Journal of International Education in Business*, *10*(2), 164–182. https://doi.org/10.1108/JIEB-09-2016-0032

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Turkle, S. (2015). *Reclaiming conversation: The power of talk in a digital age*. Penguin Books.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Welch, R., Alade, T., & Nichol, L. (2020). Using the unified theory of acceptance and use of technology (UTAUT) model to determine factors affecting mobile learning adoption in the workplace: A study of the science museum group. *International Journal on Computer Science and Information Systems*, *15*(1), 85–98.

Zhao, X., Shao, M., & Su, Y. S. (2022). Effects of online learning support services on university students' learning satisfaction under the impact of Covid-19. *Sustainability*, *14*(17), Article 10699. https://doi.org/10.3390/su141710699

About the Authors

Sun Kyong Lee (PhD, Rutgers University) is a professor at the College of Media and Communication, Korea University. Her research examines the sociocultural antecedents, processes, and consequences of human–machine communication. Lee's published work can be found in journals such as *International Journal of Human-Computer Studies*, *Computers in Human Behavior*, *Behaviour & Information Technology*, and the *Journal of Computer-Mediated Communication*.

Jongsang Ryu obtained an MA from the Graduate School of Media and Communication, Korea University, specializing in human–machine communication. His academic endeavors explore how emerging technologies—particularly AI and interactive platforms—reconfigure communication practices, cultural norms, and the strategic orientation of media firms. Ryu has presented his research on the reciprocal relationships among technology, industry evolution, and social discourse.

Yeowon Jie is a student at the College of Media and Communication, Korea University, where she is enrolled in the linked BA–MA program. She is interested in studying human–computer interaction and social media policy and aims to contribute to research in the understanding of human agency in the face of emerging media and human-centered media development.

Dong Hoon Ma is a professor at the College of Media and Communication, Korea University. His major fields of academic interest range from cultural and historical media theories, cultural studies, popular culture, and media public sphere to the landscape of future media, on which he published numerous articles and book chapters both in English and Korean.

Media and Communication
2025 • Volume 13 • Article 9534
https://doi.org/10.17645/mac.9534

ARTICLE

Open Access Journal

Support for Businesses' Use of Artificial Intelligence: Dynamics of Trust, Distrust, and Perceived Benefits

Lisa Tam [1], Soojin Kim [2], and Yi Gong [1]

[1] School of Advertising, Marketing and Public Relations, Queensland Technology of Technology, Australia
[2] School of the Arts and Media, University of New South Wales, Australia

Correspondence: Soojin Kim (soojin.kim@unsw.edu.au)

Submitted: 30 October 2024 **Accepted:** 6 March 2025 **Published:** 30 April 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

Current research on AI has extensively examined drivers that predict individuals' attitudes and behavioral intentions toward AI use. Despite this, there is limited research that explores factors that influence consumers' acceptance of AI integration into businesses. As more businesses have integrated AI systems into different aspects of their operations, consumers have experienced increasing interactions with AI systems adopted by businesses. Thus, it is critical to understand not only whether individuals trust and accept the use of AI in their everyday lives, but also whether they trust and accept the use of AI by businesses they interact with. As such, this study tests a theoretical framework developed on the basis of current research on AI and technology acceptance. This study used a survey dataset collected from a nationally representative sample of 420 Australian consumers in 2024. The findings revealed that the interplay between faith in general technology, trust and distrust in businesses' AI use, and perceived AI benefits shaped attitudes and behavioral intentions toward businesses using AI. These dynamics also contributed to the approval of businesses' use of AI. The findings offer theoretical and practical insights on how to manage these dynamics to foster positive attitudes and behavioral intentions toward businesses that use AI.

Keywords

artificial intelligence; consumers; distrust; faith in technology; perceived benefits; trust

1. Introduction

AI is defined as "a technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity, and autonomy" (Stryker & Kavlakoglu, 2024,

para. 1). The Commonwealth Scientific and Industrial Research Organisation (CSIRO) has defined AI as "a collection of interrelated technologies to solve problems autonomously and perform tasks to achieve defined objectives, in some cases without explicit guidance from a human being" (Hajkowicz et al., 2019, p. 2). Because AI can be used to complete tasks with minimal or no human intervention, there are concerns about its capabilities, biases, and security risks. As such, current research has identified factors including familiarity (Gerlich, 2023), trust (Jiang et al., 2024), perceived usefulness (Jiang et al., 2024), and perceived risk and threats (Jiang et al., 2024; Sindermann et al., 2022) as significantly influencing attitudes and behavioral intentions toward AI use.

AI has transformed the ways that consumers interact with businesses. First, AI enables businesses to deliver personalized experiences by analyzing large volumes of data (Szleter, 2024). Second, AI enables businesses to provide immediate, around-the-clock customer support with virtual assistants and chatbots (Szleter, 2024). Third, AI enables businesses to gain deeper insights into customer data to improve their products and services (Jobanputra, 2024). These capabilities have optimized operations and have made businesses more efficient by saving time and costs (Haan & Watts, 2023a). Despite this, only 32% of consumers had successfully resolved a customer service issue using AI, and 63% were frustrated with businesses' use of AI for customer support (Hyken, 2024). Consumers were reported to have "fear and frustration" toward their interactions with businesses' AI systems (Hyken, 2024, para. 11).

Despite the prevalence of AI use, it remains a challenge for businesses to understand consumers' attitudes toward AI and to gain consumers' approval for their use of AI in their business operations. Current research on AI acceptance has focused on users' acceptance of AI technologies (Kelly et al., 2023). Yet, there is limited research on consumers' perceptions of businesses that use AI technologies (Oyekunle et al., 2024). When interacting with AI systems adopted by businesses, consumers generally have concerns about privacy and security (Alhitmi et al., 2024) and AI's limited capacity in resolving consumer complaints (Agnihotri et al., 2021). As AI adoption grows in businesses across industries, such as in health, education, and manufacturing ("Exploring AI adoption," 2024), it is critical to understand consumers' concerns and investigate the dynamics of different factors that influence their support for businesses' use of AI.

Amidst the large body of research on AI acceptance, most of the studies have adopted traditional acceptance models such as the technology acceptance model (TAM) and the unified theory of acceptance and use of technology (UTAUT) to uncover drivers for AI acceptance for individuals' own use of AI (Kelly et al., 2023). Hence, there are calls for AI acceptance studies to (a) include trust and attitudes, (b) examine actual use behaviors, (c) examine users' understanding of AI technologies (Kelly et al., 2023), and (d) explore consumers' perceptions of how businesses use them (Frank et al., 2023; Jain et al., 2024). On the last point, Jain et al. (2024) described artificial intelligence consumer behavior (AI CB) as consumer behaviors influenced by the application of AI in consumer interactions and suggested that future research employ theoretical lenses to explore consumers' interpretations, perceptions, and responses to the adoption of AI technologies by businesses.

In response to the research gaps stated above, this study proposes and tests a theoretical framework that explores factors that shape consumers' attitudes and behavioral intentions toward businesses that use AI and their approval of businesses' AI use. Several propositions form the foundation of this framework. First, current research on AI is characterized by two clusters: one on individuals' acceptance of AI and the other on

consumers' behaviors related to AI (Jain et al., 2024; Kelly et al., 2023). In fact, there is a confluence between the two. Many people are open to using AI, but some are skeptical about how businesses use AI, and this can affect how many people actually use AI (Frank et al., 2023). Second, while actual AI use behaviors have been under-researched (Kelly et al., 2023), consumers have pre-existing beliefs and expectations on how businesses use AI (Frank et al., 2023; Jain et al., 2024). As such, consumers' pre-existing propensity and beliefs about general technology and businesses that use a specific technology (i.e., AI) are likely to affect their attitude and behavioral intention toward businesses that use AI. Therefore, reflecting on the need to incorporate trust, distrust, and attitude into AI acceptance research (Kelly et al., 2023), this framework shifts the focus of TAM from understanding consumers' acceptance of a technology to understanding consumers' acceptance and behavioral intention toward businesses' use of AI technology (Davis, 1989). This approach highlights the significance of gaining and cultivating (a) trust in a specific technology (i.e., AI), (b) trust in businesses that use the technology, and (c) trust in how businesses use the technology.

2. Literature Review

At present, there is a handful of research articles that examine AI in the consumption process in relation to acceptance and trust toward specific applications such as chatbots and voice assistants (Jain et al., 2024). However, these studies often yielded mixed results. On the one hand, AI has its novelty and can solve problems beyond human capabilities; on the other hand, there are perceptions of risk (Hasan et al., 2021). Of note, there is still skepticism among consumers about how businesses use AI. Hence, Frank et al. (2023) shifted the focus from examining trust in AI to examining how trust in businesses affects trust in the businesses' AI adoption. They found that when businesses that are trusted give full autonomy for AI to make decisions, trust in the businesses' AI adoption was negatively affected. A confluence of factors is at play in shaping perceptions and acceptance of businesses' use of AI. To explain consumers' acceptance of businesses' use of AI, it is crucial to unpack the dynamics behind consumers' trust and distrust in the businesses' AI use as well as their attitude and behavioral intention toward the businesses that use AI in their operations.

2.1. Faith in General Technology, Trust and Distrust in Businesses' AI Use

Researchers have found that trust is a key factor in how people accept AI. It shows how people feel about AI's dependability, skill, and safety (Bitkina et al., 2020; Choung et al., 2023; Hasija & Esper, 2022; J. Kim et al., 2021). However, there is a difference between trust in AI and trust in businesses that use AI. While consumers experience services delivered by AI technologies, it is businesses that develop, deploy, and manage the technologies (Frank et al., 2023; Gillespie et al., 2023). Trust in AI is characterized as "a social contract of assumptions between humans and machines on how a system or algorithm will perform" (Mylrea & Robinson, 2023, p. 2). According to Frank et al. (2023), trust in companies is when people are willing to put themselves at risk for a trustee (like a company) because they believe the trustee will do something important to the trustors, like providing a service that meets or exceeds their expectations. When evaluating individuals' trust in a technology, four types of trust come into play: trust in people, trust in technology, faith in general technology, and trust in a specific technology (McKnight et al., 2009). First, trust in technology refers to "individuals depending on, or being willing to depend on, the technology to accomplish a specific task" (McKnight et al., 2009, p. 2). While both trust in people and trust in technology have the same contextual conditions (i.e., risk, uncertainty, and lack of total control), one significant difference between them is the lack of moral agency in technology-related trust (McKnight et al., 2009). In other words, trust in

technology relies on the technology's functionality necessary to complete a task (McKnight, 2005) and reliability to consistently operate (McKnight et al., 2002). Second, one's propensity to trust in general technology exists when "one assumes technologies are usually consistent, reliable, functional, and provide the help needed" (McKnight et al., 2009, p. 5), and it implies "one is willing to trust *technology across situations and persons*" (p. 7, emphasis added). And lastly, trust in a specific technology refers to "one's beliefs that the target technology has the capacity (i.e., features) to complete a required task" and indicates individuals' "willingness to depend on a specific technology in uncertain, risky situations" (McKnight et al., 2009, p. 7).

There is currently a "trust gap" or "trust deficit" in companies' adoption of AI, as there are persistent risks associated with businesses' use of AI (Chakravorti, 2024). Again, how businesses use AI technologies and what they use them for is the main cause of the trust gap. Thus, businesses are advised to endorse trust-building initiatives by communicating how AI is used and what it is used for (Frank et al., 2023). Distrust is not equivalent to the absence of trust and is characterized as "the active expectation that the other party will behave in a way that violates one's welfare and security" (Cho, 2006, p. 26). This study adopts Cho's (2006) approach to measure trust and distrust separately. This approach was built on the empirical evidence that distrust often influences behavioral intentions more than trust (Cho, 2006). This study posits that faith in general technology is likely to be positively associated with trust in businesses that use AI while negatively associated with distrust in businesses that use AI, based on McKnight et al.'s (2009) study suggesting that individuals' trust in the attributes of a certain technology can be translated into attitudes and intentions of a specific technology use.

McKnight et al. (2009) conceptualized that a connection exists between faith in general technology and trust in a specific technology. Without users' willingness to depend on technology in general, it would be difficult to build trust in any type of technological advances, including AI technology. Lack of faith in general technology may exacerbate individuals' skepticism and concerns about AI risks. New technologies like AI create anxiety and distrust due to their unpredictability (Edelman, 2019). Considering mixed sentiments and attitudes toward AI (e.g., Gessl et al., 2019), this study posits that individuals with greater faith in technology tend to show higher trust in businesses' AI usage, as a general predisposition to trust technology often translates into trust in specific AI tools (McKnight et al., 2011). Empirical findings reveal that people with high faith in technology report significantly greater trust in AI-driven services (Zarifis & Fu, 2023). The social cognition theory says that people who trust technology might use a "machine heuristic," which means they think that automated systems are objective and accurate (Sundar & Kim, 2019). The trust transfer theory also states that previous positive experiences with a technology create a "reservoir of trust" extendable to new AI solutions (Glikson & Woolley, 2020). Together, these insights justify propositions to test associations between faith in technology and trust in businesses' use of AI. Conversely, this study also proposes that lower faith in technology corresponds to higher distrust of businesses' AI usage. Because faith in technology provides a baseline of trust, it mitigates the "confident negative expectations" that define distrust (Cho, 2006). Research suggests that individuals with low faith in technology demand more assurances before trusting a specific system, whereas those with high faith in technology are less prone to assume bias or harm (Lewicki et al., 1998; McKnight et al., 2011). As a result, they approach AI adoption with fewer suspicions, though not necessarily blindly.

Therefore, the following hypotheses are posited:

H1: Faith in general technology is positively associated with trust in businesses that use AI.

H2: Faith in general technology is negatively associated with distrust in businesses that use AI.

2.2. Perceived Benefits

Current research has identified a host of unique individuals' characteristics as significantly influencing AI acceptance (Kelly et al., 2023). Specifically, perceived benefits (also known as perceived usefulness) are often considered an antecedent that explains other perceptual variables (Choung et al., 2023; Kelly et al., 2023). Perceived benefits are defined as "beliefs about the positive outcomes associated with a behavior in response to a real or perceived threat" (Chandon et al., 2000; Liu et al., 2013). Perceived benefits are equivalent to perceived usefulness, which is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance" (Davis, 1989, p. 320). According to the TAM, perceived usefulness has been consistently found to be strongly significant in predicting usage, indicating that users are driven to adopt a technology because of its functions (Davis, 1989). In the context of AI adoption, perceived usefulness is conceptualized as the degree to which individuals believe that using AI will enhance their performance or provide benefits (Choung et al., 2023). It has been found as a prominent variable in influencing AI acceptance (Choung et al., 2023; Del Giudice et al., 2023; Ismatullaev & Kim, 2024; Kelly et al., 2023). In the context of individuals' use of AI, Gansser and Reich (2021) have identified four dimensions of perceived benefits, namely health, convenience (comfort), sustainability, and performance expectancies (such as increasing productivity). These factors were found to predict behavioral intentions to use products containing AI.

While Bedué and Fritzsche (2022) suggested that trust building is required to increase AI adoption, Choung et al. (2023) found the indirect effect of trust and the effects of perceived usefulness, ease of use, and attitude on intention to use AI. Rossi (2018) argued:

> Trust in the technology should be complemented by trust in those producing the technology. Yet, such trust can only be gained if companies are transparent about their data usage policies and the design choices made while designing and developing new products. (p. 130)

Multiple empirical studies confirm that trust in a business or its AI technology can heighten consumers' perceptions of usefulness, ease of use, and overall benefits. In the context of the TAM, trust can act as an antecedent or moderator that strengthens perceived usefulness and reduces uncertainty. For instance, in AI voice assistants, Choung et al. (2023) found that higher trust in AI positively influences perceived usefulness and attitudes, which in turn boosts adoption intentions. Similarly, Gefen et al. (2003) demonstrated that in e-commerce, consumer trust in an online vendor was as influential as perceived usefulness and ease of use in predicting intended usage. Therefore, if consumers trust those businesses that use AI, they are likely to see the benefits of those companies' products/services that contain AI technologies.

While trust generally amplifies perceived benefits, the converse—distrust—can have the opposite effect. Distrust not only diminishes perceived usefulness but can also heighten perceived risk, causing consumers to focus on potential harm rather than benefits. Current literature indicates that individuals who actively distrust a business's use of AI are more likely to question the technology's performance, suspect hidden motives or data mismanagement, and anticipate negative outcomes (Cho, 2006). For example, in healthcare, low trust in "AI doctors" is associated with heightened risk perceptions and reduced benefit perceptions

(Kerstan et al., 2024). Moreover, a US Gallup survey found that widespread distrust in businesses' responsible use of AI corresponded with minimal belief in AI's net benefits (Price, 2023). Hence, if consumers distrust those businesses' intention behind using AI in their products/services, it is unlikely for them to perceive benefits from using AI-embedded products/services. Therefore, we postulated the following hypotheses:

> H3: Trust in businesses that use AI is positively associated with perceived benefits of products containing AI.

> H4: Distrust in businesses that use AI is negatively associated with perceived benefits of products containing AI.

In addition, based on the above literature review suggesting the role of trust in general and specific technology, we posit that trust and distrust will be mediating between faith in general technology and perceived benefits of products containing AI:

> H5: Trust and distrust mediate between faith in general technology and perceived benefits of products containing AI.

2.3. Attitudes

Current research on technology adoption has found conflicting results about the value of attitudes in predicting adoption intention or adoption (Yang & Yoo, 2004). On the one hand, perceived usefulness was found to be highly significant in influencing usage such that attitudes offer little value in predicting use (Davis et al., 1989). On the other hand, Yang and Yoo (2004) argued that attitude still deserves attention because it is a contagious social function that facilitates influence among people and that cognitive and affective attitudes influence usage differently. They stated: "Attitude is contagious and as people work together, they express their own and listen to each other's attitudes. Therefore, organizations and managers need to care about the positive attitude change" (Yang & Yoo, 2004). According to the TAM, attitude refers to users' assessments of the desirability of using a specific technology, reflecting either positive or negative feelings (Ajzen & Fishbein, 1980).

In the context of AI, individuals' attitudes toward the technology itself have been researched and measured (Grassini, 2023; Stein et al., 2024). But there is a lack of research on attitudes toward businesses that use AI. Unlike individuals' use of AI which could be explained by perceived usefulness (Gursoy et al., 2019), attitudes toward businesses that use AI could be influenced by trust (Frank et al., 2023). Only 28% of people in the US trust businesses that use AI models with customers ("AI has a trust problem," 2024). Therefore, even if one has a positive attitude toward AI use, they could have a negative attitude toward businesses' use of AI as a result of uncertainty about how businesses use AI or what they use it for. Because trust in a new technology is a precursor to acceptance (Dirsehan & Can, 2020), this study posits that trust in how businesses use AI influences consumers' attitudes toward businesses that use AI. Current research has also identified trust as a precursor to attitude toward a business's website (Limbu et al., 2012). A systematic review shows that trust and attitudes are equally important in AI acceptance (Kelly et al., 2023). The following hypothesis will be tested:

> H6: Perceived benefits are positively associated with attitudes toward businesses that use AI.

2.4. Behavioral Intentions

In line with current literature on technology adoption based on the theory of reasoned action and the TAM (Kelly et al., 2023), this study will test if one's attitudes toward businesses that use AI predict behavioral intentions toward those businesses (Davis et al., 1989). Behavioral intention measures "the strength of one's intention to perform a specified behavior" (Davis et al., 1989, p. 984). When people have positive feelings toward an action, they are likely to perform the action. The relationship between attitudes and behaviors is mediated by behavioral intentions, but researchers have advised of the importance of ensuring a match in the operationalizations of the attitude and behavioral intention constructs (Jaccard et al., 1977; M.-S. Kim & Hunter, 1993). For example, in the context of businesses' websites, Limbu et al. (2012) tested attitudes toward a business's website and behavioral intentions to purchase from that business's website. Hence, M.-S. Kim and Hunter (1993) noted "the importance of developing proper measures of attitude, intention, and behavior" to examine the factors determining behavioral inclinations (p. 354). Acknowledging this, this study will specifically examine the association between consumers' attitudes toward AI use by businesses and behavioral intentions to support businesses that use AI:

> H7: Attitude toward businesses that use AI is positively associated with behavioral intentions to support businesses that use AI.

While attitudes may play a mediator role between perceived AI benefits for individuals and behavioral intention (Ho et al., 2013; H6 and H7), to fully understand if this is a full mediation or a partial mediation, we also posit the following hypothesis to test:

> H8: Perceived benefits are positively associated with behavioral intention.

2.5. Approval of AI Use by Businesses

Although many benefits of AI have been proposed for both consumers and businesses, consumers have been reported to remain concerned and skeptical about businesses' use of AI. A survey conducted found that consumers were concerned about AI-generated product descriptions, AI-generated product reviews, chatbots answering questions, and AI being used for recommendations and personalized advertising (Haan & Watts, 2023b). On the other hand, other consumers in the survey believed that AI could improve personalized recommendations and advertising (Haan & Watts, 2023b). AI has enormous capabilities, but how and what businesses use it for can affect consumers' perceptions and eventually the overall reputations of the businesses (Enholm et al., 2022). Enholm et al. (2022) suggested that although AI technologies could improve businesses' operational, financial, and market-based sustainability performance, there could be unintended, negative consequences such as generating biased outcomes to benefit the businesses themselves but not their customers, ultimately costing their reputations. Businesses are advised to comprehend the unintended consequences of AI systems and to adopt responsible AI governance frameworks to create enhanced business value (Perifanis & Kitsios, 2023).

While consumers' adoption of AI technologies is a hot topic that has been extensively researched, consumers' interpretations and evaluations of businesses' use of AI require further examination (Jain et al., 2024). Current research has noted that even though consumers show a positive attitude toward AI

marketing communication, they also have a neutral or slightly negative feeling toward it depending on their perceptions of what businesses use AI for (Chen et al., 2022). On the one hand, businesses were found to use AI for positive purposes, such as using it for corporate social responsibility initiatives (Wu et al., 2024) and building relationships with stakeholders (Oh & Ki, 2024). Thus, if AI is used for good causes that benefit both businesses and stakeholders, then the use of AI by businesses is positively perceived. On the other hand, AI has its "dark sides": Consumers have concerns that the benefits of AI come at the expense of privacy (Cheng et al., 2022). For example, personalized recommendations lead to data security concerns (Cheng et al., 2022). As such, how AI is used influences consumers' cognitive and affective attitudes, requiring further investigation into the extent to which consumers are willing to approve businesses' AI use. Therefore, this study posits that individuals with positive attitudes toward products/services containing AI features are likely to approve businesses' use of AI in their operations. This proposition is built on the assumption that consumers' beliefs and attitudes toward the performance of AI-embedded products/services for oneself are transferrable to their beliefs in the performance of AI for businesses (Bitkina et al., 2020). Current analyses have listed numerous benefits of AI technologies for businesses, including improving customer experience and relationships, increasing productivity and sales, and saving costs (Haan & Watts, 2023a; Weitzman, 2022). Even if there exist positive attitudes toward businesses that use AI, accepting businesses' use of AI based on these benefits may be needed before individuals develop supportive behavioral intentions. Therefore, we postulate the following hypotheses:

H9: Perceived AI benefits for individuals are positively associated with approval of businesses' use of AI.

H10: Attitude toward businesses that use AI is positively associated with approval of businesses' use of AI.

H11: Approval of businesses' use of AI is positively associated with behavioral intention.

H12: Approval of businesses' use of AI mediates the relationship between attitudes and behavioral intention.

3. Methods

3.1. Development of Survey Instruments

To test the hypotheses, an online questionnaire was created based on existing studies. First, the measurement items for "faith in general technology" were adopted from McKnight et al. (2009). Second, the survey items for "perceived benefits for individuals" were adapted from the four dimensions (i.e., influence on health, influence on convenience, influence on sustainability, and performance expectancy) identified in Gansser and Reich's (2021) study on AI acceptance. Third, the survey items for "approval of businesses' use of AI" were developed based on industry articles that analyze the benefits of AI to businesses (Haan & Watts, 2023a; Weitzman, 2022). Fourth, trust and distrust were operationalized based on survey items in Cho's (2006) study on trust and mistrust. Lastly, attitude and behavioral intentions toward businesses that use AI were adapted from Wang et al. (2023). Table 1 shows a list of survey items used.

Table 1. A list of survey items used, Cronbach alpha (α), standardized loading, mean (M), standard deviation (SD), and standard error (SE).

Variable	Survey Item	Loading	*M*	*SD*	*SE*
Faith in general technology $\alpha = 0.889$	I believe that most technologies are effective at what they are posited to do.	0.827	3.86	0.854	0.042
	A large majority of technologies are excellent.	0.843	3.92	0.855	0.042
	Most technologies have features needed for their domains.	0.809	3.84	0.806	0.039
	I think most technologies enable me to do what I need to do.	0.787	3.97	0.080	0.039
Perceived AI benefits for individuals (health $\alpha = 0.929$	A product that contains AI can increase awareness of my health and well-being.	0.846	3.20	1.123	0.055
	A product that contains AI can provide me with information that helps me make better decisions about my health and well-being.	0.909	3.22	1.119	0.055
	A product that contains AI can give me more control over my health and well-being.	0.879	3.14	1.098	0.054
	A product that contains AI can increase my chances for a healthier lifestyle.	0.869	3.20	1.110	0.054
Perceived AI benefits for individuals (convenience) $\alpha = 0.907$	It is convenient that products that contain AI automatically control and check themselves.	0.816	3.36	1.076	0.052
	It is convenient that products that contain AI can control electrical devices by a simple operation.	0.795	3.32	1.100	0.054
	It is convenient that products that contain AI can provide access to a lot of information.	0.817	3.60	1.091	0.053
	It is convenient that products that contain AI can help me proactively and without human intervention.	0.831	3.33	1.153	0.056
	It is convenient that products that contain AI can help me make better decisions.	0.808	3.37	1.101	0.054
Perceived AI benefits for individuals (sustainability) $\alpha = 0.913$	People can use products with AI to manage waste better.	0.820	3.31	1.052	0.051
	People can use products with AI to save resources.	0.916	3.48	1.062	0.052
	People can use products with AI to achieve cost savings.	0.872	3.52	1.064	0.052
	People can use products with AI to know exactly how much resources they consume (time, money, etc.).	0.793	3.54	1.060	0.052
Perceived AI benefits for individuals (performance expectancy) $\alpha = 0.913$	Products with AI can help people get things done more quickly.	0.829	3.67	1.042	0.051
	Products with AI can increase people's productivity.	0.873	3.62	1.040	0.051
	Products with AI can increase people's chances of achieving things that are important.	0.867	3.46	1.057	0.052
	Products with AI are useful in everyday life.	0.823	3.59	1.084	0.053

Table 1. (Cont.) A list of survey items used, Cronbach alpha (α), standardized loading, mean (M), standard deviation (SD), and standard error (SE).

Variable	Survey Item	Loading	*M*	*SD*	*SE*
Approval of businesses' AI use $\alpha = 0.961$	Businesses should use AI for cost savings.	0.755	3.49	1.076	0.052
	Businesses should use AI to increase productivity.	0.824	3.59	1.038	0.051
	Businesses should use AI to improve daily operations.	0.834	3.61	1.057	0.052
	Businesses should use AI to maximize profits.	0.752	3.32	1.141	0.056
	Businesses should use AI to make better decisions.	0.836	3.55	1.095	0.053
	Businesses should use AI to reduce waste.	0.807	3.72	1.067	0.052
	Businesses should use AI to create more personalized shopping experiences for consumers.	0.827	3.44	1.141	0.056
	Businesses should use AI to gather customers' data to improve services.	0.751	3.20	1.150	0.056
	Businesses should use AI to improve relationships with customers.	0.841	3.35	1.152	0.056
	Businesses should use AI to understand the customer experience.	0.834	3.42	1.133	0.055
	Businesses should use AI to be more creative and innovative.	0.843	3.55	1.083	0.053
	Businesses should use AI to enhance the quality of its products and services.	0.851	3.61	1.099	0.054
	Businesses should maximize their use of AI.	0.761	3.22	1.149	0.056
Trust in businesses that use AI $\alpha = 0.902$	Businesses use AI in a highly dependable and reliable manner.	0.808	3.10	1.064	0.052
	Businesses are responsible and reliable in their use of AI.	0.824	3.23	1.117	0.054
	Businesses promote customers' benefits as well as their own in their use of AI.	0.881	3.20	1.108	0.054
	Businesses will not engage in any kinds of exploitative and damaging behaviors to customers through their use of AI.	0.828	3.25	1.098	0.054
Distrust in businesses that use AI $\alpha = 0.918$	Businesses exploit their customers' vulnerability through their use of AI.	0.827	3.43	0.051	1.049
	Businesses engage in damaging and harmful behaviors to customers to pursue their own interests through their use of AI.	0.882	3.31	0.049	1.001
	The way businesses use AI is irresponsible and unreliable.	0.856	3.25	0.050	1.017
	Businesses use AI in a deceptive and fraudulent way.	0.874	3.27	0.053	1.076
Attitudes $\alpha = 0.917$	Buying from businesses that use AI is a good idea.	0.881	3.13	1.043	0.051
	Buying from businesses that use AI is a wise idea.	0.920	3.07	1.039	0.051
	I feel positive about buying from businesses that use AI.	0.866	3.06	1.134	0.055
Behavioral intentions	I intend to buy from businesses that use AI more frequently.	—	2.88	1.168	0.057
	I am willing to spend more buying from businesses that use AI.	—	2.62	1.261	0.062

3.2. Data Collection

Upon approval from the first author's university's ethics committee (#2024-8995-20339), an online survey was administered to a nationally representative sample (by age and gender) of 456 Australian consumers in September 2024. The sample was recruited by Qualtrics. The respondents received remuneration based on their agreement with Qualtrics. After removing incomplete and straight-lining responses, 420 responses were retained for data analysis. Table 2 shows the demographic characteristics of the sample. The mean and standard deviation for age is 48.29 and 18.717, respectively.

Table 2. Demographic characteristics of the sample.

Individual-Level Variables	*N*	Percent
Age	420	
18–20	15	3.6%
21–30	78	18.6%
31–40	81	19.3%
41–50	66	15.7%
51–60	50	11.9%
61–70	59	14%
Above 70	71	16.9%
Gender		
Male	202	48.1%
Female	214	51%
Non-binary	3	0.7%
Other	1	0.2%
Education		
Less than high school	29	6.9%
High school graduate	89	21.2%
TAFE certificate or diploma	104	24.8%
Some university	21	5%
Bachelor's degree	134	31.9%
Master's degree	34	8.1%
Doctorate	6	1.4%
Other	3	0.7%
Annual pre-tax income		
Less than AUD 30,000	86	20.5%
AUD 30,001–60,000	113	26.9%
AUD 60,001–90,000	90	21.4%
AUD 90,001–120,000	48	11.4%
More than AUD 120,000	59	14%
Prefer not to answer	24	5.7%
Employment status		
Full-time	170	40.5%
Part-time	67	16%
Casual	20	4.8%
Not working	96	22.9%
Other	67	16%

Note: TAFE = Technical and Further Education.

3.3. Data Analysis

The data analysis process involved several steps. First, the data were analyzed using SPSS version 28. Specifically, the means, standard deviations, and standard errors for each item, as well as the reliability (Cronbach's alpha) for each variable, were calculated (as shown in Table 1). Second, as the measurement items were adapted or created based on existing studies, exploratory factor analysis (EFA) was run using principal component analysis (PCA) and Oblimin rotation, followed by confirmatory factor analysis (CFA) to check the validity of measurement items. The standardized loadings were reported in Table 1. Weighted composites were created based on these loadings. Third, as approval of businesses' use of AI was a newly conceptualized variable, the 13 items used were examined using PCA. The Kaiser-Meyer-Oklin value was 0.957 and Barlett's test of sphericity (Bartlett, 1954) reached statistical significance ($\chi^2 = 4887.918$, $df = 78$, $p < 0.001$). PCA analysis showed that there is one component with eigenvalue exceeding 1, explaining 68.213% of the variance. Lastly, structural equation modeling (SEM) was run on AMOS version 28 to test all the hypotheses in the hypothesized model. Age and gender were used as control variables. Hu and Bentler's (1999) joint criteria (CFI > 0.95, SRMR ≤ 0.10, or RMSEA < 0.06 and SRMR ≤ 0.10) were used to assess model fit. To test the mediation for H9, Holmbeck's (1997) procedure was adopted for testing three models, i.e., model with no mediator (Figure 1), model with full mediation (Figure 2), and model with partial mediation (Figure 3).

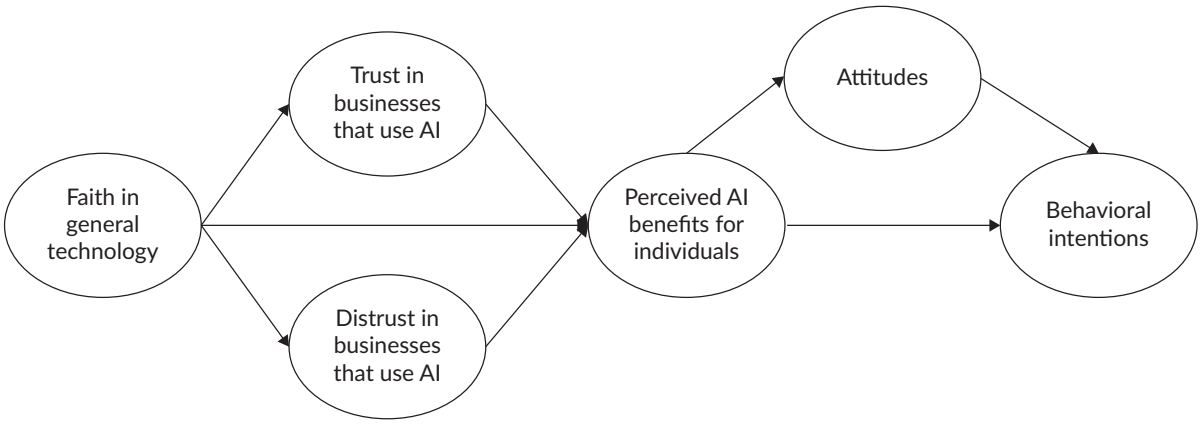


Figure 1. Model with no mediator.

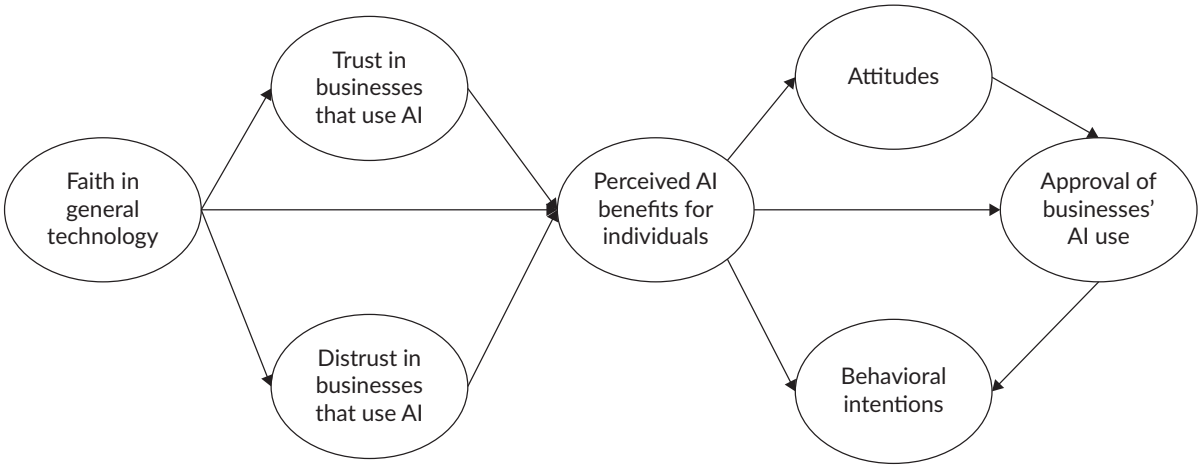


Figure 2. Model with full mediation.

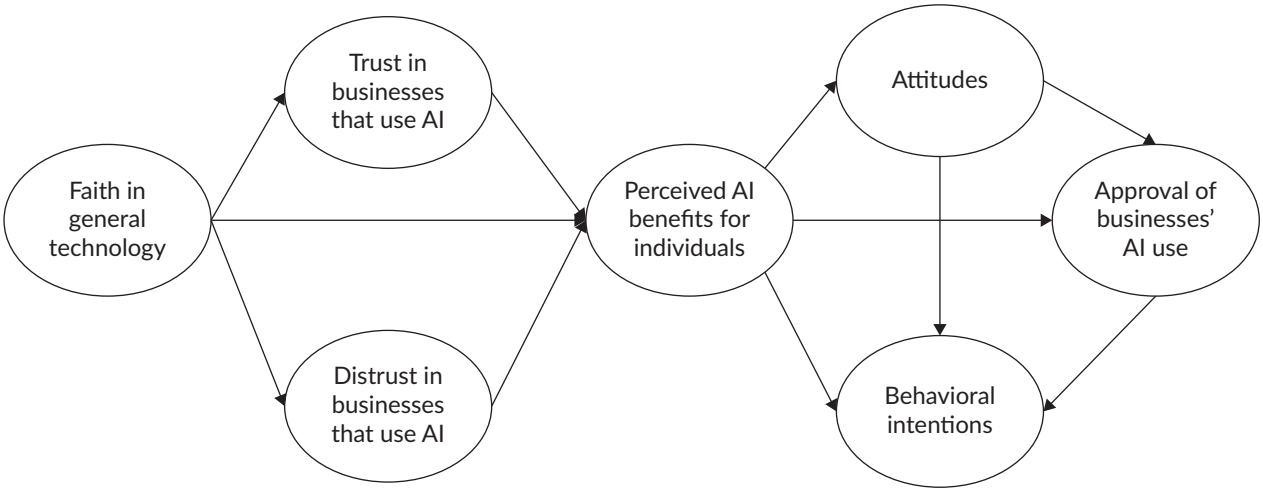


Figure 3. Model with partial mediation.

4. Results

Holmbeck's (1997) procedure for testing mediation was used to test the mediation hypothesized in H12. The first model (Figure 1), which does not have approval of businesses' AI use, showed a good model fit ($\chi^2 = 161.858$, $df = 37$, CFI = 0.961, RMSEA = 0.090, SRMR = 0.0373). The second model (Figure 2), with full mediation of approval of businesses' AI use between attitudes and behavioral intention, was found to have an acceptable fit ($\chi^2 = 259.298$, $df = 45$, RMSEA = 0.107, SRMR = 0.0412). Finally, the third model (Figure 3), with a partial mediation, resulted in a good model fit ($\chi^2 = 178.406$, $df = 44$, CFI = 0.964, RMSEA = 0.085, SRMR = 0.0359). Therefore, Figure 3 was accepted for testing hypotheses based on the fit indices. However, H12 was found insignificant (Figure 4).

Findings from the hypotheses tested are reported as follows. A positive association between faith in general technology and trust in businesses that use AI was found, so H1 was supported ($\beta = 0.348$, $p < 0.001$). In contrast, a negative association between faith in general technology and distrust was found (H2:

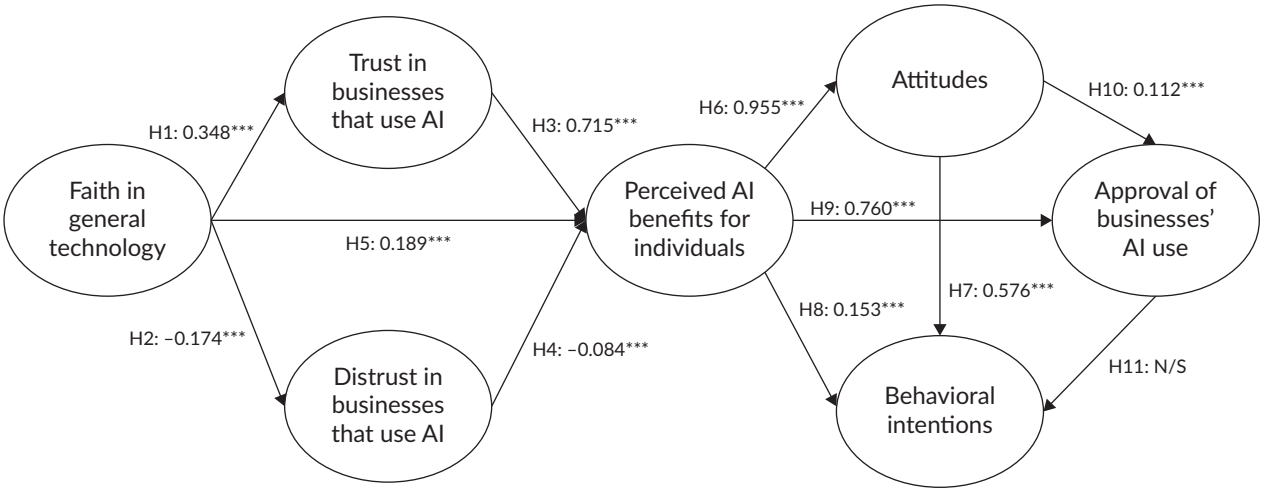


Figure 4. Results from the hypothesized model tested. Notes: N/S = non-significant; * $p < 0.05$, ** $p < 0.01$, $p < 0.001$; χ^2 [df] = 178.406[44]; CFI = 0.964; RMSEA = 0.085; SRMR = 0.0359.

$\beta = -0.174$, $p < 0.001$). Trust has a positive relationship with perceived benefits of products/services containing AI (H3: $\beta = 0.715$, $p < 0.001$) while distrust has a negative relationship with perceived benefits of products/services containing AI (H4: $\beta = -0.084$, $p < 0.001$). H5 predicting the mediating role of trust (H5a) and distrust (H5b) was tested; however, it turned out to be a partial mediation due to the path between faith in general technology and perceived AI benefits for individuals (H5: $\beta = 0.189$, $p < 0.001$). H6 predicting a positive relationship between perceived benefits and attitudes toward businesses that use AI was supported ($\beta = 0.955$, $p < 0.001$). As predicted, there was also a positive association between attitude toward businesses that use AI and behavioral intentions to support those businesses (H7: $\beta = 0.576$, $p < 0.001$). Perceived AI benefits for individuals were positively associated with behavioral intention (H8: $\beta = 0.153$, $p < 0.001$). Perceived AI benefits for individuals were also positively associated with approval of businesses' use of AI (H9: $\beta = 0.760$, $p < 0.001$). Attitude toward businesses that use AI is positively associated with approval of businesses' use of AI (H10: $\beta = 0.112$, $p < 0.001$). H11 predicting a positive relationship between approval of businesses' use of AI and behavioral intention was not supported. Therefore H12 suggesting the mediating role of approval of businesses' use of AI between attitudes and behavioral intention was not supported.

Regarding the effects of control variables, age and gender, there was a negative association between age and trust ($\beta = -0.294$, $p < 0.001$). There was also a negative relationship between gender and trust ($\beta = -0.100$, $p < 0.001$). Age and behavioral intention turned out a negative relationship ($\beta = -0.179$, $p < 0.001$). Finally, gender and approval turned out a negative association ($\beta = -0.065$, $p < 0.001$).

5. Discussion

In response to consumers' conflicting views about businesses' use of AI, this study tested a framework that examines the dynamics of faith in general technology, trust, distrust, and perceived benefits in influencing attitudes and behavioral intentions toward businesses that use AI. These dynamics also contributed to approval for businesses' use of AI. Current research has either examined consumers' acceptance of AI use for themselves (Kelly et al., 2023) or consumers' perceptions of interactions with AI-enabled applications adopted by businesses (Jain et al., 2024). Such research often results in general recommendations being made, such as improving transparency and streamlining processes to foster acceptance (Frank et al., 2023; Gillespie et al., 2023). Notably, consumers are skeptical about businesses' use of AI because they are uncertain about what businesses use AI for and how they use it (Haan & Watts, 2023b).

This study found the crucial roles of faith in general technology and trust in businesses in influencing individuals' perceptions of AI benefits for themselves. These perceived benefits ultimately influence attitudes, approval of businesses' AI use, and behavioral intentions toward the businesses that use AI. As predicted, attitudes were positively associated with behavioral intention to support the businesses that use AI as well as with approval of businesses' AI use. However, approval of businesses' AI use does not translate into consumers' behavioral intention to support the businesses.

Our study redirects scholarly attention to the importance of building and cultivating trust in businesses. Earlier we noted that there is a "trust gap" due to the risks associated with businesses' use of AI (Chakravorti, 2024). Our study departs from existing studies that focus on trust in and acceptance of specific technology features. To make consumers understand and accept a specific technology's functionality and benefits,

businesses ought to earn consumers' trust in their operations. While trust in a specific technology is important in terms of ensuring reliability and consistency in delivering a task, as technology evolves, focusing on task-oriented trust may be myopic. As it is businesses that develop, manage, and use certain technologies in their products/services, trust in those businesses *across situations and technologies* is essential. Trust in businesses is fundamental in businesses' relationships with customers for the long term. Thus, businesses are advised to endorse trust-building initiatives by communicating how AI is used and what it is used for (Frank et al., 2023).

Our study also found that faith in general technology serves as an antecedent to trust and distrust in businesses that use AI as a specific technology. While the associations among faith in general technology, trust, perceived benefits, attitudes, approval of businesses' AI use, and behavioral intention may show linear patterns toward AI optimism, the relationship between faith in general technology and distrust shows that when individuals do not tend to believe in general technologies, it will be challenging to address individuals' skepticism, anxiety, and distrust in a specific technology. Individuals who distrust businesses that use AI are likely to be more doubtful about AI benefits from products/services that contain AI features. From a communication perspective, future research should examine how variations in how businesses communicate their use of AI influence trust, attitudes, and behavioral intentions.

From a theoretical perspective, this study advances current research on technology adoption by assessing consumers' perceptions of technology adoption by businesses. Because businesses are a third party in control of developing, deploying, and managing AI (Frank et al., 2023; Gillespie et al., 2023), the framework tested shows the significance of trust and distrust in businesses in influencing support for businesses that use AI. Individuals may be able to assess the perceived usefulness and perceived ease of use of a new technology if they are in control. But unless businesses clearly disclose how they use AI, individuals' behavioral inclinations toward their use of AI will be dependent on trust in how businesses use AI. Future research may consider integrating perceived use and trust into existing theoretical frameworks such as the theory of reasoned action and the theory of technology acceptance.

6. Limitations and Future Directions

The limitations of this study are as follows. First, the study is tested in the context of consumers' perceptions toward businesses' general use of AI. It is possible that consumers have varying degrees of perceptions and behavioral inclinations toward specific businesses (Frank et al., 2023). Thus, future research could examine if individuals' perceptions about specific businesses influence their perceptions toward how they use AI. Second, constructs such as approval of businesses' use of AI were newly conceptualized and operationalized. Even though the conceptualizations and operationalizations were developed based on existing research, future studies should refine these constructs and measures. Third, this study has not tested existing frameworks such as the theory of reasoned action or the theory of technology acceptance (Kelly et al., 2023) in full; instead, it tested a framework that is perceived to be suitable for the research context. Future studies should explore the possibilities of using an existing framework to examine consumers' perceptions of AI use by businesses. Lastly, there is a host of antecedent variables that could be examined to extend the findings of this study, such as perceived risk (Liu et al., 2013) and privacy concerns and ethics (Mylrea & Robinson, 2023).

7. Conclusion

There is a plethora of research that has proposed different factors influencing consumers' acceptance of AI use. In the context of consumers' support for businesses that use AI, this study has identified the significance of trust and distrust as mediating factors between individuals' perceived AI benefits and behavioral inclinations. Moreover, consumers who are optimistic about the benefits of AI are generally also optimistic about how businesses use AI. Consumers are aware that AI use by businesses is inevitable, but how they use AI is often unregulated (Mylrea & Robinson, 2023). Future research should further theorize consumers' evaluation of businesses' AI use as a central variable in shaping acceptance of businesses' AI adoption.

Acknowledgments

The authors appreciate the valuable suggestions provided by the academic editors and reviewers. An AI writing assistant, QuillBot, was used to proofread the article.

Funding

Publication of this article in open access was made possible through the institutional membership agreement between the University of New South Wales and Cogitatio Press.

Conflict of Interests

The authors declare no conflict of interests.

References

AI has a trust problem. Here's how to fix it. (2024, September 4). *Harvard Business Review*. https://hbr.org/sponsored/2024/09/ai-has-a-trust-problem-heres-how-to-fix-it

Agnihotri, D., Kulshreshtha, K., & Tripathi, V. (2021). A study on firms' communication based on artificial intelligence and its influence on customers' complaint behavior in social media environment. *IOP Conference Series: Materials Science and Engineering*, *1116*(1), Article 012180. https://doi.org/10.1088/1757-899X/1116/1/012180

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Prentice-Hall.

Alhitmi, H. K., Mardiah, A., Al-Sulaiti, K. I., & Abbas, J. (2024). Data security and privacy concerns of AI-driven marketing in the context of economics and business field: An exploration into possible solutions. *Cogent Business & Management*, *11*(1), Article 2393743.

Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society*: *Series B (Methodological)*, *16*(2), 296–298.

Bedué, P., & Fritzsche, A. (2022). Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management*, *35*(2), 530–549. https://doi.org/10.1108/JEIM-06-2020-0233

Bitkina, O. V., Jeong, H., Lee, B. C., Park, J., Park, J., & Kim, H. K. (2020). Perceived trust in artificial intelligence technologies: A preliminary study. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *30*(4), 282–290. https://doi.org/10.1002/hfm.20839

Chakravorti, B. (2024, May 3). AI's trust problem: Twelve persistent risks of AI that are driving skepticism. *Harvard Business Review*. https://hbr.org/2024/05/ais-trust-problem

Chandon, P., Wansink, B., & Laurent, G. (2000). A benefit congruency framework of sales promotion effectiveness. *Journal of Marketing*, *64*(4), 65–81. https://doi.org/10.1509/jmkg.64.4.65.18071

Chen, H., Chan-Olmsted, S., Kim, J., & Mayor Sanabria, I. (2022). Consumers' perception on artificial intelligence applications in marketing communication. *Qualitative Market Research*, *25*(1), 125–142. https://doi.org/10.1108/QMR-03-2021-0040

Cheng, X., Lin, X., Shen, X. L., Zarifis, A., & Mou, J. (2022). The dark sides of AI. *Electronic Markets*, *32*(1), 11–15. https://doi.org/10.1007/s12525-022-00531-5

Cho, J. (2006). The mechanism of trust and distrust formation and their relational outcomes. *Journal of Retailing*, *82*(1), 25–35. https://doi.org/10.1016/j.jretai.2005.11.002

Choung, H., David, P., & Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982–1003.

Del Giudice, M., Scuotto, V., Orlando, B., & Mustilli, M. (2023). Toward the human-centered approach. A revised model of individual acceptance of AI. *Human Resource Management Review*, *33*(1), Article 100856. https://doi.org/10.1016/j.hrmr.2021.100856

Dirsehan, T., & Can, C. (2020). Examination of trust and sustainability concerns in autonomous vehicle adoption. *Technology in Society*, *63*, Article 101361. https://doi.org/10.1016/j.techsoc.2020.101361

Edelman. (2019). *2019 Edelman AI survey*. https://www.edelman.com/research/2019-artificial-intelligence-survey

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, *24*(5), 1709–1734. https://doi.org/10.1007/s10796-021-10186-w

Exploring AI adoption in Australian businesses. (2024, December 17). *Australian Government Department of Industry, Science and Resources*. https://www.industry.gov.au/news/exploring-ai-adoption-australian-businesses

Frank, D. A., Jacobsen, L. F., Søndergaard, H. A., & Otterbring, T. (2023). In companies we trust: Consumer adoption of artificial intelligence services and the role of trust in companies and AI autonomy. *Information Technology and People*, *36*(8), 155–173. https://doi.org/10.1108/ITP-09-2022-0721

Gansser, O. A., & Reich, C. S. (2021). A new acceptance model for artificial intelligence with extensions to UTAUT2: An empirical study in three segments of application. *Technology in Society*, *65*, Article 101535. https://doi.org/10.1016/j.techsoc.2021.101535

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, *27*(1), 51–90. https://doi.org/10.2307/30036519

Gerlich, M. (2023). Perceptions and acceptance of artificial intelligence: A multi-dimensional study. *Social Sciences*, *12*(9), Article 502. https://doi.org/10.3390/socsci12090502

Gessl, A. S., Schlögl, S., & Mevenkamp, N. (2019). On the perceptions and acceptance of artificially intelligent robotics and the psychology of the future elderly. *Behaviour and Information Technology*, *38*(11), 1068–1087. https://doi.org/10.1080/0144929X.2019.1566499

Gillespie, N., Curtis, C., Pool, J., & Lockey, S. (2023, February 23). A survey of over 17,000 people indicates only half of us are willing to trust AI at work. *The Conversation*. https://theconversation.com/a-survey-of-over-17-000-people-indicates-only-half-of-us-are-willing-to-trust-ai-at-work-200256

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Grassini, S. (2023). Development and validation of the AI attitude scale (AIAS-4): A brief measure of general attitude toward artificial intelligence. *Frontiers in Psychology*, *14*, Article 1191628. https://doi.org/10.3389/fpsyg.2023.1191628

Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, *49*, 157–169. https://doi.org/10.1016/j.ijinfomgt.2019.03.008

Haan, K., & Watts, R. (2023a, April 24). How businesses are using artificial intelligence. *Forbes*. https://www.forbes.com/advisor/business/software/ai-in-business

Haan, K., & Watts, R. (2023b, July 20). Over 75% of consumers are concerned about misinformation from artificial intelligence. *Forbes*. https://www.forbes.com/advisor/business/artificial-intelligence-consumer-sentiment

Hajkowicz, S. A., Karimi, S., Wark, T., Chen, C., Evans, M., Rens, N., Dawson, D., Charlton, A., Brennan, T., Moffatt, C., Srikumar, S., & Tong, K. J. (2019). *Artificial intelligence: Solving problems, growing the economy and improving our quality of life*. Commonwealth Scientific and Industrial Research Organisation. https://www.csiro.au/-/media/D61/Reports/AI-Roadmap/19-00346_DATA61_REPORT_AI-Roadmap-_7_.pdf

Hasan, R., Shams, R., & Rahman, M. (2021). Consumer trust and perceived risk for voice-controlled artificial intelligence: The case of Siri. *Journal of Business Research*, *131*, 591–597. https://doi.org/10.1016/j.jbusres.2020.12.012

Hasija, A., & Esper, T. L. (2022). In artificial intelligence (AI) we trust: A qualitative investigation of AI technology acceptance. *Journal of Business Logistics*, *43*(3), 388–412. https://doi.org/10.1111/jbl.12301

Ho, L.-H., Hung, C.-L., & Chen, H.-C. (2013). Using theoretical models to examine the acceptance behavior of mobile phone messaging to enhance parent–teacher interactions. *Computers & Education*, *61*, 105–114. https://doi.org/10.1016/j.compedu.2012.09.009

Holmbeck, G. N. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: Examples from the child-clinical and paediatric psychology literatures. *Journal of Consulting and Clinical Psychology*, *65*(4), 599–610.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hyken, S. (2024, August 11). Beyond the hype: What 1,000 U.S. customers really think about AI. *Forbes*. https://www.forbes.com/sites/shephyken/2024/08/11/beyond-the-hype-what-1000-us-customers-really-think-about-ai

Ismatullaev, U. V. U., & Kim, S. H. (2024). Review of the factors affecting acceptance of AI-infused systems. *Human Factors*, *66*(1), 126–144. https://doi.org/10.1177/00187208211064707

Jaccard, J., King, G. W., & Pomazal, R. (1977). Attitudes and behavior: An analysis of specificity of attitudinal predictors. *Human Relations*, *9*, 817–824.

Jain, V., Wadhwani, K., & Eastman, J. K. (2024). Artificial intelligence consumer behavior: A hybrid review and research agenda. *Journal of Consumer Behaviour*, *23*(2), 676–697. https://doi.org/10.1002/cb.2233

Jiang, P., Niu, W., Wang, Q., Yuan, R., & Chen, K. (2024). Understanding users' acceptance of artificial intelligence applications: A literature review. *Behavioral Sciences*, *14*(8), Article 671. https://doi.org/10.3390/bs14080671

Jobanputra, K. (2024, August 22). Customer service: How AI is transforming interactions. *Forbes*. https://www.forbes.com/councils/forbesbusinesscouncil/2024/08/22/customer-service-how-ai-is-transforming-interactions

Kelly, S., Kaye, S. A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*, Article 101925. https://doi.org/10.1016/j.tele.2022.101925

Kerstan, S., Bienefeld, N., & Grote, G. (2024). Choosing human over AI doctors? How comparative trust associations and knowledge relate to risk and benefit perceptions of AI in healthcare. *Risk Analysis*, *44*(4), 939–957. https://doi.org/10.1111/risa.14216

Kim, J., Giroux, M., & Lee, J. C. (2021). When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. *Psychology and Marketing*, *38*(7), 1140–1155. https://doi.org/10.1002/mar.21498

Kim, M.-S., & Hunter, J. E. (1993). Relationships among attitudes, behavioral intentions, and behavior. *Communication Research*, *20*(3), 331–364.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *The Academy of Management Review*, *23*(3), 438–458. https://doi.org/10.2307/259288

Limbu, Y. B., Wolf, M., & Lunsford, D. (2012). Perceived ethics of online retailers and consumer behavioral intentions: The mediating roles of trust and attitude. *Journal of Research in Interactive Marketing*, *6*(2), 133–154. https://doi.org/10.1108/17505931211265435

Liu, M., Brock, J. L., Shi, G., Chu, R., & Tseng, T. (2013). Perceived benefits, perceived risk, and trust: Influences on consumers' group buying behaviour. *Asia Pacific Journal of Marketing and Logistics*, *25*(2), 225–248. https://doi.org/10.1108/13555851311314031

McKnight, D. H. (2005). Trust in information technology. In G. B. Davis (Ed.), *The Blackwell encyclopedia of management* (Vol. 7, pp. 329–331). Blackwell.

McKnight, D. H., Carter, M., & Clay, P. (2009). Trust in technology: Development of a set of constructs and measures. *DIGIT 2009 proceedings* (Article 10). Association for Information Systems. http://aisel.aisnet.org/digit2009/10

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2), Article 12. https://doi.org/10.1145/1985347.1985353

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

Mylrea, M., & Robinson, N. (2023). Artificial intelligence (AI) trust framework and maturity model: Applying an entropy lens to improve security, privacy, and ethical AI. *Entropy*, *25*(10), Article 1429. https://doi.org/10.3390/e25101429

Oh, J., & Ki, E. J. (2024). Can we build a relationship through artificial intelligence (AI)? Understanding the impact of AI on organization–public relationships. *Public Relations Review*, *50*(4), Article 102469. https://doi.org/10.1016/j.pubrev.2024.102469

Oyekunle, D., Matthew, U. O., Preston, D., & Boohene, D. (2024). Trust beyond technology algorithms: A theoretical exploration of consumer trust and behavior in technological consumption and AI projects. *Journal of Computer and Communications*, *12*(6), 72–102. https://doi.org/10.4236/jcc.2024.126006

Perifanis, N. A., & Kitsios, F. (2023). Investigating the influence of artificial intelligence on business value in the digital era of strategy: A literature review. *Information*, *14*(2), Article 85. https://doi.org/10.3390/info14020085

Price, R. (2023, October 17). Americans are skeptical that AI will be used responsibly. *Digital Content Next*. https://digitalcontentnext.org/blog/2023/10/17/americans-are-skeptical-that-ai-will-be-used-responsibly

Rossi, F. (2018). Building trust in artificial intelligence. *Journal of International Affairs*, *72*(1), 127–134.

Sindermann, C., Yang, H., Elhai, J. D., Yang, S., Quan, L., Li, M., & Montag, C. (2022). Acceptance and fear of artificial intelligence: Associations with personality in a German and a Chinese sample. *Discover Psychology*, *2*(1), Article 8. https://doi.org/10.1007/s44202-022-00020-y

Stein, J. P., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M. (2024). Attitudes towards AI: Measurement and associations with personality. *Scientific Reports*, *14*(1), Article 2909. https://doi.org/10.1038/s41598-024-53335-2

Stryker, C., & Kavlakoglu, E. (2024). *What is AI?* IBM. https://www.ibm.com/topics/artificial-intelligence

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In S. Brewster & G. Fitzpatrick (Eds.), *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Article 538). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300768

Szleter, P. (2024, August 9). Five ways AI can transform how businesses interact with customers. *Forbes*. https://www.forbes.com/councils/forbestechcouncil/2024/08/09/five-ways-ai-can-transform-how-businesses-interact-with-consumers

Wang, C., Ahmad, S. F., Bani Ahmad Ayassrah, A. Y. A., Awwad, E. M., Irshad, M., Ali, Y. A., Al-Razgan, M., Khan, Y., & Han, H. (2023). An empirical evaluation of technology acceptance model for artificial intelligence in e-commerce. *Heliyon*, *9*(8), Article e18349. https://doi.org/10.1016/j.heliyon.2023.e18349

Weitzman, T. (2022, November 22). The top five ways AI is transforming business. *Forbes*. https://www.forbes.com/councils/forbesbusinesscouncil/2022/11/21/the-top-five-ways-ai-is-transforming-business

Wu, L., Chen, Z. F., & Tao, W. (2024). Instilling warmth in artificial intelligence? Examining publics' responses to AI-applied corporate ability and corporate social responsibility practices. *Public Relations Review*, *50*(1), Article 102426. https://doi.org/10.1016/j.pubrev.2024.102426

Yang, H. D., & Yoo, Y. (2004). It's all about attitude: Revisiting the technology acceptance model. *Decision Support Systems*, *38*(1), 19–31. https://doi.org/10.1016/S0167-9236(03)00062-9

Zarifis, A., & Fu, S. (2023). Re-evaluating trust and privacy concerns when purchasing a mobile app: Re-calibrating for the increasing role of artificial intelligence. *Digital*, *3*(4), 286–299. https://doi.org/10.3390/digital3040018

About the Authors

Lisa Tam (PhD, Purdue University) is a senior lecturer at the Queensland University of Technology, Australia. Her research explores the dynamics of power and influence between organizations and publics. Her research has conceptualized and operationalized new constructs, including country-of-origin relationship (CoOR), power mutuality, power discrepancy, and conspiratorial thinking.

Soojin Kim (PhD, Purdue University) is an associate professor at the University of New South Wales, Australia. Her research investigates the behavior of publics and stakeholders to find connections between public/stakeholder insights and organizations' strategies for facilitating meaningful engagement and collaboration. She explores diverse organizational-public relationship contexts.

Yi Gong is a researcher in public relations at the Queensland University of Technology's Business School in Brisbane, Australia. Her research examines situational and dispositional subjective norms shaping young adults' engagement with government communication.

Media and Communication
2025 • Volume 13 • Article 9644
https://doi.org/10.17645/mac.9644

ARTICLE

Open Access Journal

Exploring the Challenges of Generative AI on Public Sector Communication in Europe

Alessandro Lovari [1] and Fabrizio De Rosa [2]

[1] Department of Political and Social Sciences, University of Cagliari, Italy
[2] Independent Researcher, Italy

Correspondence: Alessandro Lovari (alessandro.lovari@unica.it)

Submitted: 17 November 2024 **Accepted:** 5 February 2025 **Published:** 14 May 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

This study investigates how emerging digital technologies, particularly generative AI tools, are transforming public sector communication in Europe, highlighting the profound intersection between public organizations, AI, and human interactions. In particular, it explores the opportunities and risks that public sector communicators face as they deal with and integrate digital platforms and AI-driven tools into their strategies and practices in a contemporary scenario characterized by the spread of disinformation and a growing distrust toward institutions. The article gathers insights from in-depth interviews with leading public sector communicators working for European governments and EU institutions. Findings reveal that generative AI is seen as a transformative tool for governments and public institutions, with communicators emphasizing both benefits and risks, as well as the importance of adopting ethical practices and new responsibilities toward citizens, institutions, and mass media. From the interviews, generative AI tools emerged as game-changers in message delivery and content production, demanding greater professionalism and new competencies and skills to integrate these technologies into public sector communication strategies and to counteract the threats posed by disinformation campaigns and platformization. The study provides valuable insights into the evolving role of generative AI in public-sector communication, addressing the scarcity of research in this field. As the adoption of generative AI becomes inevitable, and policy frameworks like the EU AI Act develop, communicators must ensure transparency and trust to align public sector communication with democratic values and foster meaningful dialogue in new digital-media arenas. Implications for theory and practice are discussed.

Keywords

communicators; disinformation; ethics; generative AI; public communication; public sector communication; trust

1. Introduction

In the post-Covid-19 era, the role of public sector communication (Canel & Luoma-aho, 2018) has undergone profound transformations due to the turbulence of socio-political scenarios, the spread of disinformation and misinformation, the rapid technological advancements, and the increasing influence of digital platforms and AI on private and public organizations (OECD, 2021; van Dijck et al., 2018; Zerfass et al., 2024; Zerfass et al., 2023). In the European context, social media platforms are increasingly considered the primary sources for news consumption by many citizens. A Eurobarometer (2023) survey reveals that 37% of European citizens stay informed regularly by checking their newsfeeds on social media, an 11% increase compared to the 2022 edition of the report. At the same time, 42% of respondents read news on online media websites and apps (European Parliament, 2023b). These trends span all age groups and most of the EU member states, highlighting a broad shift towards online information and digital media arenas (Badham et al., 2024) that also impacts governments and public sector organizations (PSOs) as well as their communication staff.

Furthermore, the growing relevance and penetration of generative AI (GenAI) tools have the potential to become a game-changer also in the public sector (van Noordt & Misuraca, 2022). This technology is a double-edged sword (Zerfass et al., 2024; Zuiderwijk et al., 2021): On the one hand, these tools can offer new opportunities for improving public services and enhancing public communication efforts (Dwivedi et al., 2023; Larsen & Følstad, 2024) but, on the other hand, it can also pose unprecedented ethical challenges and risks of amplifying disinformation potentially threatening the foundation of informed public debate and democratic decision-making globally (Jungherr & Schroeder, 2023; World Economic Forum, 2024).

In this context, within the theoretical framework of public sector communication (Canel & Luoma-aho, 2018; Luoma-aho & Canel, 2020), this study aims to investigate the impact of GenAI tools through the voices of prominent EU institutions and governments' communicators in order to identify new responsibilities, challenges, risks, and ethical considerations in the relationship with citizens and the media. In a field constantly reshaped by rapid technological innovation, this research explores issues at the intersection of institutions, AI, and human interaction. Through in-depth interviews, the article recognizes these actors as strategic players in the development of public communication strategies at national and European levels, shaping the international landscape in public communication.

2. Literature Review

In recent years, numerous scholars worldwide have emphasized that public sector communication has regained significant importance, particularly in the wake of the Covid-19 pandemic (Lovari et al., 2020; OECD, 2021). This renewed centrality was particularly evident in public institutions and governments where communication was strategically planned and delivered to diverse publics through a multichannel and multilevel approach, with digital platforms playing a crucial role in both ordinary and crisis situations (Coombs, 2020; Lovari & Belluati, 2023). These communication flows enabled institutions not only to inform citizens and the media swiftly but also to actively listen and respond with reliable information, thereby upholding public values and safeguarding common goods. This process has become more visible in contemporary media ecologies, characterized by hybrid media systems and the growing power of digital platforms (Chadwick, 2013; van Dijck et al., 2018).

The OECD (2021) highlighted that when public sector communication is strategically managed, it becomes foundational for enhancing democratic processes, promoting citizen participation, public decision-making, and building trust in institutions. This perspective underscores the idea that public sector communication should not be viewed merely as a function of message transmission, but rather as a vital mechanism for public policy development and civic engagement (Canel & Luoma-aho, 2018). Consequently, governments and public institutions are encouraged to invest in communication as a means of building relationships with stakeholders and a strategic leverage for fostering mutual understanding and collaboration, especially in times of global threats (OECD, 2021). Indeed, PSOs face exceptional challenges due to a complex interplay of factors. First of all, societies across the globe are experiencing an increasing level of skepticism toward public institutions. According to the Edelman Trust Barometer (2023), governments are far less trusted than private companies worldwide. The report reveals a significant trust disparity between governments and private companies, showing companies leading with a 62% trust level, 11 points higher than the government at 51%. Distrust in government is notable across 16 of 28 surveyed countries, with trust in media and governments ranked low due to perceived unethical behavior and incompetence (Edelman, 2023). Furthermore, the report shows that 46% of citizens see governments as a source of misleading rather than trustworthy information. This perception directly impacts the effectiveness of public communication strategies, emphasizing the necessity for transparency, accuracy, and public trust building as central themes in the evolving role of public communicators (OECD, 2021).

The current landscape is further complicated by the polarization of public opinion and the spread of disinformation on digital platforms, requiring public communication professionals to manage a delicate balance of maintaining public trust while effectively communicating in an increasingly skeptical and polarized society (World Economic Forum, 2024). Indeed, disinformation and misinformation can significantly influence and distort the public debate, leading to misinformed opinions, polarized discussions, and eroded trust in the public sphere (Kim & Gil de Zúñiga, 2020; Lovari, 2020). This trend is part of a larger phenomenon, the so-called information disorder (Wardle & Derakhshan, 2017). While the historical effects of rumors and false content are well-known, Wardle and Derakhshan (2017, p. 4) argue that:

> We are witnessing something new: information pollution at a global scale; a complex web of motivations for creating, disseminating and consuming these 'polluted' messages; a myriad of content types and techniques for amplifying content; innumerable platforms hosting and reproducing this content.

Moreover, the Covid-19 pandemic marked a turning point in how disinformation is perceived in the public discourse. Once considered a marginal issue or a topic for a limited group of experts (i.e., journalists, fact-checkers), it has now become a public problem (Gusfield, 1981) and a key item on the agendas of governments at the international level. Information disorder and the diffusion of polluted messages are particularly prevalent in the context of digital media. In fact, the algorithms and the inherent characteristics of digital platforms can amplify these harmful narratives (Benesch, 2023; UN Interregional Crime and Justice Research Institute, 2020). Overall, the World Economic Forum has highlighted the growing disinformation on digital media as a major global threat. According to the report, this phenomenon "is emerging as the most severe global risk anticipated over the next two years, foreign and domestic actors alike will leverage misinformation and disinformation to further widen societal and political divides" (World Economic Forum, 2024, p. 7).

Moreover, digital platforms today represent the connective tissue of contemporary society, and their impact is also clearly visible in public sector and public sector communication (Lovari & Valentini, 2020). The expansion of digital connectivity underscores the evolving landscape of public communication and the critical role of digital proficiency for public information professionals. In fact, digital platforms offer governments and institutions new communicative environments for informing citizens directly, for enhancing citizens' participation and policy development, and for being more transparent and accountable to public opinion and mass media (Haro-de-Rosario et al., 2018; Lovari & Valentini, 2020; Silva et al., 2019). Therefore, they play an important role in redefining relationships, power dynamics, and communication strategies involving PSOs and their stakeholders.

However, several scholars have highlighted that PSOs are increasingly dependent on digital platforms and social media logic (Olsson & Eriksson, 2016; van Dijck & Poell, 2013). Indeed, the influence of these platforms on the visibility of public sector communication, combined with the non-transparent management of data and the opacity of associated algorithms, favors processes of datafication and platformization (Helmond, 2015; Reutter, 2022). Platformization refers to the process and effects of digital platforms' impact on contemporary society. In particular, van Dijck (2020) stressed how this process can pose challenges for the public sector, with the risk of treating common goods, such as health or education, as privatized assets encapsulated within the platform ecosystem that operates on market logics, following the principles of digital capitalism. All these factors directly impact public sector communication practices and citizens' consumption patterns. In this regard, digital platforms can also create inequalities and pose vulnerabilities for public communication when communicators fail to comprehend the complexities and potential issues associated with their superficial application or inaccurate use (Ducci & Lovari, 2021).

In this context, the rapid penetration of GenAI can pose both novel and traditional challenges for the public sector and public sector professionals.

2.1. AI in the Public Sector and the Impact on Communication Management

AI is defined as systems that display intelligent behavior by analyzing their environment and taking actions—with some degree of autonomy—to achieve specific goals (European Commission, 2025). Gil de Zúñiga et al. (2024, p. 317) define AI as "the tangible real-world capability of non-human machines or artificial entities to perform, task solve, communicate, interact, and act logically as it occurs with biological humans." This definition addresses the strong influence of AI on communication research, as well as its impact on their sectors and informative practices (Ertem-Eray & Cheng, 2025). Indeed, the topic of AI is currently experiencing significant hype and rapid technological advancements across various social domains (Audétat, 2022), spanning from health to education, cultural practices, science, and financial services (Kennedy et al., 2023). The so-called "AI spring," characterized by high expectations for these technologies, seems to have bloomed into "AI summer," in which AI tools are widely used, meeting the expectations of different stakeholders (Toll et al., 2020).

AI has also become a salient issue at the policy level, with initiatives emerging at transnational, international, and national levels (European Commission, 2023a; OECD, 2024). It is increasingly recognized as a promising technology (Konrad et al., 2016), incorporating and translating diverse narratives and storytelling approaches by heterogeneous actors. These narratives frame AI in various ways, ranging from apocalyptic visions to sustainability-focused perspectives and its role as a "service" for humanity. For instance, Tzachor et al.

(2020) pointed out the importance of using AI tools for pandemic prevention and response, as well as to counteract online disinformation surrounding Covid-19.

The rapid success and adoption of AI tools are closely tied to the development of GenAI, a technology that generates new content in response to prompts. Indeed, GenAI is a category of AI that can create new content, such as texts, images, and videos, through text-to-image generators and Large Language Models. These systems are "developing fresh, human-like material that can be engaged with and consumed, rather than just numerical forecasts or internal rules" (García-Peñalvo & Vázquez-Ingelmo, 2023, p. 7).

The increased visibility and media coverage of GenAI tools have spotlighted AI's capabilities, leading to significant academic and public debate (Lorenz et al., 2023). Indeed, this technology is revolutionizing various sectors, including education, healthcare, journalism, and communication (Esposito, 2021; Gil de Zúñiga et al., 2024). The use of AI tools and technologies has also expanded in PSOs and governments, although its adoption has been slower than in the private sector (Bowen, 2024; Desouza et al., 2020; Madan & Ashok, 2023) due to the presence of specific organizational and legal barriers (Selten & Klievink, 2024). In fact, the adoption of AI in the public sector follows a different pace, driven by the specific challenges, objectives, and stakeholders' priorities inherent to PSOs (Kuziemski & Misuraca, 2020; van Noordt & Misuraca, 2022). This might be due to the fact that public sector entities prioritize values such as transparency, and equity, embedding these principles into the design and implementation of AI systems (Wang et al., 2024). Additionally, AI implementation demands accountability while navigating ethical and social barriers, particularly those linked to trust in AI technologies (Desouza et al., 2020; Zuiderwijk et al., 2021).

Notwithstanding this assumption, some authors have argued that in the present day, traditional public sector policy-making, service provision, and governance procedures can be rapidly transformed with the introduction of GenAI technologies (Bright et al., 2024; Salah et al., 2023).

Moreover, GenAI-based platforms are transforming how governments and PSOs inform and engage with citizens, enabling more personalized interactions that allow citizens to express concerns, provide feedback, and even participate in policy development (Pislaru et al., 2024). Indeed, digital tools such as chatbots, conversational agents, and AI-enabled forums can enhance government responsiveness, enhancing a dialogue that is more calibrated to diverse communities' needs and fostering citizens' trust (Dwivedi et al., 2023). In terms of implementation, conversational agents have struggled to establish a foothold within public sector contexts (Androutsopoulou et al., 2019; Zuiderwijk et al., 2021). Their adoption brings unique challenges, including the complexity of GenAI projects and the imperative to ensure transparency, fairness, and public value (Larsen & Følstad, 2024). Furthermore, large language model-enabled chatbots can mimic authoritative sources, generating voices and anthropomorphic imitations of humans, hallucinations and misleading answers that can intoxicate the public arena (Jungherr & Schroeder, 2023).

The impact of GenAI tools poses challenges for institutions, society, and individuals, underscoring the need to balance opportunities against potential risks (Dwivedi et al., 2023; World Economic Forum, 2024). In addition, the integration of these tools in the public sector's governance can optimize administrative workflows, minimizing inefficiencies, and allowing public officials and civil servants to focus on strategic and citizen-centered initiatives, increasing public value (Hjaltalin & Sigurdarson, 2024). However, this technological advancement also introduces potential risks such as data privacy, algorithmic bias, ethical and

legal issues, and accountability for technology misuse (Dwivedi et al., 2023; OECD, 2024). Interactive governance, therefore, requires not only active participation but also a balanced approach that integrates GenAI within ethical and cooperative frameworks, ensuring effective and transparent communication between government entities and citizens (Bowen, 2024). Recent research into these "AI tensions" aims to deepen understanding of the waves of adoption and diffusion of AI tools in PSOs (Madan & Ashok, 2023).

Another challenge and possible threat related to GenAI is the spread of disinformation. These tools can significantly strengthen the effectiveness of viral disinformation campaigns, making it easier to produce and disseminate highly tailored and convincing false information (Pérez Dasilva et al., 2021), targeting specific groups and sparking societal and political tensions (Ferrari et al., 2023). In fact, the 2024 World Economic Forum report finds that GenAI's role in creating realistic yet fabricated content, such as campaign videos, could lead to severe consequences, ranging from protests to radicalization, challenging the integrity of democratic processes (Formosa et al., 2024) and the stability of societies globally (World Economic Forum, 2024).

In January 2024, the President of the European Commission, Ursula von der Leyen, emphasized these challenges posed by GenAI in spreading disinformation, underscoring the necessity of upholding responsibilities by large internet platforms to manage the content they disseminate. These concerns have also been included in specific legislation related to AI platforms. The EU was the first supranational organization to develop a regulatory framework for AI "to make sure that AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly" (European Parliament, 2023a). The European Commission's initial proposal, developed in 2021, did not specifically address GenAI systems. However, the emergence of AI tools such as ChatGPT has prompted a revision to include such technologies (Ferrari et al., 2023). The legislation reached a political agreement on 8 December 2023 and approached the final approval stages in 2024 (Chee, 2024). Thus, the EU has positioned itself as a pioneer, understanding the importance of its role as global standard-setter (European Commission, 2023a). The "AI Act" aims to be a worldwide model for leveraging AI's advantages while mitigating its risks, such as job automation, the spread of online misinformation, and threats to national security, imposing new transparency obligations on major AI systems (Council of the EU, 2024; European Commission, 2025).

Society, organizations, and the job market will be radically transformed by the GenAI revolution, and these changes will need communicators to help adjust to these new realities (Haefner et al., 2020; Smillie & Scharfbillig, 2024; Zerfass et al., 2024). Despite the current hype surrounding GenAI, limited attention has been devoted to the relationship between AI and public sector communication at the international level. A systematic literature review by Ertem-Eray and Cheng (2025, p. 4) highlights this oversight, emphasizing the need for more comprehensive and multidisciplinary studies to clarify its applications to the communication field: "Analyzing research topics provides information about common and underrepresented topics that require further investigation." Indeed, even though communication represents a promising topic in scholars' research agenda related to AI and public sector ecologies (Jungherr & Schroeder, 2023; Zuiderwijk et al., 2021), only limited insights can be found in studies dedicated to public relations and strategic communication, where GenAI has sparked lively debate within the profession (McCorkindale, 2024; Panda et al., 2019; Smith & Waddington, 2023; Zerfass et al., 2024). Furthermore, initial insights have been presented in the field of communication professionals (Zerfass et al., 2020). For instance, Zerfass et al. (2019), in a survey involving 2,700 practitioners across Europe, found that three-quarters of communication

professionals believe that AI will change the communication profession as a whole. Communicator leaders forecast greater changes in the field of communication due to AI compared to their unit leaders or team members. The main challenges that emerged from the survey included securing the competencies of communication practitioners (56%), followed by addressing barriers related to various aspects of organizational infrastructure, such as ICT, budget, and responsibilities (54%). Interestingly, data showed that professionals working in governmental organizations rate competencies and organizational challenges for implementing AI higher than professionals working for private companies or agencies.

In this context, this article seeks to overcome the scarcity of studies specifically addressing the perception of the impact of GenAI tools in public sector communication, an under-researched topic in scholarly strategic communication, and contribute to addressing the "lack of understanding of the AI phenomenon within the public administration" (Madan & Ashok, 2023, p. 12). In particular, the study investigates how leading European communication professionals are navigating the contemporary digital communication ecologies and examines the main challenges and tensions posed by the increasing reliance on and influence of GenAI-driven communication. Thus, it responds to the recommendation to analyze how communication professionals in different media and communication sectors "can best realign their roles and relationships between the publics and technology in their work" (Ertem-Eray & Cheng, 2025, p. 14).

In particular, the research questions that guided this study are:

> RQ1: Which challenges do professionals identify regarding the implementation of GenAI in public communication management?
>
> RQ1a: How are communicators perceiving the impact of these digital technologies in their work?
>
> RQ2: What competencies should communicators develop in order to strategically manage public sector communication in the face of AI-driven technological advancements and socio-political turbulence?

3. Methodology

This study adopts a qualitative methodology to examine the emerging practices related to the use of AI in European governments and supranational organizations. Indeed, qualitative research is particularly suited to exploring new or multifaceted phenomena where variables are not easily quantifiable (Bryman, 2012), allowing for a deep understanding of perspectives, experiences, and contexts, which is essential when exploring a dynamic field such as GenAI in public sector communication.

In particular, the study was carried out using in-depth semi-structured interviews (Johnson, 2001) with elite publics (Hertz & Imber, 1995). These qualitative techniques were adopted to gather detailed data while allowing flexibility in the conversation to explore unexpected insights or themes that may arise. This approach is particularly suited to exploring the nuances and complexities inherent in the evolving role of public communicators amidst rapid contemporary social changes and technological disruptions, facilitating an in-depth exploration of professionals' perspectives. According to Hertz and Imber (1995), elite professionals, due to their positions, experiences, and insights, can provide a depth and richness of

information that is often unattainable from other sources. Their insights are not merely opinions but are grounded in their extensive experience and professional expertise, making them extremely relevant for understanding the current and future landscape of public sector communication, a field where understanding the strategic drivers and institutional frameworks is critical.

Elite publics were chosen using a snowball sampling technique (Biernacki & Waldorf, 1981) from the Club of Venice forum, founded in 1986 under the Italian Presidency of the Council of the EU. In particular, the Club of Venice is the "informal gathering of the Directors-General, Directors, and Heads of the information and communication services of the EU Member States and the EU Institutions" (Club of Venice, 2013). These high-level professionals are at the forefront of public sector communication in Europe. They have a deep understanding of the challenges and dynamics within this field due to their experience at the national and European levels, and they hold significant roles in communication and media relations within EU institutions and national governments.

These professionals were invited to participate in the study during the Club of Venice conference held in Venice in November 2023. They were contacted via email in January 2024; a second round of messages was sent if the first email was not replied to; other potential participants were recruited to enlarge the panel of experts using Club of Venice members who had already taken part in the study. In October 2024, a total number of 14 professionals were interviewed in English using Microsoft Teams. Participants include directors of communication in ministries, heads of communication units for national departments, heads of communication at EU institutions, and former heads of public information for governments and ministries of European countries. Interviewees reside in 13 different countries, with a good balance between Northern, Central, and Southern Europe and a wide range of cultural and professional perspectives (see Table 1).

The interview grid was developed from a literature review on public sector communication and PSO culture (Canel & Luoma-aho, 2018; OECD, 2021), and was structured in four sections (trends in public sector communication; digital platforms and GenAI challenges; strategies for fighting disinformation; communicators' identity and future landscapes). The interviews' average length was 42 minutes, a sufficient duration to address the complexities of the topic while being respectful of the interviewees' time constraints due to their professional roles. A thematic analysis was applied to interviews (Braun & Clarke, 2006), combining manual and computer-assisted techniques using NVivo 14 software. The interviews were analyzed starting with automatic transcription. Afterward, the interviews were reviewed again by listening to the audio and integrating missing parts manually. Finally, they were imported into NVivo for coding and analysis. The codes and subcodes were identified after careful discussion among the authors of this study, based on both manifest meanings and recurring patterns in the interviews. The coding scheme adopted for the interview analysis is: (a) Social media impact on public communication; (b) Trust; (c) Disinformation and misinformation; (d) GenAI views (Opportunity; Risk; AI ethical considerations and standards; Adoption of GenAI in the organization); (e) Communicator' role (Competencies and skills; International and national co-operation; Impact on intangible assets). Subcodes are specified only for the part of the study reported in this manuscript. Authors independently coded the same set of transcripts, compared their coding, and resolved discrepancies through consensus.

Using NVivo software, it was also possible to find connections between items by comparing codes, which were further supplemented by manual analysis from the researchers to understand the qualitative nature of

Table 1. Participants information.

Interviewee	Gender	EU Country	Position	Institution	Years of experience in public sector communication
1	Female	Ireland	Communications Manager	Government	10+
2	Female	Germany	Head of Communication Unit	Government	20+
3	Male	Portugal	Head of Communication Unit	Government	20+
4	Male	The Netherlands	Director of Communication	Government	20+
5	Male	Romania	Senior Communication Expert	Government	20+
6	Male	France	Head of Communication	International Organization	20+
7	Female	Croatia	Head of Public Information and Relations	Government	20+
8	Female	Slovenia	Former Head of Public Information	Government	20+
9	Male	Sweden	Director of Communication	Government	20+
10	Male	Italy	Head of Communication	EU Institution	20+
11	Male	Latvia	Deputy-Head of Communication	Government	10+
12	Male	The Netherlands	Former Head of Public Information	Government	20+
13	Male	Malta	Director of Communication	Government	15+
14	Male	Belgium	Head of Communication	EU Institution	20+

the links. Moreover, NVivo allowed for the identification of word trees from the interviews, facilitating the identification of thematic associations and repeated contexts in which specific words appeared.

4. Results and Discussions

All interviews showed a strong awareness of the technological revolution and the impact of digital technologies within governments and EU institutions. Participants underscored how digital platforms and GenAI tools are not only technologies but also communicative and social environments (Zerfass et al., 2024) that must be deeply understood and used with a critical approach. The interviewees highlighted that the use of AI tools should be approached carefully, avoiding both techno-enthusiastic perspectives—as seen during the first phase of social media adoption (Lovari & Valentini, 2020)—or catastrophic ones that predict significant job losses, as reported in the early debate on the use of GenAI (Council of the EU, 2023).

All the interviewees agreed that GenAI represents a game-changer for their profession in terms of message delivery and communication campaign productions, requiring a new sense of responsibility and an ethical approach, thus introducing new challenges at the organizational level and in relations with citizens and media, as can be seen from the following word tree (Figure 1).

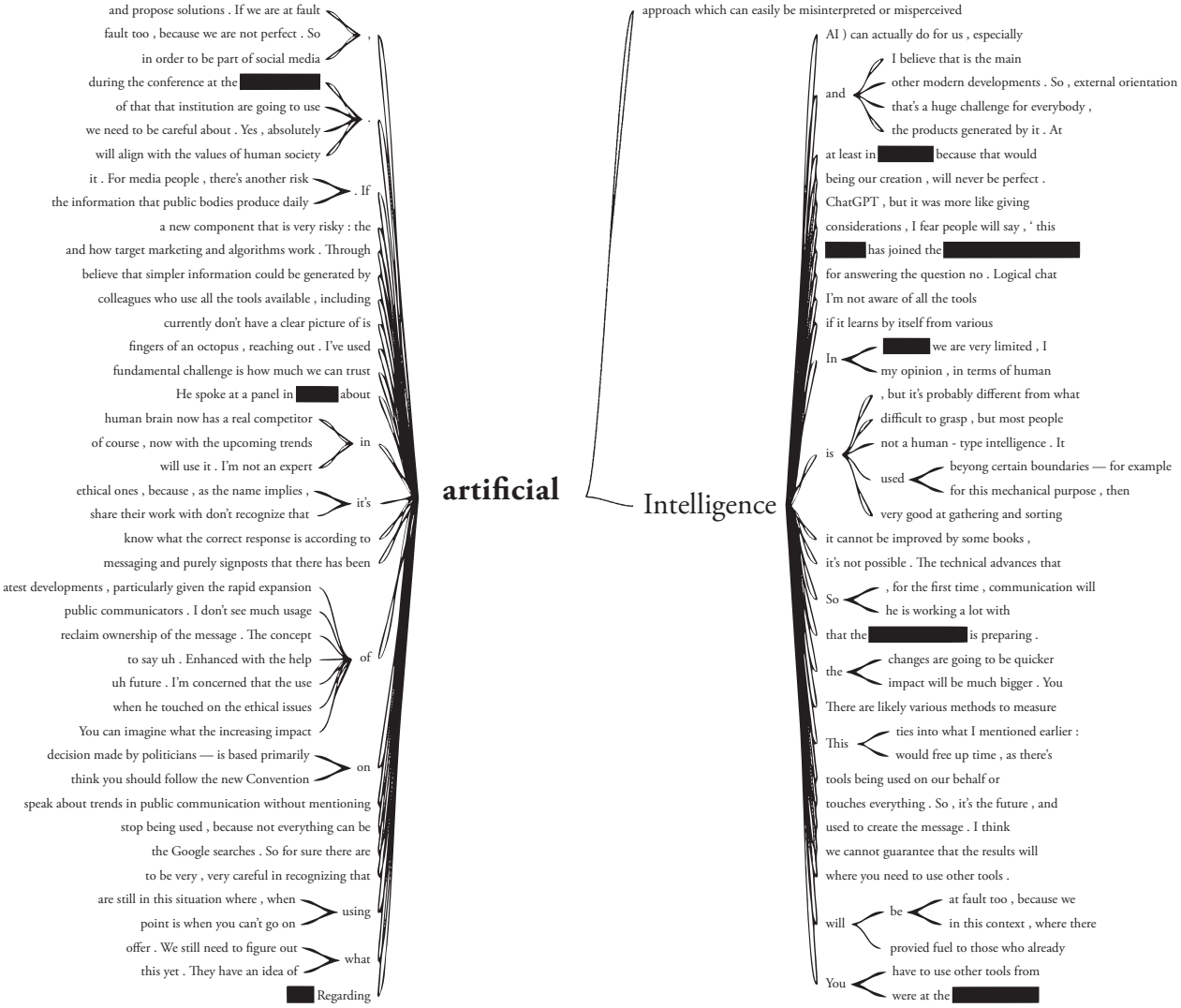


Figure 1. Word tree of the keyword "Artificial" extracted from qualitative interview analysis using NVivo 14.

This word tree shows the use of "AI" in relation to trust challenges, ethical aspects, and practical implications for public sector professionals in managing communication activities. A key theme that emerges is the role of trust in the use of AI. Phrases like "how much we can trust" (Interviewee 5) highlight a recurring concern about how much professionals can really rely on AI-generated decisions or AI-enabled information for their communication campaigns and media relations. For instance:

> I think the basic challenge of using AI is how much we can trust…the Artificial Intelligence and the products made by AI. In this sense, it is not, like, 100% sure; you know ChatGPT and similar platforms….So I believe we cannot completely rely on it or rather we should define to what extent AI can be, you know, actively involved in our work as public communicators. (Interviewee 5)

This excerpt directly ties into the discussion of the reliability of GenAI-based tools, especially when it comes to making decisions in both ordinary and crisis situations (Tzachor et al., 2020). Additionally, the link between human intelligence and GenAI is emphasized, as seen in the word tree phrase "human brain now has a real competitor" (Interviewee 10). This sentence highlights the perception that AI is becoming

increasingly sophisticated, with it being seen not only as a communicative partner (Esposito, 2021) but also as a genuine competitor to human intellect.

Analyzing the AI perception code (GenAI view) and its related subcodes (AI opportunity, AI risk, AI ethical considerations), interviewees expressed a limited skepticism about these tools' impact on public sector communication, maintaining a cautious, sometimes critical, approach. Out of 14 interviews, "AI opportunity" was cited more frequently than "AI risk" in the majority of interviews.

On the one hand, there is a consensus on the opportunities presented by the impact of GenAI on public communication. For instance, these included: enhancing efficiency in communication management; faster data analysis; creating communication plans; engaging different publics; monitoring social media sentiment; stimulating creativity in the production of visuals, videos, and messages; preparing internal staff training; and assisting in strategic and complex tasks that these professionals undertake regularly in their jobs (Ertem-Eray & Cheng, 2025). These potential applications and examples showcase what these elite professionals are already doing or plan to do with this emerging technology, suggesting trends that might become standard practices in European institutions, as reported in the following sentences:

> I have read and learned a lot about AI. I think that we are going to use more and more of these kinds of tools that artificial intelligence is going to provide us….You can generate texts, you can generate images and you can put that information on your websites, on your social media. (Interviewee 3)

> Bureaucracy creates a lot of information, data, and issues. Public communicators, most of the time, are spending their energy trying to understand, sort out, select, and process the information that the public bodies they work for produce on a daily basis. If artificial intelligence is used for this, let's say for a mechanical purpose, then I am pretty sure that the smart algorithms will help public communicators to quickly understand the key information. (Interviewee 5).

> With AI, I think the aim of public information would be to reach out to record numbers of people. You might reach a lot of people, but then the message would not get across. That is why I think it is very interesting to use AI in this field because it can target people. In the future, public information will be on this data, so AI will focus on gathering people's data. I imagine it is like the fingers of an octopus reaching out. Far, far, far, far across everywhere. (Interviewee 13).

The last excerpt highlights the process of "ultra-targetization" of citizens enabled by AI tools, representing a significant potential advancement of public sector communication in crafting tailored messages to publics. Indeed, the "octopus," with its many tentacles reaching out in various directions, symbolizes the extensive reach and precision targeting capabilities that AI brings to communication. Each tentacle represents a channel or a demographic segment, allowing for tailored communication strategies that effectively engage diverse publics.

Thematic analysis of the interviews also reported several risks and potential threats related to the use of GenIA for society and public sector communication. The majority of the interviewees emphasized the risks stemming from GenAI, particularly in relation to the issue of disinformation, as illustrated in Figure 2.

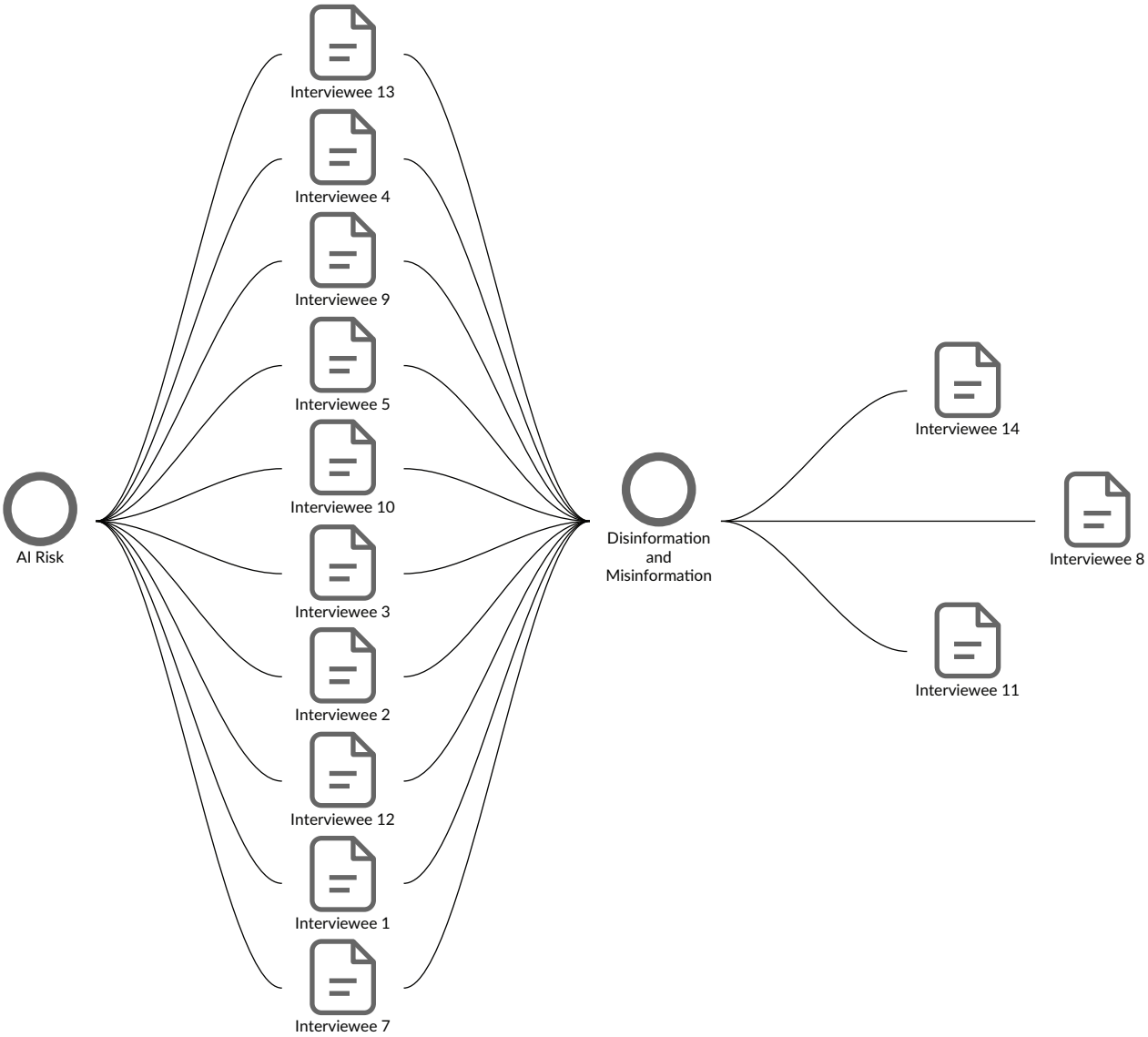


Figure 2. Comparison of codes between "AI Risk" and "Disinformation and Misinformation" from qualitative interview analysis using NVivo 14.

Specifically, the interviews highlighted risks associated with the creation of deepfakes, hallucinations, misleading content, and incorrect information that AI could generate, ultimately leading to the spread of disinformation in digital platforms (Pérez Dasilva et al., 2021; World Economic Forum, 2024). According to some interviewees, GenAI platforms could automate the generation of false content, accelerating its dissemination and multiplying the volume of misleading information online (Jungherr & Schroeder, 2023), as reported in these excerpts:

> Disinformation will be a huge challenge. AI is making it. It's so easy and so cheap and so possible for everyone with a computer to alter the truth in millions of possible ways and to spread this through millions of channels. And it costs nothing. This is not happening all the time, but it will eventually. There's a risk that AI will transform the core of the public discourse, so we will never know if what we are seeing is real or not. How will I be able to trust that you are human, and not a bot? (Interviewee 9)

> For example, today, you need troll farms to spread disinformation, but soon AI could do this without any human intervention, which could significantly increase the spread of disinformation. This will likely make it harder for us to deliver our own messages, as others will be able to use AI to distribute even more disinformation with greater impact. (Interviewee 2)

Additionally, another insight from the interviewees' responses concerns the transparency of the sources used by GenAI tools in content creation. Respondents underlined that digital platforms should disclose their sources to allow users to recognize whether the content is false or not. Moreover, human oversight should always be present to verify sources and information. From this perspective, many professionals highlighted the important role of public communicators in fulfilling this key function to guarantee transparency and accountability in the government's message production and sharing. This approach ensures that GenAI serves as an aid rather than a substitute for human expertise in public sector communication, due to the strategic role this function plays in the democratic debate (Formosa et al., 2024).

Furthermore, interviewees emphasized the importance of media and digital literacy and the need for governments and EU institutions to provide citizens with the knowledge and tools to detect and combat disinformation:

> You should be careful when using AI, especially with sensitive topics. You must tell people when you are using AI. If you use AI as a discovery tool, in five years, no one will be concerned about where the information is coming from. You'll go to ChatGPT, type in a question, and get an answer without really thinking about how that information was generated or what the sources are. That's something we need to pay attention to. AI should tell you where it's getting that information. (Interviewee 3)

> That's a huge challenge for everybody, especially for government institutions. As you know, government institutions are always very slow and not so quick to adapt to all these changes. Everything is moving too fast, and I'm afraid that we might have situations where we would have problems to deal with. It's very dangerous to receive information and not know if it's disinformation, if it's true, or if it's not true, and you don't know the source of where the information is coming from. (Interviewee 7)

Another subcode was related to the ethical impact and implications of GenAI for public sector communication. The six interviews revealed a direct relation (Figure 3) and a wide set of different views, ethical considerations, and professional standards stemming from the use and integration of GenAI technologies into public information practices toward citizens and the media.

These professionals highlighted that the capability of these tools in processing vast amounts of personal data for targeted messaging raises privacy and ethical dilemmas (Bowen, 2024), not only limited to the field of public communication:

> If I used AI tools for public communication purposes, my main ethical dilemma would be data protection….You already know how target marketing and algorithms work, and when they are powered by AI, it will be much, much bigger than that. (Interviewee 13)

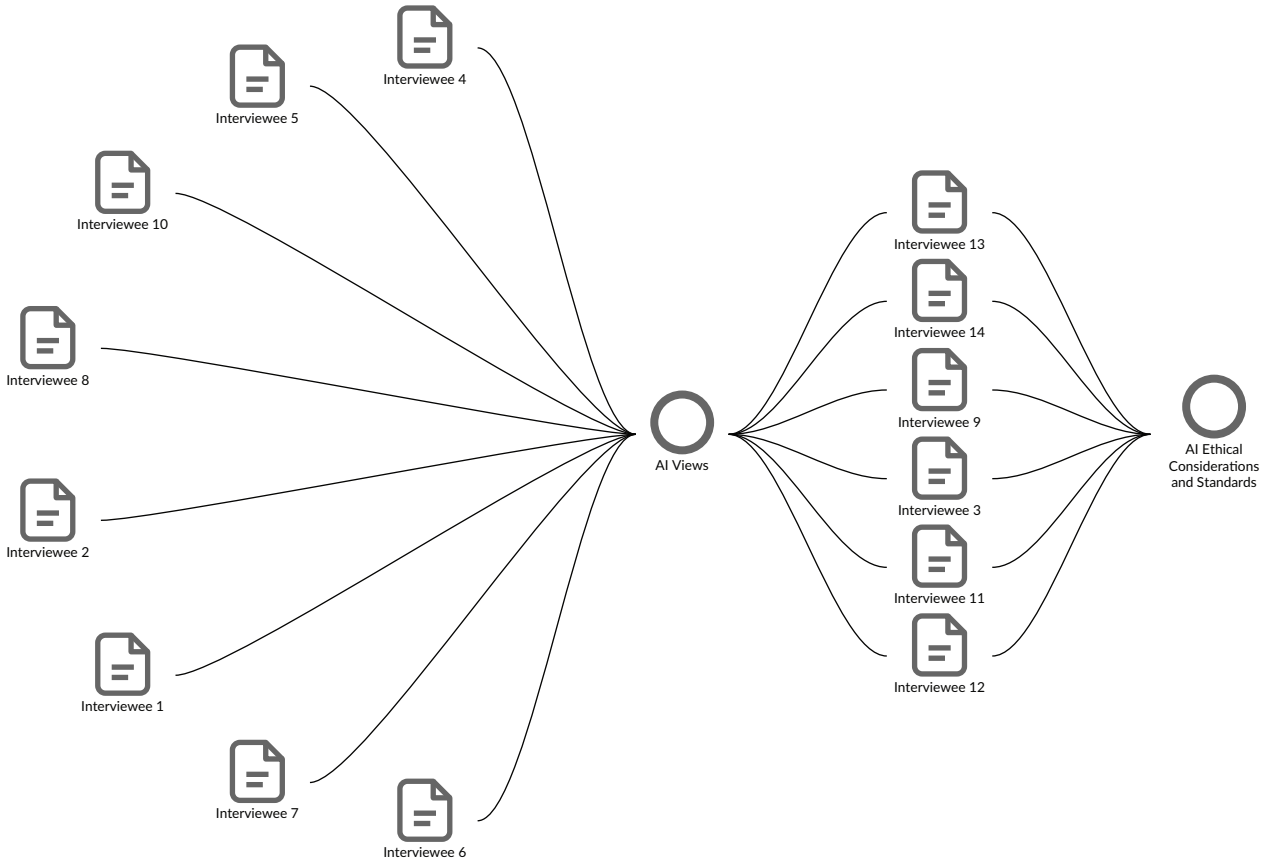


Figure 3. Comparison of codes between "AI Views" and "AI Ethical Considerations and Standards" from qualitative interview analysis using NVivo 14.

> You have to consider ethics. The high risks that come with AI are not just about technology; there are ethical considerations, and institutions, including governments and the EU, have been discussing these challenges for years. (Interviewee 14)

Some interviewees, moreover, do not currently perceive greater ethical risk for communication management, particularly due to the role played by the European Commission in AI regulation (Ferrari et al., 2023), and a greater sense of responsibility that these professionals experience while performing their strategic roles. For instance:

> The major issue is whether, when using AI, especially in relation to visuals, and if we are moving towards AI-generated interpretations in videos, etc…we are actually presenting these things as facts or if we are using this tool to make information more accessible to people. That would be the ethical issue. (Interviewee 11)

> I don't see a major ethical conflict with this because we are using it mainly for inspiration. The end product is always controlled and further developed by a desk officer. For instance, if we use an AI-generated image, we review it carefully; we never just take anything AI produces and send it out without oversight. We still take full responsibility for everything we do, and we don't have any unfiltered AI-generated content going out. (Interviewee 9)

According to most of the interviewees, ensuring the accuracy, transparency, and authenticity of AI-generated communication is crucial for maintaining public trust (Bowen, 2024). This requires verifying the information developed by AI tools in communication management (UN Interregional Crime and Justice Research Institute, 2020) and safeguarding that messages and campaigns align with the essential key principles of public communication (Bowen & Lovari, 2021; OECD, 2021) before publishing or sharing them.

Regarding RQ2, respondents collectively underscore the need to possess a balanced skill set, blending traditional communication expertise with digital literacy, and an adaptive approach to emerging technologies such as GenAI. The interviewees stated that to address contemporary socio-technological challenges and the rampant impact of GenAI platforms, public sector communicators should have a set of fundamental competencies and skills to navigate the modern communication landscape (Figure 4).

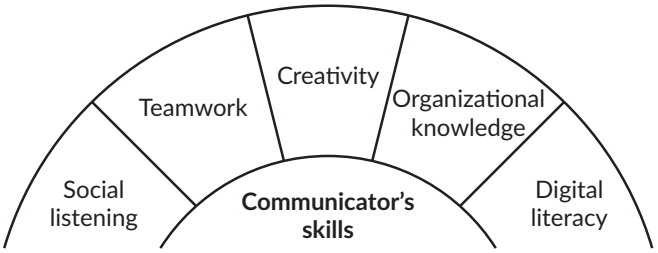


Figure 4. Diagram of the communicator's skills based on recurring words from the interviews.

First, social listening is a crucial skill. Listening is a central theme in public communication research (Macnamara, 2016), but it gains new meaning with the development of digital platforms and AI-generated communication. By understanding citizens' needs and concerns while leveraging digital tools, public communication can become more user-centered and effective (Smillie & Scharfbillig, 2024). Indeed:

> What we really need are people who understand and can bridge the gap between technical aspects on the one hand and citizens on the other, understanding their needs and the kind of information they seek. (Interviewee 4)

> We need people able to cross the gap to understand what citizens need to hear, to focus on the citizens, not on the institution. (Interviewee 5)

Teamwork is equally important, as communication is often the result of collaborative efforts with other professionals inside the organization but also in an inter-institutional perspective, as shown during the Covid-19 pandemic, thus having an impact on policy-making (Lovari, 2020; OECD, 2021). This skill is particularly highlighted by those professionals who have already started collaborating with other practitioners to integrate digital technologies into their activities. For instance:

> In my country, we are trying to promote co-operation between communicators and policymakers. First of all, communication is seen as a strategic role, not as a supporting role. On the one hand, you can already build communication in your policies. On the other hand, being a part of the development of a policy makes it easier for you to understand the audience, the solutions presented, the opposition, etc. So that's a very important future role to be more present in policy process, not just receive a package, a product that you have to sell. (Interviewee 11)

> I wouldn't forget to mention the importance of teamwork. Nowadays, there's no place anymore for solo work or single players. That's not possible; work is done in a team, and the solid skill of working in a team, which is not always easy, is critical at the international level. (Interviewee 6)

Then, creativity will be strategic since it enables the innovation needed to find original solutions to inform and engage the public, and it relies not only on choosing the effective prompts for GenAI tools or integrating other online solutions. It is also connected to the capacity of professionals to assemble and integrate traditional and innovative communication practices, including data management and multimedia content production (Zerfass et al., 2024):

> We need people to be innovative and to think about new ways of communicating with people. We also need people who are developing new methods of communication, so you know, people who are literally writing the code and coming up with whatever algorithms and apps and things are. So we find that there's an expansion of the need for communications. (Interviewee 1)

This approach is essential for crafting effective campaigns and messages for the evolving informative needs of citizens, collaborators, and mass media, thus synchronizing public sector communication with society.

Moreover, digital skills (mostly related to social media and GenAI) are becoming increasingly relevant for public communicators who need to stay updated not only on the technological evolution of digital platforms but also on their communicative implications and uses, as highlighted in this interview's excerpt.

> Given that AI is an essential tool for the future, I believe it should be a fundamental part of every communicator's training, covering different aspects of AI and how to use it effectively. (Interviewee 9)

Digital literacy is important not only for strategically managing platforms and communication automation (Zerfass et al., 2023) but also for detecting potential online crises and disinformation practices that could harm PSOs or the general population (Coombs, 2020; Lovari & Valentini, 2020; OECD, 2021).

Finally, a thorough understanding of PSOs' functioning and organizational culture (Canel & Luoma-aho, 2018), is considered by the interviewees both fundamental and strategic for enhancing the quality of public communication in these turbulent and challenging times. Indeed:

> It's important to engage with today's platforms and social media, but the most crucial aspect is a solid academic background—knowledge of international relations, political sciences, and the global context. This includes understanding crises, which are international in nature, and acquiring a strong foundation in public diplomacy. (Interviewee 14)

> I think someone who is able to walk the line and understand what is happening within institutions and governments, while also having a strong sense of how to connect with and reach the other side: the people. (Interviewee 2)

> If that person will not understand how the public institutions are operating in regard to the citizen's needs, then they will not be able to catch.. to present the empathy necessary to help pass the message to the audience. (Interviewee 5)

All these skills and competencies require a constant process of training by public sector communicators that, according to most of the respondents, necessitate a proactive approach by professionals, as well as playing an active role in different networks at the international and national levels.

Lastly, the majority of respondents highlighted the important role played by public sector communication in nurturing intangible assets to improve the relationships with citizens and the media (Canel & Luoma-aho, 2018). In particular, communicators will play an important role in enhancing and maintaining trust in government and institutions in a current scenario characterized by polarization and disinformation fueled by digital platforms and GenAI tools (OECD, 2021; World Economic Forum, 2024).

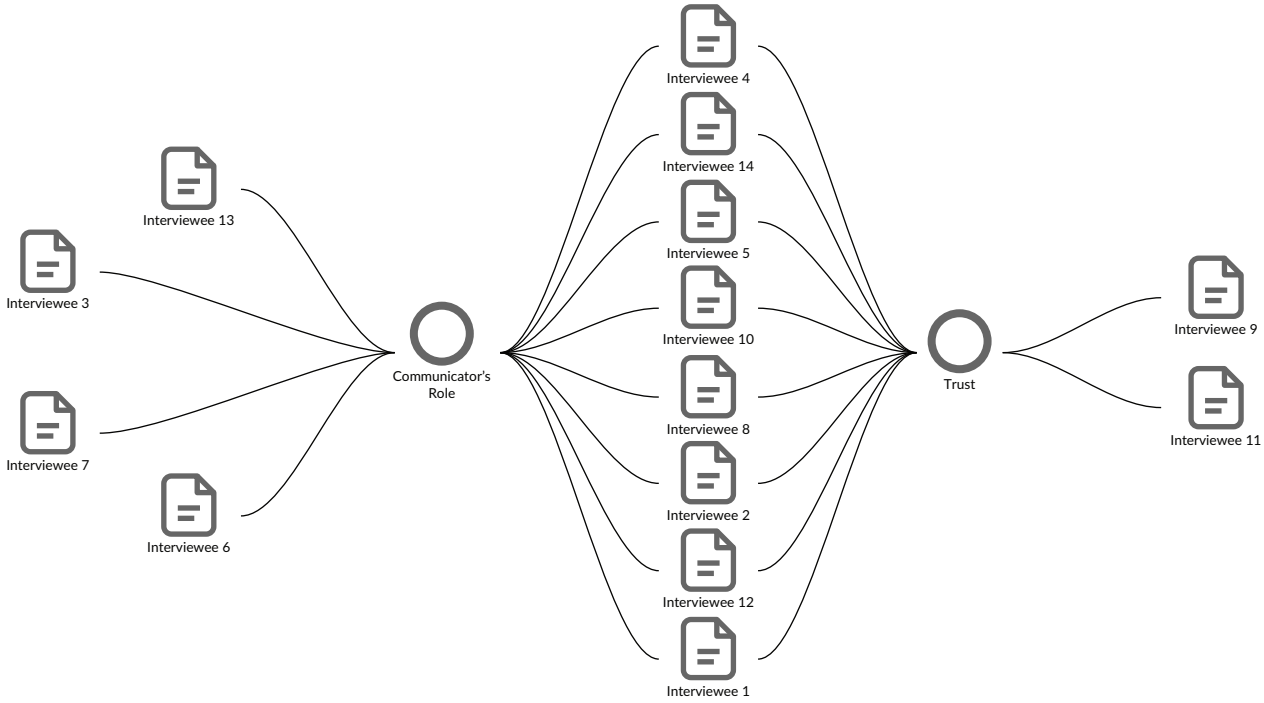


Figure 5. Comparison of codes between "Communicator's Role" and "Trust" from qualitative interview analysis using NVivo 14.

Communicators must be able to build and maintain a relationship of trust and credibility with the different publics inside and outside the organizations (citizens, media, employees, etc.), improving the quality and authenticity of their communication flows:

> We should be the trusted voice of the government. The means…will change. We have far more means than we have in the past, but these means should not distract us from where we are (positioned) in the world of communication….We really should be the trusted voice of the government. So that's not evolving. That will be the same. We should always stay people of flesh and blood. (Interviewee 4)

5. Conclusions

This article aimed to investigate the impact of GenAI tools in European public sector communication through in-depth interviews with elite communicators working for EU Institutions and European Governments.

Despite some limitations, such as the limited number of interviewees and the fact that interviewing elites provides valuable insights and trends but sometimes fails to reveal organizational issues or diverse organizational cultures, this article represents one of the first studies focusing on the impact of GenAI tools on communication management in governments and public organizations at the European level. It focuses on the perspective of communication professionals, allowing us to deeply understand how these professionals are increasingly compelled to strategically realign their intricate communication practices in response to the rapid and transformative evolution of technologies, particularly in their dynamic interactions with the turbulent external environment.

This study has interesting implications for strategic communication. First, it responds to the need to extend the research on AI to explore the implication of GenAI across different communication sectors and various industries (Ertem-Eray & Cheng, 2025), investigating the challenges these tools provide to governments and public sector communication, shedding light on their institutional, organizational, and cultural specificities. Also, the empirical evidence from this research can add a valuable contribution to public communication scholarship, particularly regarding the diffusion and adoption of GenAi technologies in the transformation of organizational cultures within the public sector (Canel & Luoma-Aho, 2018) under the pressure of internal and external factors, such as information crisis (Kim & Gil de Zúñiga, 2020), citizen's expectations, media coverage, and international regulations (Ferrari et al., 2023). Lastly, this study contributes to increasing the knowledge of the digital media-arena framework (Badham et al., 2024) and how these online communicative spaces, in particular the artificial digital media-arena, can be strategically managed by public sector communicators in order to relate and communicate with citizens, media, and other strategic stakeholders.

The manuscript also presents practical implications for professionals working in government and PSOs. The Shakespearean question "AI or not AI" is no longer relevant since AI and its integration into institutional communication have become inevitable and are also fueled by intensive international media coverage. Today, the crucial question is how, when, and where GenAI solutions are affecting PSOs. This trend will impact communications and relationships with citizens and the media, particularly in light of the approval of the EU AI Act (Council of the EU, 2024) and the evolving nature of these technologies. Interviews conducted in this study explored how these leading professionals perceive and experience the changing landscape of public sector communication influenced by the disruptive impact of these digital platforms. In addition, the findings highlighted the essential competencies and skills for rethinking the role of public communicators. Thus, GenAI's ability to aggregate and analyze data can redefine the activities of public communicators, shifting informative content production towards more data-driven approaches (OECD, 2021). However, this shift also requires that these professionals develop a new set of skills to interpret AI-generated insights accurately and apply them responsively and effectively in their communication and media relations strategies.

Nowadays, public communicators are evolving into centaur communicators, hybrid figures competent at navigating between analog and digital environments. They combine the practices and logics of legacy media with those characterizing social media (van Dijck et al., 2018) and digital media arenas where chatbots and AI solutions will be predominant also for public sector communication (Badham et al., 2024). This hybrid role underscores the necessity for communicators to adapt and integrate diverse communication tools and platforms, maintaining the essence of traditional media while embracing the new possibilities offered by digital advancements (Ducci & Lovari, 2021). Such evolution requires communicators to undergo a paradigm

shift towards a new level of professionalism, a more fluid and dynamic role, and a new sense of ethics and responsibility toward their organizations and society at large (Bowen, 2024; Smillie & Scharfbillig, 2024). Indeed, some of the established practices and skills that have been solidified over the years now appear obsolete due to societal risks, the rapid pace of digital innovation, and the activism of connected citizens and algorithmic media routines. It seems increasingly evident that recent machine learning and big data-based algorithms are able to participate in communication. Today, algorithms can act as communicative partners (Esposito, 2021). What matters is "how" communicators will partner and strategically engage with them to enhance public sector communications in ordinary and emergency situations. In fact, leveraging AI-generative tools for public communication requires committed human oversight, responsible application, and rigorous fact-checking while continuously evaluating potential risks through the lens of professional ethics and accountability (OECD, 2024). Finally, public sector communicators must act as a "steady rock" in a rapidly evolving digital information landscape, polluted by fake content, deepfakes (Pérez Dasilva et al., 2021), and disinformation. In an era increasingly reshaped by GenAI, their role is critical in maintaining integrity and trust within democratic societies, while enabling meaningful conversations in the public sphere for the benefit of all.

Acknowledgments

We would like to express our sincere gratitude to the anonymous reviewers for their thoughtful feedback and constructive suggestions, which have contributed to strengthening this work. We also extend our appreciation to all the participants who generously shared their time and insights during the interviews, providing valuable perspectives that enriched this study. AI was employed in a small portion of the text (less than 10%) for language clarity and checking using ChatGPT. AI-assisted texts were double-checked by the authors to correct selected wordings. The authors, safeguarding human oversight and expertise in the final output, performed the overall composition of the article manually.

Conflict of Interests

The authors declare no conflict of interest.

References

Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly*, *36*(2), 358–367.

Audétat, M. (2022). Promising technosciences in the economy of attention: Why have pessimistic stories of disruption and "artificial intelligence" performed so well? *TECNOSCIENZA: Italian Journal of Science & Technology Studies*, *13*(2), 35–56.

Badham, M., Luoma-aho, L., & Valentini, C. (2024). A revised digital media–arena framework guiding strategic communication in digital environments. *Journal of Communication Management*, *28*(2), 226–246.

Benesch, S. (2023). Dangerous speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 185–197). Digital Communication Research. https://doi.org/10.48541/dcr.v12.11

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research*, *10*(2), 141–163.

Bowen, S. A. (2024). "If it can be done, it will be done": AI ethical standards and a dual role of public relations. *Public Relations Review*, *50*(1), 1–13.

Bowen, S. A., & Lovari, A. (2021). Ethics in government public relations and modern challenges for public sector organizations. In M. Lee, G. Neeley, & K. Stewart (Eds.), *The practice of government public relations* (pp. 175–195). Routledge.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bright, J., Enock, F., Esnaashari, S., Francis, J., Hashem, Y., & Morgan, D. (2024). Generative AI is already widespread in the public sector: Evidence from a survey of UK public sector professionals. *Digital Government: Research and Practice*, *6*(1), Article 2. https://doi.org/10.1145/3700140

Bryman, A. (2012). *Social research methods*. Oxford University Press.

Canel, M., & Luoma-aho, V. (2018). *Public sector communication: Closing gaps between citizens and public organizations*. Wiley.

Chadwick, A. (2013). *The hybrid media system: Politics and power*. Oxford University Press.

Chee, F. Y. (2024, February 2). Europe within reach of landmark AI rules after nod from EU countries. *Reuters*. https://www.reuters.com/technology/france-now-backing-eu-ai-rules-eu-source-says-ahead-bloc-endorsement-2024-02-02

Club of Venice. (2013). *Factsheet*. https://www.affarieuropei.gov.it/media/2648/club-of-venice.pdf

Coombs, W. T. (2020). Public sector crises: Realizations from Covid-19 for crisis communication. *Partecipazione e Conflitto*, *13*(2), 990–1001.

Council of the European Union. (2023). *ChatGPT in the public sector: Overhyped or overlooked?* (Research paper). https://www.consilium.europa.eu/media/63818/art-paper-chatgpt-in-the-public-sector-overhyped-or-overlooked-24-april-2023_ext.pdf

Council of the European Union. (2024, May 21). *Artificial Intelligence Act: Council gives final green light to the first worldwide rules on AI* [Press release]. https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai

Desouza, K. C., Dawson, G. S., & Chenok, D. (2020). Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. *Business Horizons*, *63*(2), 205–213.

Ducci, G., & Lovari, A. (2021). The challenges of public sector communication in the face of the pandemic crisis: Professional roles, competencies, and platformization. *Sociologia Della Comunicazione*, *2021*(61), 9–19. https://doi.org/10.3280/SC2021-061002

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashwari, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice, and policy. *International Journal of Information Management*, *71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Edelman. (2023). *2023 Edelman trust barometer*. https://www.edelman.com/trust/2023/trust-barometer

Ertem-Eray, T., & Cheng, Y. (2025). A review of artificial intelligence research in peer-reviewed communication journals. *Applied Sciences*, *15*(3), Article 1058. https://doi.org/10.3390/app15031058

Esposito, E. (2021). *Artificial communication: How algorithms produce social intelligence*. MIT Press.

European Commission. (2023a, December 8). *Statement by commissioner Breton—The European AI Act is here!* [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_23_6471

European Commission. (2023b, December 9). *Commission welcomes political agreement on Artificial Intelligence Act* [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473

European Commission. (2025). *A European approach to artificial intelligence*. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

European Parliament. (2023a). *EU AI Act: First regulation on artificial intelligence*. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

European Parliament. (2023b). *Media & news survey 2023*. https://europa.eu/eurobarometer/surveys/detail/3153

Ferrari, F., van Dijck, J., & van den Bosch, A. (2023). Observe, inspect, modify: Three conditions for generative AI governance. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448231214811

Formosa, P., Kashyap, B., & Sahebi, S. (2024). Generative AI and the future of democratic citizenship. *Digital Government: Research and Practice*. Advance online publication. https://doi.org/10.1145/3674844

García-Peñalvo, F. J., & Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(4), 7–16. https://doi.org/10.9781/ijimai.2023.07.006

Gil de Zúñiga, H., Goyanesd, M., & Durotoyeb, T. (2024). A scholarly definition of artificial intelligence (AI): Advancing AI as a conceptual framework in communication research. *Political Communication*, *41*(2), 317–334.

Gusfield, J. R. (1981). *The culture of public problems: Drinking-driving and the symbolic order*. University of Chicago Press.

Haefner, N., Wincent, J., Parida, V., & Gassmann, O. (2020). Artificial intelligence and innovation management: A review, framework, and research agenda. *Technological Forecasting Social Change*, *162*, Article 120392.

Haro-de-Rosario, A., Sáez-Martín, A., & Caba-Perez, M. C. (2018). Using social media to enhance citizens engagement with local government: Twitter or Facebook? *New Media & Society*, *20*(1), 29–49. https://doi.org/10.1177/1461444816645652

Helmond, A. (2015). The platformization of the web: Making web data platform ready. *Social Media + Society*, *1*(2). https://doi.org/10.1177/2056305115603080

Hertz, R., & Imber, J. B. (Eds.). (1995). *Studying elites using qualitative methods*. Sage.

Hjaltalin, I. T., & Sigurdarson, H. T. (2024). The strategic use of AI in the public sector: A public values analysis of national AI strategies. *Government Information Quarterly*, *41*, Article 101914. https://doi.org/10.1016/j.giq.2024.101914

Johnson, J. M. (2001). In-depth interviewing. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context and method* (pp. 103–119). Sage.

Jungherr, A., & Schroeder, R. (2023). Artificial intelligence and the public arena. *Communication Theory*, *33*(2/3), 164–173. https://doi.org/10.1093/ct/qtad006

Kennedy, B., Tyson, A., & Saks, E. (2023). *Public awareness of artificial intelligence in everyday activities*. Pew Research Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2023/02/PS_2023.02.15_AI-awareness_REPORT.pdf

Kim, J. N., & Gil de Zúñiga, H. (2020). Pseudo-information, media, publics, and the failing marketplace of ideas: Theory. *American Behavioral Scientist*, *65*(2), 163–179. https://doi.org/10.1177/0002764220950606

Konrad, K., van Lente, H., Groves, C., & Selin, C. (2016). Performing and governing the future in science and technology. In U. Felt, R. Fouche, C. A. Miller, & L. Smith-Doerr (Eds.), *The handbook of science and technology studies* (pp. 465–493). MIT Press.

Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision making in democratic settings. *Telecommunications Policy*, *44*(6), Article 101976. https://doi.org/10.1016/j.telpol.2020.101976

Larsen, A. G., & Følstad, A. (2024). The impact of chatbots on public service provision: A qualitative interview study with citizens and public service providers. *Government Information Quarterly*, *41*(2), Article 101927. https://doi.org/10.1016/j.giq.2024.101927

Lorenz, P., Perset, K., & Berryhill, J. (2023). *Initial policy considerations for generative artificial intelligence* (OECD Artificial Intelligence Papers, No. 1). OECD. https://doi.org/10.1787/fae2d1e6-en

Lovari, A. (2020). Spreading (dis)trust: Covid-19 misinformation and government intervention in Italy. *Media and Communication*, *8*(2), 458–461. https://doi.org/10.17645/mac.v8i2.3219

Lovari, A., & Belluati, M. (2023). We are all Europeans: EU institutions facing the Covid-19 pandemic and information crisis. In G. La Rocca, M.-E. Carignan, & G. B. Artieri (Eds.), *Infodemic disorder: Covid-19 coping strategies in Europe, Canada and Mexico* (pp. 65–96). Palgrave MacMillan.

Lovari, A., D'Ambrosi, L., & Bowen, S. A. (2020). Re-connecting voices: The (new) strategic role of public sector communication after Covid-19 crisis. *Partecipazione e Conflitto*, *13*, 970–989. https://doi.org/10.1285/i20356609v13i2p970

Lovari, A., & Valentini, C. (2020). Public sector communication and social media: Opportunities and limits of current policies, activities, and practices. In V. Luoma-aho & M. J. Canel (Eds.), *Handbook of Public Sector Communication* (pp. 315–328). Wiley.

Macnamara, J. (2016). *Organizational listening: The missing essential in public communication*. Peter Lang.

Madan, R., & Ashok, M. (2023). AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly*, *40*(1), Article 101774. https://doi.org/10.1016/j.giq.2022.101774

McCorkindale, T. (2024). *Generative AI in organizations: Insights and strategies from communication leaders*. Institute for Public Relations. https://instituteforpr.org/wp-content/uploads/Generative-AI-in-Organizations-Insights-and-Strategies-from-Communic

OECD. (2021). *OECD report on public communication: The global context and the way forward*. https://doi.org/10.1787/22f8031c-en

OECD. (2024). *Assessing potential future artificial intelligence risks, benefits and policy imperative*. https://www.oecd-ilibrary.org/deliver/3f4e3dfb-en.pdf?itemId=%2Fcontent%2Fpaper%2F3f4e3dfb-en&mimeType=pdf

Olsson, E. K., & Eriksson, M. (2016). The logic of public organizations' social media use: Toward a theory of 'social mediatization.' *Public Relations Inquiry*, *5*(2), 187–204. https://doi.org/10.1177/2046147X16654454

Panda, G., Upadhyay, A. K., & Khandelwal, K. (2019). Artificial intelligence: A strategic disruption in public relations. *Journal of Creative Communications*, *14*(3), 196–213. https://doi.org/10.1177/0973258619866585

Pérez Dasilva, J., Meso Ayerdi, K., & Mendiguren Galdospin, T. (2021). Deepfakes on Twitter: Which actors control their spread? *Media and Communication*, *9*(1), 301–312. https://doi.org/10.17645/mac.v9i1.3433

Pislaru, M., Vlad, C. S., Ivascu, L., & Mircea, I. I. (2024). Citizen-centric governance: Enhancing citizen engagement through artificial intelligence tools. *Sustainability*, *16*(7), Article 2686. https://doi.org/10.3390/su16072686

Reutter, L. (2022). Constraining context: Situating datafication in public administration. *New Media & Society*, *24*(4), 903–921.

Salah, M., Abdelfattah, F., & Al Halbusi, H. (2023). Generative artificial intelligence (ChatGPT & bard) in public administration research: A double-edged sword for street-level bureaucracy studies. *International Journal of Public Administration*. Advance online publication. https://doi.org/10.1080/01900692.2023.2274801

Selten, F., & Klievink, B. (2024). Organizing public sector AI adoption: Navigating between separation and integration. *Government Information Quarterly*, *41*(1), Article 101885.

Silva, P., Tavares, A. F., Silva, T., & Lameiras, M. (2019). The good, the bad and the ugly: Three faces of social media usage by local governments. *Government Information Quarterly*, *36*(3), 469–479. https://doi.org/10.1016/j.giq.2019.05.006

Smillie, L., & Scharfbillig, M. (2024). *Trustworthy public communications*. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/695605

Smith, A. B., & Waddington, S. (2023). *Artificial intelligence (AI) tools and the impact on public relations (PR) practice*. Chartered Institute of Public Relations.

Toll, D., Lindgren, I., Melin, U., & Madsen, C. Ø. (2020). Values, benefits, considerations and risks of AI in government: A study of AI policy documents in Sweden. *eJournal eDemocracy Open Government*, *12*(1), 40–60.

Tzachor, A., Whittlestone, J., Sundaram, L., & O' hE'igeartaigh, S. (2020). Artificial intelligence in a crisis needs ethics with urgency. *Nature Machine Intelligence*, *2*, 365–366. https://doi.org/10.1038/s42256-020-0195-0

UN Interregional Crime and Justice Research Institute. (2020). *Stop the virus of disinformation: The risk of malicious use of social media during Covid-19 and the technology options to fight it*. https://digitallibrary.un.org/record/3927039?v=pdf

van Dijck, J. (2020). Governing digital societies: Private platforms, public values. *Computer Law & Security Review*, *36*, Article 105377. https://doi.org/10.1016/j.clsr.2019.105377

van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication*, *1*(1), 2–14. https://doi.org/10.17645/mac.v1i1.70

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. *Government Information Quarterly*, *39*(3), Article 101714.

Wang, J., Kiran, E., Aurora, S. R., Simeone, M., & Lobo, J. (2024). ChatGPT on ChatGPT: An exploratory analysis of its performance in the public sector workplace. *Digital Government: Research and Practice*. Advance online publication. https://doi.org/10.1145/3676281

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe.

World Economic Forum. (2024). *Global risks report 2024*. https://www.weforum.org/publications/global-risks-report-2024

Zerfass, A., Buhmann, A., Laborde, A., Moreno, A., Romenti, S., & Tench, R. (2024). *European communication monitor 2024: Managing tensions in corporate communications in the context of geopolitical crises, artificial intelligence, and managerial learning*. European Public Relations Education and Research Association. https://www.communicationmonitor.eu/2024/11/23/ecm-european-communication-monitor-2024

Zerfass, A., Hagelstein, J., & Tench, R. (2020). Artificial intelligence in communication management: A cross-national study on adoption and knowledge, impact, challenges and risks. *Journal of Communication Management*, *24*(4), 377–389. https://doi.org/10.1108/jcom-10-2019-0137

Zerfass, A., Tench, R., Verčič, D., Moreno, A., Buhmann, A., & Hagelstein, J. (2023). *European Communication Monitor 2023: Looking back and ahead—15 years of research on strategic communication*. European Public Relations Education and Research Association; European Association of Communication Directors. https://www.communicationmonitor.eu/2023/09/07/ecm-european-communication-monitor-2023

Zerfass, A., Verčič, D., Verhoeven, P., Moreno, A., & Tench, R. (2019). *European Communication Monitor 2019: Exploring trust in the profession, transparency, artificial intelligence and new content strategies*. European Public Relations Education and Research Association; European Association of Communication Directors. https://www.communicationmonitor.eu/2019/05/23/ecm-european-communication-monitor-2019

Zuiderwijk, A., Yu-Chen, C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda. *Government Information Quarterly*, *38*, Article 101577. https://doi.org/10.1016/j.giq.2021.101577

About the Authors

Alessandro Lovari (PhD) is an associate professor of sociology of communication at the Department of Political and Social Sciences, University of Cagliari (Italy), where he is the coordinator of the Doctoral Program in Research and Social Innovation. He is the vice-chair of the Organizational Strategic Communication Section in ECREA. Lovari's research, focused on public sector communication, public relations, and health communication, investigates digital technologies' impact on organizational practices and citizens' behaviors. He is author of more than 100 publications in books, encyclopedias, and journals.

Fabrizio De Rosa is a public communication expert and independent researcher with experience in managing multichannel campaigns for national and international institutions. His work focuses on the intersection of public communication, governance, technological innovation, and human rights. Currently, he collaborates with national institutions to develop strategic awareness initiatives that enhance citizen engagement with digital public services. His research explores cyber threats, fact-checking tools, disinformation campaigns, and AI's impact on public sector communication in Europe.

Media and Communication
2025 • Volume 13 • Article 9523
https://doi.org/10.17645/mac.9523

ARTICLE

Open Access Journal

How Generative AI Went From Innovation to Risk: Discussions in the Korean Public Sphere

Sunghwan Kim and Jaemin Jung

Moon Soul Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, Republic of Korea

Correspondence: Jaemin Jung (nettong@kaist.ac.kr)

Submitted: 29 October 2024 **Accepted:** 6 January 2024 **Published:** 17 April 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

Technological progress breeds both innovation and potential risks, a duality exemplified by the recent debate over generative artificial intelligence (GAI). This study examines how GAI has become a perceived risk in the Korean public sphere. To explore this, we analyzed news articles (N = 56,468) and public comments (N = 68,393) from early 2023 to mid-2024, a period marked by heightened interest in GAI. Our analysis focused on articles mentioning "generative artificial intelligence." Using the social amplification of risk framework (Kasperson et al., 1988), we investigated how risks associated with GAI are amplified or attenuated. To identify key topics, we employed the bidirectional encoder representations from transformers model on news content and public comments, revealing distinct media and public agendas. The findings show a clear divergence in risk perception between news media and public discourse. While the media's amplification of risk was evident, its influence remained largely confined to specific amplification stations. Moreover, the focus of public discussion is expected to shift from AI ethics and regulatory issues to the broader consequences of industrial change.

Keywords

AI; amplification stations; ChatGPT; generative AI; public discourse; risk amplification; risk attenuation; risk communication

1. Introduction

Technological progress always has two sides, offering opportunities for innovation while posing significant risks. Beck (2012) predicted that 21st-century society would face threats not from traditional "dangers" but from "risks" shaped by human activity. In his concept of a global risk society, Beck (2012) highlighted risks

such as climate change, financial crises, and terrorism, emphasizing how these challenges are amplified in a knowledge-driven world.

Today, global risks are increasingly evident, exacerbated by advances in science and technology. One prominent example is the growing controversy surrounding generative artificial intelligence (GAI), which has recently captured public and market attention. GAI, a technology capable of generating content based on user input, is regarded as a game-changer in various industries, with profound social and cultural implications. Microsoft co-founder Bill Gates has described AI's development as the most significant technological advancement in decades, likening its impact to the advent of the iPhone (Gates, 2023; Nolan, 2023).

However, alongside its transformative potential, GAI has also sparked concerns and uncertainties. For instance, when OpenAI unveiled ChatGPT, media outlets speculated that the dominance of Google's search engine could be under threat. Reports from Google executives referred to this as a "code red," signalling a potential crisis (Cuthbertson, 2022; Grant & Metz, 2022). More broadly, the rapid proliferation of GAI has led to calls for caution, exemplified by the open letter titled *Pause Giant AI Experiments* and signed by AI researchers and tech leaders (Bengio et al., 2023). Such developments underscore the relevance of Beck's "global risk society" in contemporary contexts.

A review of existing literature has categorized the risks and controversies surrounding GAI into several key areas: (a) lack of market regulation and urgent regulatory needs; (b) poor content quality, disinformation, deepfakes, and algorithmic bias; (c) job losses due to automation; (d) breaches of privacy and data security; (e) social manipulation and erosion of ethics; (f) widening socioeconomic inequality; and (g) technology-related stress (Wach et al., 2023). Wach et al. (2023) argues that it is imperative to examine the social and ethical implications as GAI continues to develop.

This study investigates the risks associated with GAI, focusing on its amplification in the Korean public sphere. We analyze the period from the launch of ChatGPT in January 2023 through June 2024, during which public discourse on GAI surged. Using the social amplification of risk framework (SARF), a well-established theory in risk communication research, this study examines how GAI risks are amplified or attenuated in Korean society.

SARF posits that an individual's perception of risk is influenced by social and cultural factors, with amplification occurring through "social stations" such as media, experts, civil society, and personal networks. The framework emphasizes that risk perception varies across different social and cultural environments, making it particularly relevant to South Korea. As a global ICT leader with 97.4% internet penetration and advanced mobile connectivity (International Telecommunication Union, n.d.), South Korea represents a unique case for studying GAI. Korean companies like Naver and Kakao dominate local news and internet service platforms, and their entry into the GAI market, where they compete with global ICT companies such as Google, OpenAI, and Amazon, offers valuable insights into responses in an ICT-sensitive society (Internet Trend, n.d.).

This study contributes to the literature by employing topic modelling to analyze news coverage based on SARF and examining public comments to capture direct societal reactions. By empirically exploring the relationship between media and public discourse, which is traditionally viewed as a mechanism of risk amplification, this research sheds light on the dynamics of risk communication in the context of GAI.

2. Theoretical Framework

2.1. SARF

The definition of "risk" has evolved throughout history, reflecting shifts in societal and cultural contexts. Risk can be viewed objectively, as a quantifiable phenomenon assessed through technical expertise, or subjectively, as a construct shaped by societal values and perceptions (Kim, 2006).

SARF was first conceptualized by Kasperson et al. (1988) to explore how risk-related events, initially assessed as minor or technical by experts, can evolve into major societal concerns. The framework integrates psychological, organizational, social, and communicative processes to explain how risk signals are amplified or attenuated through "amplification stations," such as media, government, and public discourse. These stations act as filters that shape the transmission and reception of risk signals, ultimately influencing public perception and societal responses (Kasperson et al., 1988; Song et al., 2012).

SARF's utility lies in its ability to analyze the ripple effects of risk events, originally likened to waves in a pond. With the advent of the internet and social media, these ripple effects have become increasingly complex and interconnected, requiring a nuanced understanding of how digital platforms act as amplification stations (Chung, 2011; Kasperson et al., 2022). SARF has also been instrumental in understanding the cultural and societal dimensions of risk perception, highlighting how values, beliefs, and institutional responses interact to shape risk dynamics.

2.2. Works related to SARF

SARF has been extensively applied to diverse contexts, including natural disasters, technological risks, and health crises. From the perspective of communication studies, the concept of risk has been increasingly linked to the dynamics of public discourse and societal responses. This connection, especially in risk communication research, led to the emergence of SARF.

Early studies primarily analyzed traditional media's role in risk amplification. For instance, Renn (1991) and Pidgeon et al. (2003) examined how media framing influences societal reactions to environmental hazards. Similarly, Crespi and Taibi (2020) highlighted how German news media amplified perceptions of earthquake risks in Italy by emphasizing uncertainty and dramatic outcomes. SARF offers risk communication scholars a useful conceptual tool for examining the social experience of risk by extending our understanding of news media as a component of the framework (Binder et al., 2014).

The rise of social media has significantly influenced SARF research, as platforms like X (formerly Twitter) and Facebook act as dynamic amplification stations. Fellenor et al. (2017, 2020) explored X's role in amplifying public concern during the 2012 bubonic plague outbreak in the UK. They demonstrated how social media blurs boundaries between journalists and consumers, enabling the rapid dissemination and amplification of risk signals.

Schmid-Petri et al. (2023) collected and analysed tweets about Covid-19 vaccination among German, Russian, Turkish, and Polish groups to measure information gaps between specific demographic groups

about vaccination during the Covid-19 pandemic. Using SARF, E. W. J. Lee et al. (2023) identified 11 key themes based on tweets in public discourse on Covid-19, including health impacts, economic consequences, and public calls for action. Survey research building on SARF identified the importance of online discussion in influencing the spread of risk information during the early stages of Covid-19 outbreaks when publics rely primarily on social media for information (J. Lee et al., 2023; Zhang & Cozma, 2022).

In the context of emerging technologies, SARF has been employed to analyze public perceptions of AI and digital innovations. Recent studies on AI have begun to explore its societal and ethical implications. Neri and Cozman (2020) examined shifts in public sentiment toward AI by analyzing X data over a decade. Park et al. (2022) analyzed AI-related news articles from Korea and the United States, employing topic modelling to identify dominant narratives. Beltran et al. (2024) examined GAI usage guidelines across several countries, highlighting risks such as data privacy concerns, security threats, and public trust issues. By analyzing a sample of 501 of the most-viewed YouTube videos about AI, Schwarz (2024) found that frames with a higher emphasis on the societal threat of AI were more likely to be viewed and commented on by users. Furthermore, Leiter et al. (2023) and Taecharungroj (2023) utilized social media analysis to capture the rapid evolution of public discourse surrounding ChatGPT within short timeframes.

2.3. Research Questions

Despite growing interest in SARF applications in digital technologies, a limited number of studies have addressed its role in analyzing GAI discourse, particularly in Korea. This study applies SARF to examine the public discourse on GAI in Korea. By focusing on interactions within amplification stations such as news media and public commentary, the framework enables a systematic analysis of how risk perceptions surrounding GAI are amplified or attenuated in the Korean context. Using the bidirectional encoder representations from transformers (BERT) model, a machine learning algorithm, this study seeks to identify the thematic structures and dynamics influencing the public's risk perception of GAI. By leveraging SARF, this study seeks to address the following research question:

> RQ1: Is there a difference between how the media and the public perceive risk in the Korean public sphere?

SARF posits that amplification stations, such as media, shape risk signals differently depending on societal and cultural contexts. By analyzing news articles and public comments, this question examines the disparities in risk perception between news media and the public. Understanding these differences can provide insights into how media narratives influence public discourse on GAI risks.

The second research question focuses on identifying whether risk perceptions of GAI technologies are heightened or diminished through interaction with amplification stations. By employing the BERT model, this study explores the thematic structures and dynamics underlying public discourse, assessing the areas where risks are most amplified or attenuated. These findings can reveal key factors driving societal responses to GAI risks in Korea. As such, we propose the following research question:

> RQ2: Perceiving the risk of GAI, have the Korean public been amplified or attenuated through the amplification station?

Through these research questions, this study contributes to a deeper understanding of how SARF can be utilized to analyze the evolving dynamics of risk amplification in the context of emerging technologies.

3. Methods

3.1. Dataset

To analyze the Korean public sphere, it is essential to understand the unique news consumption patterns in Korea. Koreans rely on internet portal services for news more than citizens of any other country. According to the Reuters Journalism Institute, 69% of Korean users access news through search engines and news aggregators, more than double the global average of 33% across 46 countries (Newman et al., 2022).

In this study, we collected news distributed via Naver (http://www.naver.com), which holds the largest market share among portal services in Korea. For data collection, we focused on in-linked news on Naver. The dataset includes news articles published between January 1, 2023, and June 30, 2024, a period during which issues related to GAI began to gain significant attention.

We extracted news articles containing the keywords "Generative AI" or "Generative Artificial Intelligence" (in Korean: 생성형AI or 생성형 인공지능). Python was used to facilitate data collection. As a result, we gathered 56,468 articles from 115 media outlets, including national and local dailies, business newspapers, broadcasting, online news platforms, magazines, and news agencies.

To examine public reactions, we also collected comments posted on the analyzed articles. From the various comment-sorting options available on Naver—such as sorting by empathy, newest, oldest, empathy ratio, and many nested replies—the top 10 comments with the highest empathy were selected for analysis. In this context, empathy is measured as the number of likes minus the number of dislikes. A total of 68,393 comments were collected through this process.

3.2. Preprocessing Data

We collected data comprising news articles and their associated comments. Since the collected data is text-based and unstructured, it was necessary to preprocess it by removing unspecified words and breaking the text into morphological units. We focused on analyzing Korean lexical morphemes (e.g., nouns, adjectives, and verbs) with two or more syllables while filtering out words deemed unimportant or lacking meaningful content.

News articles can be analyzed at either the sentence level or the article level. For identifying distinct topics within an article and enabling more detailed topic analysis, sentence-level analysis offers a greater advantage. Following preprocessing, the articles were segmented into sentence units, generating a total of 1,781,121 documents.

In contrast, comments, which typically consist of short sentences or paragraphs and often include profanity, were analyzed as whole units rather than being divided into sentence-level components like the articles.

Comments with three words or fewer were excluded from the analysis because they were generally too ambiguous to be effectively interpreted. This resulted in 46,662 documents.

3.3. Analysis With BERT Model

The objective of this study was to use text-mining techniques to investigate how the risks associated with GAI are amplified or attenuated in public discourse. Text mining methods are broadly categorized into two types: topic frequency analysis, which identifies frequently occurring words in a text; and association frequency analysis, which examines the frequency and correlation of co-occurring words. While topic frequency analysis is effective for identifying prevalent topics, it falls short in revealing relationships between them. Given the study's focus on interactions between amplification stations, association word frequency analysis was deemed more suitable.

Topic modelling enables the grouping of documents with similar meanings and the clustering of words with shared contexts into distinct topics. For instance, in a collection of documents on a specific topic, certain words are expected to appear more frequently than others. The most commonly used approach for this purpose is the latent Dirichlet allocation model. However, this model has a significant limitation in that it does not account for word order or sentence structure.

To address this limitation, we adopted the BERT topic modelling technique, which enhances the embedding performance of textual data (Grootendorst, 2022). BERT leverages robust contextual embeddings within the BERT framework, combined with a class-based term frequency-inverse document frequency algorithm. This approach facilitates the comparison of term importance within dense clusters and the development of refined term representations (Sánchez-Franco & Rey-Moreno, 2022).

4. Results

4.1. Analysis of News Articles and Comments

Analysis of the 56,468 news articles in this study reveals a steady increase in frequency over time, interspersed with sudden spikes during specific periods. These concentrated bursts of news coverage can be interpreted as the media's efforts to set an agenda and amplify certain topics. Major events related to GAI, highlighted by the media, act as risk signals.

To understand the overall themes within the dataset, we applied the BERT model to both news articles and comments. To ensure meaningful clustering and avoid the creation of numerous small clusters, we merged similar topics and capped the maximum number of topics at 60.

4.1.1. Topic Analysis of News Article

First, we conducted BERT model testing on the news articles, extracting the six topics with the highest document frequencies (see Table 1 and Figure 1). Among the analyzed documents, Topic 0 had the largest representation, encompassing 59,835 documents. We labelled this topic "enterprise business," as it focused on how organizations respond to GAI adoption. Articles discussed activities by Kakao, a leading Korean ICT

company, as well as developments in advertising and cloud utilization. Key terms included: cloud, advertising, big, engine, forum, Kakao, open, office, startup, and technology. Examples of related headlines include: "ChatGPT Writes Advertise Copy: Get the Point vs not Creative" and "Kakao Opens Beta for Korean ChatGPT Da-Daum: Developed as a Prototype."

Topic 1 centred on "GAI and robots," exploring the integration of GAI technologies into robotics and deep learning services. Key terms included: artificial intelligence, robot, intelligent, language, learning, image, English, interpreter, and Siri. This topic included headlines such as: "Indigenous Cloud Companies Laugh at Last Year's Results: Public-AI Demand Is Bigger This Year" and "MS Combines Generative AI Co-Pilots in Office: Changing the Way We Work."

Topic 2 was labelled "game changer of the GAI era" and showcased innovative business models such as Microsoft's co-pilot. Prominent terms included: cloud, game, centre, Copilot, processing, ultra, data, chain, and core. Examples of related headlines are: "The Keyword of the Year is Speed: The Era of 6G, Robot, and AI is Upon us" and "Talk to It, Play It Music, and It Will Draw a Picture for You."

Topic 3 focused on "Samsung and Hynix," emphasizing the role of Korean semiconductor companies in the GAI landscape. Key terms included: electronics, chairman, generation, Samsung, academic, Hynix, certification, group, and session. Relevant headlines are: "AI-Driven Semiconductor Big Bang: K-Semiconductor, Opportunities and Risks" and "Morgan Stanley: Samsung Electronics and SK Hynix Are Also AI Beneficiaries."

Topic 4 addressed changes in the stock market, highlighting fluctuations in financial performance linked to GAI developments. Key terms included: profit, operating, increase, contrast, net profit, forecast value, and decrease. Relevant headlines include: "Gartner: Worldwide IT Spending Next Year to Increase 8% Over This Year, Led by AI Investments" and "OpenAI, Which Was in the Red, Expects 1.3 Billion Won in Annual Revenue on ChatGPT Jackpot."

Table 1. Topic analysis of news articles.

Topic	Label	Weights	Documents	Keywords
0	Enterprise business	0.034	59,835	cloud, advertising, big, engine, forum, Kakao, open, office, startup, technology
1	GAI and robot	0.022	39,657	artificial intelligence, robot, intelligent, language, learning, image, English, interpreter, Siri
2	Game changer of GAI era	0.018	32,170	cloud, game, centre, Copilot, processing, ultra, data, chain, core
3	Samsung and Hynix	0.015	27,510	electronics, chairman, generation, Samsung, academic, Hynix, certification, group, session
4	Changes in the stock market	0.014	24,729	profit, operating, increase, contrast, net profit, forecast value, decrease
5	Stock investment	0.013	22,550	investment, stock, management, fund, investment trust, ant (small cap investors), dividend, share

Notes: "Documents" refers to the number of sentence-level articles assigned to each topic; "Weights" is the number of documents in each topic divided by the total number of documents (N = 1,781,121).

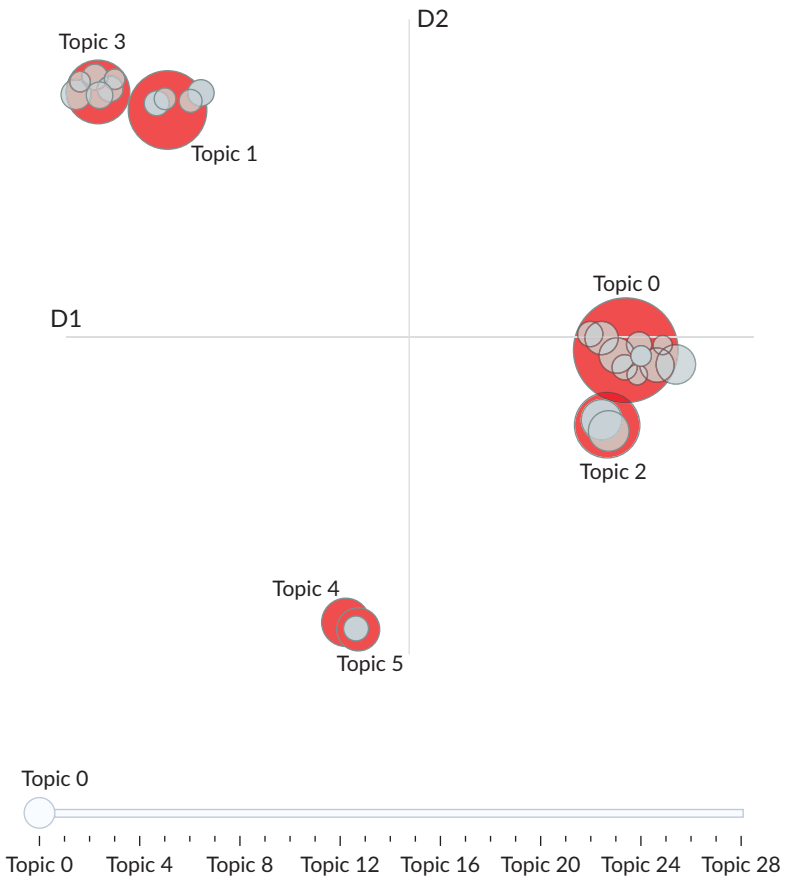


Figure 1. Intertopic distance map of news articles. Note: The red circles are the top six topics.

Topic 5 covered stock investment, emphasizing investing in the stock market for the masses. Key terms were: investment, stock, management, fund, investment trust, ant (small cap investors), dividend, and share. Relevant headlines were: "Stock Investment AI Will Also Become a Game Changer" and "Amazon Invests in Companies Combining AI and Robots: Creates 1.3 Trillion Won Fund."

Beyond these top six topics, additional themes such as chatbot services, smartphones, and information security also emerged prominently.

4.1.2. Topic Analysis of News Comments

BERT model analysis of news comments was conducted using the same approach as for the news articles. While the labelling of news articles predominantly highlights industry-related topics such as semiconductors and smartphones, the comments are primarily dominated by negative themes, including gaming regulation, cryptocurrency losses, and fake news (see Table 2 and Figure 2).

Topic 0 contained the largest number of comments (n = 3,771) and was labelled "short selling." This topic reflects responses to the impact of the GAI outbreak on South Korean semiconductor companies. Key terms include: stock, short selling, ant (small cap investors), stock price, Samsung Electronics, semiconductor, Hynix, shareholder, stock market, and investment. Examples of comments within this topic include: "Korea's Samsung and Hynix should support and grow alongside small Korean companies developing system AI and

semiconductors. This might be one way to stay ahead of the competition from Taiwan and the United States," "I'm sure short stock sellers will soon exploit this article to incite people to short secondary batteries," and "meanwhile, attention is being drawn to Nvidia and SK Hynix's monopoly on production."

Topic 1, labelled "iPhone and Galaxy," focuses on the trends and implications of GAI adoption by major smartphone companies like Apple and Samsung. Key terms include: Apple, iPhone, Galaxy, Nvidia, Samsung, innovation, Google, smartphone, Jobs, and features. Example comments include: "When the iPhone was first released, I thought, What's the big deal about having the internet on your phone?," "when AI like ChatGPT becomes cheaper and apps using it become available, it will be a different world," and "Samsung would be bigger than Nvidia if they went to the US, but the market is not big because of short stock selling on a tilted playing field."

Topic 2, which focuses on "ICT companies," discusses the challenges and opportunities faced by firms such as Naver, Kakao, and Google in the GAI era. Key terms include: Naver, Kakao, search, Google, advertising, blog, stock listing, Coupang, shopping, and search engine. Example comments include: "The metaverse may not have shaken any major companies, but ChatGPT has sparked concerns that even prominent firms like Naver, Kakao, and Google might struggle to survive," "Bard or GPT shows a lot of hallucinations and performance drops for Korean prompts. Even if it's Konglish [Korean English], Bard shows good performance for the same English prompt," and "the lack of Korean data is more of a problem with Naver than with Google or OpenAI looking down on Korea."

Topic 3, labelled "game regulation," explores public opinions on the regulation of gaming and related technologies. Key terms include: drawing, game, regulation, author, web comics, copyright, technology, graphics, cloud, and work (of art). Example comments are: "All those who advocate for AI regulation that will never happen will be labelled as enemies of AI and disappear," "is there a gaming service that lets you play with an AI that learns without a player?," and "comic and web comic authors are worried about copyright. While they're happy about the progress in AI, they want to talk about the copyright problems that come with it."

Topic 4 addresses "cryptocurrencies and damages" and focuses on issues such as fraud, financial losses, and compensation linked to cryptocurrencies. Key terms include: coin, compensation, bitcoin, bank, principal, disaster, fraud, people, impeachment, and finance. Some comment examples are: "Is Nvidia stock selling well now, or did it sell well during the coin craze 4 years ago?," "labor industry will gradually reduce jobs by robots, and AI will reduce jobs in white-colored jobs such as simple office work, etc. Banks, media companies are overflowing," and "that ChatGPT is going to be a ball of fire, a human-made disaster. It reminds me of the movie Terminator. A terrible world where machines overpower humans and keep them as servants!"

Topic 5, labelled "mobility," discusses the advancement of autonomous driving technology as electric cars become more widespread. Key terms were: Tesla, electric car, bus, metaverse, driving, battery, self-driving, car, taxi, and battery. Some examples follow: "If you want to invest in real AI, buy Tesla stock," "GAI will allow us to create an infinite amount of VR AR content in the Metaverse. GAI is the key to the metaverse, and once the device revolution comes, we'll never see humans crossing the street with their necks craned," and "in a few years, we can imagine wearing Vision Pro and taking Tesla into self-driving mode and if we get in an accident: who's to blame? The driver, Tesla, Apple?"

Beyond the top six topics, additional themes emerged, including judgment and politics, fake news, marriage and childbirth, and English and GAI.

Table 2. Topic analysis of public comments.

Topic	Label	Weight	Documents	Keywords
0	Short selling	0.081	3,771	stock, short selling, ant (small cap investors), stock price, Samsung electronics, semiconductor, Hynix, shareholder, stock market, Investment
1	iPhone and galaxy	0.065	3,035	Apple, iPhone, Galaxy, Nvidia, Samsung, innovation, Google, smartphone, (Steve) Jobs, features
2	ICT companies	0.031	1,442	Naver, Kakao, search, Google, advertising, blog, stock listing, Coupang, shopping, search engine
3	Game regulation	0.030	1,400	drawing, game, regulation, author, web comics, copyright, technology, graphics, cloud, work (of art)
4	Cryptocurrencies and damages	0.028	1,325	coin, compensation, bitcoin, bank, principal, disaster, fraud, people, impeachment, finance
5	Mobility	0.027	1,253	Tesla, electric car, bus, metaverse, driving, battery, self-driving, car, taxi, battery

Notes: "Documents" refers to the number of comments assigned to each topic; "Weights" is the number of documents in each topic divided by the total number of documents (N = 46,662).

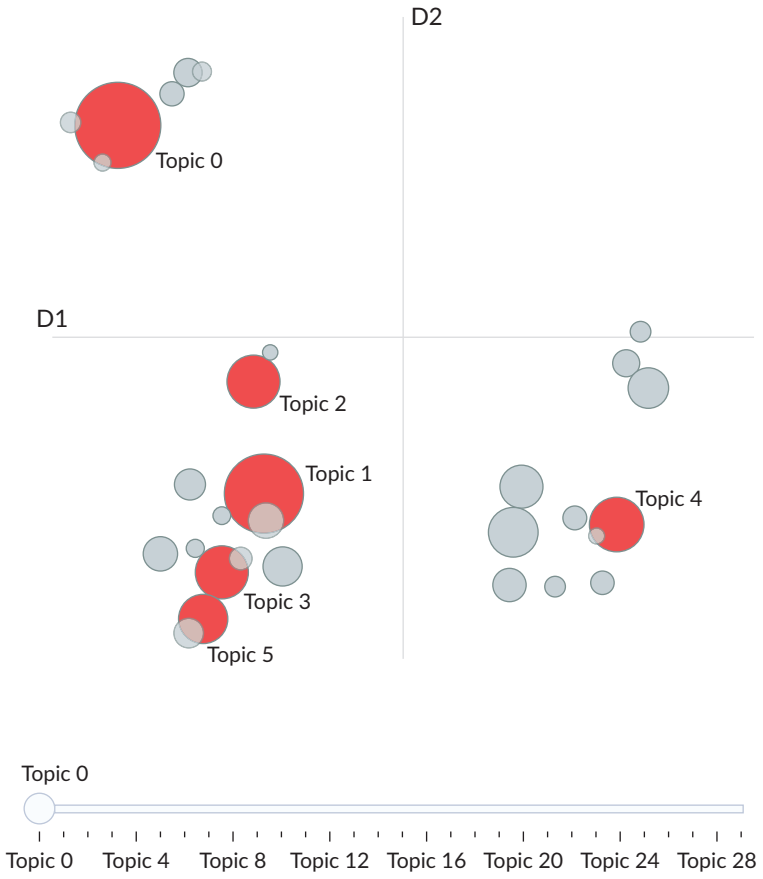


Figure 2. Intertopic distance map of the top 30 topics in news comments. Note: The red circles are the top six topics.

4.2. Identified Risk and the Public

BERT model analysis identified the communication flows in news articles and public comments that act as amplification stations in the Korean public sphere. However, this analysis was limited in its ability to pinpoint the processes underlying the amplification or attenuation of activity within the SARF. To adopt a more empirical approach, we assessed the relevance of the analysis results regarding established GAI-related risk factors.

We began by identifying GAI-related risk factors through a literature review. Wach et al. (2023) outlined seven controversies and threats associated with GAI from a management and economics perspective. Similarly, Beltran et al. (2024) analyzed government-issued guidelines for GAI usage in Australia, Canada, New Zealand, the United Kingdom, and South Korea, identifying 22 risk factors. Beltran et al. (2024) further measured how these guidelines reflected the risk factors, finding that the South Korean government's guidelines weighted leakage (41.7%) and hallucination (25%) most heavily, followed by privacy, intellectual property, and bias concerns (see Table 3).

Building on these studies, we empirically measured the differences in risk perceptions between media, which reflects expert perspectives, and public discourse from a SARF perspective. This involved comparing their alignment with the analysis results and the risk factors defined by Wach et al. (2023) and Beltran et al. (2024). We vectorized Korean data from news articles and comments within each topic. We also vectorized the 22 risk factors' names and definitions. The cosine similarity between these vectors was then calculated using the following formula where A = vector of the 22 risk factors and B = vector of BERT model results:

$$\text{cosine}_{\text{similarity}(A,B)} = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Table 3. Risks related to GAI.

Controversies and risks of GAI	Identified risk	Definition
Poor quality, lack of quality control, disinformation, deepfake content, algorithmic bias	Authenticity	GAI can intensify the spread of fake information.
	Explainability	Difficulty interpreting and understanding GAI outputs, leading to challenges in identifying errors and trust issues.
	Hallucination	GAI may generate nonsensical or incorrect outputs.
	Harmful content	Content produced by GAI could be violent, offensive, or harmful.
	Public trust	The use of GAI raises significant concerns about public trust and may lead to its erosion.
	Quality of training data	GAI can produce erroneous outputs due to inadequate or low-quality training data.
	Misuse	Potential for using GAI in plagiarism or cheating. GAI may exacerbate the digital divide, impacting individuals and communities with varying access to and acceptance levels of this technology.

Table 3. (Cont.) Risks related to GAI.

Controversies and risks of GAI	Identified risk	Definition
Widening socio-economic inequalities	Bias	GAI may show unfair favouritism or discrimination against certain individuals or groups.
	Digital divide	GAI may exacerbate the digital divide, impacting individuals and communities with varying access to and acceptance levels of this technology.
	Environmental	GAI incurs a substantial environmental cost, mainly due to the significant greenhouse gas emissions associated with its development and use.
	Income inequality and monopolies	GAI exacerbates income disparities by favouring those with AI proficiency and resources and potentially leading to resource and power monopolization by large companies.
	Industry disruption	GAI can transform competitive dynamics across industries, potentially leading to market dominance by a few players.
	Over-reliance	Users can become extremely dependent on GAI, hindering critical thinking and problem-solving skills.
Personal data violation, social surveillance, and privacy violation	Cybersecurity	The vulnerability of GAI to unauthorized access, manipulation, and data theft poses significant threats to the integrity and confidentiality of operations and sensitive data.
	Leakage	Dissemination of sensitive information or intellectual properties of the organization.
	Privacy	GAI may lead to the loss, alteration, or unauthorized disclosure of personal data and infringe on individuals' privacy rights.
AI-related technostress	Governance	Issues with human control over AI behaviour, interoperability, and data fragmentation. The use of GAI raises significant concerns about public trust and may lead to its erosion.
	Prompt engineering	The quality of prompts can lead to errors or misunderstandings in AI responses.
Automation-spurred job losses	Labor market	Potential for job displacement and unemployment as a consequence of the integration and advancement of GAI in various sectors.
	Professional standards	Using GAI to complete tasks requiring a professional license (e.g., medical diagnosis or legal advice) can breach regulations or professional guidelines.
No regulation of the AI market and urgent need for regulation	Intellectual property	GAI can contravene copyrights, trademarks, or patents.
Social manipulation, weakening ethics, and goodwill	Liability and accountability	Using GAI can involve an unclear assignment of responsibility for GAI errors or harm.

Notes: "Controversies and risks of GAI" was cited in Wach et al. (2023) and "Identified risk and Definition" was cited in Beltran et al. (2024).

The analysis revealed noticeable differences in the similarity scores between news articles and comments, as outlined in Table 4. For news articles, the topic with the highest similarity score was cybersecurity, which

Table 4. Differences in GAI risk perception by amplification station.

Rank	Korea's government	News articles	Comments
1	Leakage	Cybersecurity (0.576)	Misuse (0.507)
2	Hallucination	Misuse (0.558)	Industry disruption (0.506)
3	Privacy	Industry disruption (0.518)	Governance (0.492)
4	Intellectual property	Governance (0.510)	Labor market (0.483)
5	Bias	Hallucination (0.508)	Leakage (0.482)

Note: The numbers in parentheses are the cosine similarity between the data and the identified risk.

scored 0.576. This was followed by misuse with a similarity score of 0.558, industry disruption at 0.518, governance at 0.510, and hallucination at 0.508. In contrast, for comments, the highest similarity score was observed for misuse, with a score of 0.507. This was closely followed by industry disruption, which scored 0.506, and governance, which had a similarity score of 0.492. Furthermore, the topics of labor market and leakage were significant, with scores of 0.483 and 0.482, respectively.

These results highlight the differing emphases placed on specific risk factors by news articles and public comments, reflecting variations in the perception and prioritization of GAI-related risks between these two amplification stations.

The average similarity value across all topics was higher for news articles (0.280) than for comments (0.212). Among news articles, the highest average similarity scores were observed for misuse (0.294), followed by cybersecurity (0.257), prompt engineering (0.257), professional standards (0.251), and industry disruption (0.245). In the case of comments, the highest averages were noted for industry disruption (0.366), cybersecurity (0.360), misuse (0.350), professional standards (0.325), prompt engineering (0.312), and liability and accountability (0.311). However, these results are not significant since similarity is generally considered meaningful only when the measure is greater than or equal to 0.5.

Upon visualizing the similarity analysis results through heatmaps, notable differences emerged (see Figures 3 and 4). The similarity results between comments and risk factors showed relatively high similarity to specific topics (Topic 11), but overall low similarity between other topics. Conversely, the heatmap between articles and risk factors demonstrated a relatively distinct pattern of similarity across many topics.

From a qualitative perspective, there were also notable differences in how risk factors were prioritized by the two amplification stations. Analysis of news articles identified cybersecurity as the most significant risk factor, while comments highlighted misuse as the top concern, alongside labor market and leakage issues. Both news articles and comments recognized misuse and industry disruption as key concerns. However, neither amplification station gave significant attention to the South Korean government's top-ranked risk factors: leakage, hallucination, privacy, intellectual property, and bias.

Despite these differences, the two amplification stations demonstrated some alignment in their classifications of risk factors. The controversies and risks of GAI, as outlined by Wach et al. (2023), were addressed in both stations within five overarching categories: (a) poor quality, lack of quality control, disinformation, deepfake content, and algorithmic bias; (b) widening socio-economic inequalities; (c) personal data violations, social surveillance, and privacy breaches; (d) AI-induced technostress; and (e) job losses driven by automation. This

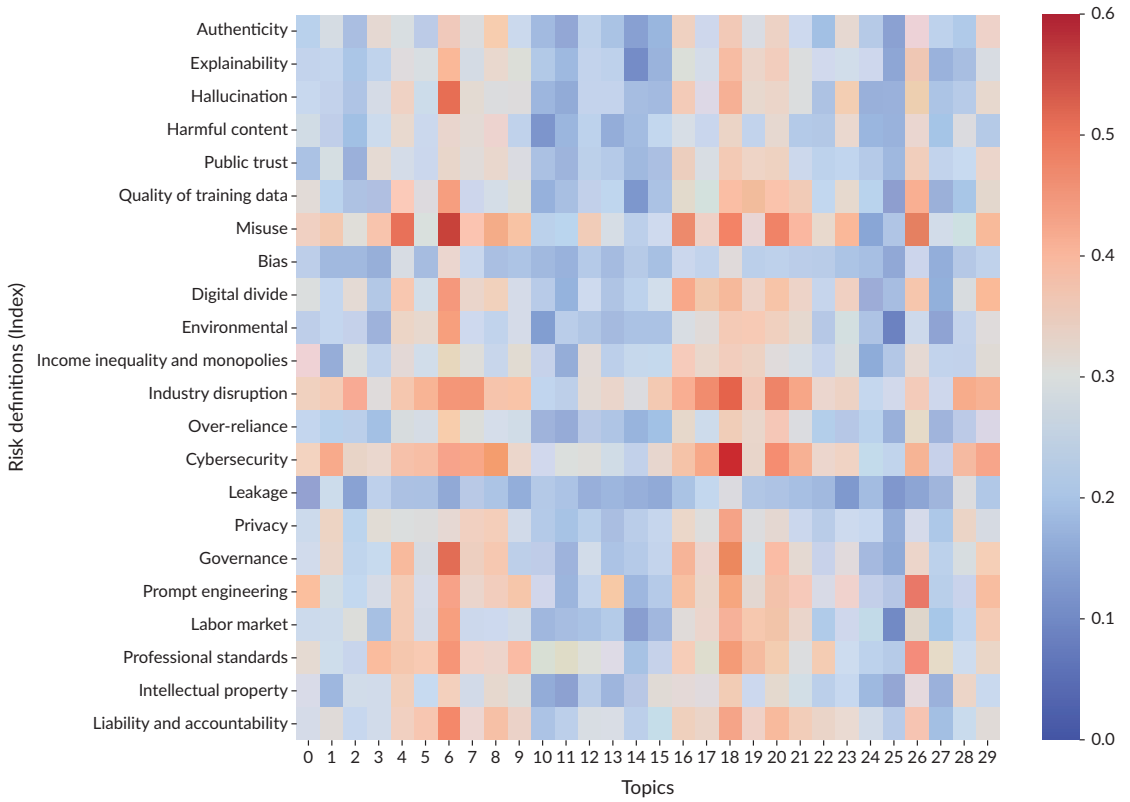


Figure 3. Similarity analysis: Identified risk keywords and BERT model of the top 30 topics in news articles.

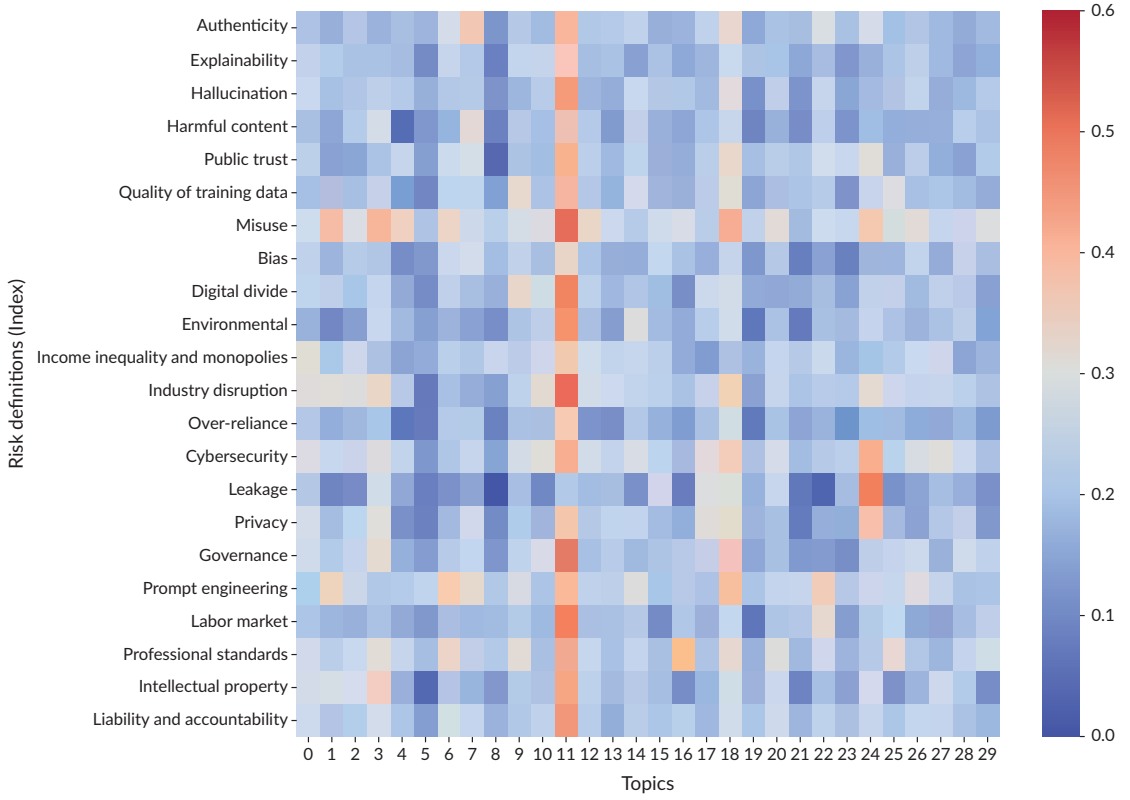


Figure 4. Similarity analysis: Identified risk keywords and BERT model of the top 30 topics in comments.

alignment underscores shared concerns about certain critical risks, even as the emphasis on specific factors varies between news articles and public comments.

5. Conclusion and Discussion

5.1. Implications of Research

Using the framework of the SARF, this study examined how the risks associated with GAI are amplified or attenuated within the Korean public sphere. Specifically, to evaluate differences in risk perception between the media and the public, we employed the BERT model analysis to identify themes in news articles and public comments. The findings yield several important implications for understanding risk dynamics.

First, the analysis of news articles and comments reveals that perceptions of GAI-related risks differ based on the amplification station. From the SARF perspective, media outlets—often shaped by expert contributions—traditionally act as amplification stations. The BERT model results show that media coverage has largely focused on keywords highlighting the impacts of GAI on various industries. Specifically, the most prominent topics in news articles include robotics, semiconductors, and smartphones, reflecting the media's emphasis on industry-level consequences of GAI. In contrast, public comments, which function as a public amplification station, set an alternative agenda for GAI-related risks, spotlighting issues such as gaming regulations, cryptocurrency concerns, and the spread of fake news. Moreover, they also reflect doubts about the relevance of learning foreign languages in an era dominated by GAI technologies. This contrast highlights the divergence in focus between media narratives and public discourse.

Second, the definition of risk factors for GAI further supports the idea that the public's perception of risk differs significantly from that of the media. To objectively identify these differences, we applied a similarity measure for each topic instead of relying solely on thematic analysis. Both the public and the media, as amplification stations, address major risk themes associated with GAI but differ in the specific risks they prioritize. Notably, the public tends to amplify concerns about misuse and labor market disruptions, whereas the media and government emphasize other risks such as cybersecurity and industry disruption.

Third, the study found that the amplification effect of news media within the amplification station was limited. Consistent with prior SARF-based studies, the topics amplified by news media were not always reflected in public comments. This suggests that the risk agenda set by the media does not necessarily align with the risk perceptions expressed in public discourse.

This selective amplification and attenuation of risk, arising from the interaction between news articles and comments, is likely to influence the public forum's response to GAI-related risks. These dynamics will, in turn, shape the subsequent phases of the SARF, including the ripple effects and impact phase. Based on the findings, public debate on the risks of GAI is expected to converge around specific topics and is likely to prioritize the impacts of industry changes over broader issues such as the ethics of AI, global regulatory frameworks, or transformations in the knowledge economy.

5.2. Limitations and Suggestions

This study builds upon existing academic research to identify new dimensions of social risk perception. However, several limitations should be acknowledged. First, this study did not establish a fully objective metric to quantify the degree of risk amplification or attenuation in news articles and comments. Although the SARF inherently lacks tools for measuring the extent of these processes, we attempted to address this gap by using word similarity as a quantitative measure. While this approach offered a partial solution, future research could benefit from incorporating more diverse and robust methodologies to enhance the analysis and provide a clearer understanding of amplification and attenuation dynamics.

Second, the study relied on risk factors identified in prior studies to measure the similarity between news articles and comments. While this approach yielded valuable insights, it may not have captured the broader or evolving spectrum of risks discussed in public forums. Public perceptions of risk are dynamic and multifaceted, often influenced by emerging issues and contextual shifts. Future studies could develop improved metrics and frameworks that account for the variability and complexity of public discourse on risk-related topics.

Third, the role of social media as an amplification station remains a significant challenge in risk communication research. This study focused on analyzing news comments, which are inherently shaped by the agendas set by news agencies. As a result, they may not fully reflect the public's active, autonomous responses to risk signals. To overcome this limitation, future research could expand in scope to include other social media. Platforms such as X, Facebook, or YouTube may provide a more comprehensive and diverse perspective on public engagement with risk-related issues, offering insights into how risks are perceived and debated across different digital spaces.

By addressing these limitations, future studies can contribute to a more nuanced and holistic understanding of how risks are communicated, perceived, and amplified in the public sphere. This understanding can in turn inform more effective strategies for risk management and communication.

Acknowledgments

The authors used generative artificial intelligence (ChatGPT 4.0; DeepL) partly for the coding in Python, translating, and proofreading.

Conflict of Interests

The authors declare no conflict of interests. In this article, editorial decisions were undertaken by Jeong-Nam Kim (University of Oklahoma).

References

Beck, U. (2012). Global risk society. In G. Ritzer (Ed.), *The Wiley-Blackwell encyclopedia of globalization*.

Beltran, M. A., Ruiz Mondragon, M. I., & Han, S. H. (2024, June). Comparative analysis of generative ai risks in the public sector. In H.-C. Liao, D. Duenas Cid, M. A. Macadar, F. Bernardini (Eds.), *dg.o '24: Proceedings of the 25th Annual International Conference on Digital Government Research* (pp. 610–617). https://doi.org/10.1145/3657054.3657125

Bengio, Y., Russell, S., & Musk, E. (2023). *Pause giant AI experiments: An open letter*. Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments

Binder, A. R., Cacciatore, M. A., Scheufele, D. A., & Brossard, D. (2014). The role of news media in the social amplification of risk. In H. Cho, T. Reimer, & K. A. McComas (Eds.), *The Sage handbook of risk communication* (pp. 69–85). Sage. https://doi.org/10.4135/9781483387918.n10

Chung, I. J. (2011). Social amplification of risk in the Internet environment. *Risk Analysis*, *31*(12), 1883–1896. https://doi.org/10.1111/j.1539-6924.2011.01623.x

Crespi, I., & Taibi, M. (2020). Cultural differences and social amplification of risk of a tourism destination: Foreign media coverage after 2016/2017 earthquakes in central Italy. *Italian Sociological Review*, *10*(2), 201–237. https://doi.org/10.13136/isr.v10i2.337

Cuthbertson, A. (2022, December 2). 'Google is done': World's most powerful AI chatbot offers human-like alternative to search engines. *Independent*. https://www.independent.co.uk/tech/ai-chatbot-chatgpt-google-openai-b2237834.html

Fellenor, J., Barnett, J., Potter, C., Urquhart, J., Mumford, J., & Quine, C. (2017). The social amplification of risk on Twitter: The case of ash dieback disease in the United Kingdom. *Journal of Risk Research*, *21*(10), 1163–1183. https://doi.org/10.1080/13669877.2017.1281339

Fellenor, J., Barnett, J., Potter, C., Urquhart, J., Mumford, J., & Quine, C. P. (2020). 'Real without being concrete': The ontology of public concern and its significance for the Social Amplification of Risk Framework (SARF). *Journal of Risk Research*, *23*(1), 20–34. https://doi.org/10.1080/13669877.2018.1501598

Gates, B. (2023). *The age of AI has begun*. Gates Notes. https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

Grant, N., & Metz, C. (2022, December 21). A new chat bot is a 'code red' for Google's search business. *The New York Times*. https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv. https://doi.org/10.48550/arXiv.2203.05794

International Telecommunication Union. (n.d.). *Korea (Rep. of)*. https://datahub.itu.int/data/?e=KOR

Internet Trend. (n.d.). *Inteonen teurendeu ripoteu*. http://www.internettrend.co.kr/trendForward.tsp

Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., Kasperson, J. X., & Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis*, *8*(2), 177–187.

Kasperson, R. E., Webler, T., Ram, B., & Sutton, J. (2022). The social amplification of risk framework: New perspectives. *Risk Analysis*, *42*(7), 1367–1380. https://doi.org/10.1111/risa.13926

Kim, Y.-W. (2006). Risk society and risk communication: Reflexivity on risk and the need of communication. *Keomyunikeisyeon iron*, *2*(2), 192–232.

Lee, E. W. J., Zheng, H., Goh, D. H., Lee, C. S., & Theng, Y. L. (2023). Examining Covid-19 tweet diffusion using an integrated social amplification of risk and issue-attention cycle framework. *Health Communication*, *39*(3), 493–506. https://doi.org/10.1080/10410236.2023.2170201

Lee, J., Choi, J., & Britt, R. K. (2023). Social media as risk-attenuation and misinformation-amplification station: How social media interaction affects misperceptions about Covid-19. *Health Communication*, *38*(6), 1232–1242. https://doi.org/10.1080/10410236.2021.1996920

Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., & Eger, S. (2023). ChatGPT: A meta-analysis after 2.5 months. *Machine Learning with Applications*, *16*, Article 100541. https://doi.org/10.1016/j.mlwa.2024.100541

Neri, A., & Cozman, F. G. (2020). The role of experts in the public perception of risk of artificial intelligence. *AI & Society*, *35*(3), 663–673. https://doi.org/10.1007/s00146-019-00924-9

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). *Reuters institute digital news report 2022*. Reuters Institute; University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022

Nolan, B. (2023, January 28). ChatGPT has only been around for 2 months and is causing untold chaos. *Business Insider*. https://www.businessinsider.com/chatgpt-ai-chaos-openia-google-creatives-academics-2023-1

Park, J. H., Kim, M. S., & Kim, J. H. (2022). How does the media deal with artificial intelligence? Analyzing articles in Korea and the US through big data analysis. *Jeongbosiseutem yeongu*, *31*(1), 175–195.

Pidgeon, N., Kasperson, R. E., & Slovic, P. (2003). *The social amplification of risk*. Cambridge University Press.

Renn, O. (1991). *Risk communication and the social amplification of risk*. Springer.

Sánchez-Franco, M. J., & Rey-Moreno, M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. *Psychology & Marketing*, *39*(2), 441–459. https://doi.org/10.1002/mar.21608

Schmid-Petri, H., Bürger, M., Schlögl, S., Schwind, M., Mitrović, J., & Kühn, R. (2023). The multilingual Twitter discourse on vaccination in Germany during the Covid-19 pandemic. *Media and Communication*, *11*(1), 293–305. https://doi.org/10.17645/mac.v11i1.6058

Schwarz, A. (2024). The mediated amplification of societal risk and risk governance of artificial intelligence: Technological risk frames on YouTube and their impact before and after ChatGPT. *Journal of Risk Research*. Advance online publication. https://doi.org/10.1080/13669877.2024.2437629

Song, H.-R., Cho, H.-M., Lee, Y.-K., & Kim, W.-J. (2012). A study on the conceptualization, structural analysis and domain establishment of risk communication. *Bunjaeng haegyeol yeongu*, *10*(1), 65–100. http://doi.org/10.16958/drsr.2012.10.1.65

Taecharungroj, V. (2023). "What can ChatGPT do?" Analyzing early reactions to the innovative AI chatbot on Twitter. *Big Data and Cognitive Computing*, *7*(1), Article 35. https://doi.org/10.3390/bdcc7010035

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(2), 7–24. https://doi.org/10.15678/EBER.2023.110201

Zhang, X. A., & Cozma, R. (2022). Risk sharing on Twitter: Social amplification and attenuation of risk in the early stages of the Covid-19 pandemic. *Computers in Human Behavior*, *126*, Article 106983. https://doi.org/10.1016/j.chb.2021.106983

About the Authors

Sunghwan Kim is a former newspaper journalist and currently works as the social collaboration leader at the IT platform company Kakao. He is a doctoral student at the Moon Soul Graduate School of Future Strategy at the Korea Advanced Institute of Science and Technology (KAIST). His research focuses on digital journalism and risk communication in the digital society, with a particular emphasis on risk assessment, public perception, and response strategies.

Jaemin Jung (PhD, University of Florida) is a professor at the Moon Soul Graduate School of Future Strategy at the Korea Advanced Institute of Science and Technology (KAIST). His research focuses on media management, media economics, and the impact of AI on journalism and media industries, with a keen interest in exploring how AI technologies are reshaping the landscape of news production and media consumption.

Media and Communication
2025 • Volume 13 • Article 9419
https://doi.org/10.17645/mac.9419

cogitatio

ARTICLE

Open Access Journal

AI Transparency: A Conceptual, Normative, and Practical Frame Analysis

Sónia Pedro Sebastião [1] and David Ferreira-Mendes Dias [2]

[1] Centro de Administração e Políticas Públicas, Instituto Superior de Ciências Sociais e Políticas, Universidade de Lisboa, Portugal
[2] Instituto Superior de Ciências Sociais e Políticas, Universidade de Lisboa, Portugal

Correspondence: Sónia Pedro Sebastião (ssebastiao@iscsp.ulisboa.pt)

Submitted: 15 October 2024 **Accepted:** 17 December 2024 **Published:** 3 April 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

This study aims to dissect the normative discourse about artificial intelligence (AI) transparency using frame analysis. By employing a predominantly deductive, qualitative, and interpretative approach, the research leverages a qualitative frame analysis informed by a literature review on AI ethics and transparency. The study examines various AI ethical frameworks and regulations—China's Next Generation Artificial Intelligence Development Plan, the OECD's Recommendation of the Council on Artificial Intelligence, the White House's Blueprint for an AI Bill of Rights, and the EU's Artificial Intelligence Act—to understand how transparency is framed, transparency's objects, the defined accountability, and the responsible entities for ensuring transparency in the production of AI information. The study highlights transparency as a core ethical principle for trustworthy AI, emphasising its importance in final outputs and throughout AI development and deployment stages for fostering public trust. The findings indicate variability in language, priorities, and approaches to transparency across different frameworks, influenced by their socio-political, economic, and cultural contexts. Despite encouraging transparency as an ethical principle, the study notes a need for concrete guidance for its practical implementation across different AI applications. This gap underscores the need for critical examination and improvement in governance to enhance transparency and accountability in AI development and deployment. The innovative methodological approach, combining qualitative frame analysis with a theory-driven codebook, offers a novel template for investigating key concepts and issues in AI ethics and governance.

Keywords

accountability; artificial intelligence; ethical frameworks; regulation; transparency

1. Introduction

Pioneers of artificial intelligence (AI) sustained that a machine could simulate any aspect of learning or intelligence if precisely described (Lungarella et al., 2007, p. 2). AI then made a name for itself in specialised transport systems and industrial and commercial sectors.

The definition of AI is challenging due to the complexity of human intelligence. AI has different evolutionary stages and "can be classified into analytical, human-inspired and humanized AI depending on its cognitive, emotional, and social competencies" (Kaplan & Haenlein, 2020, p. 39).

Questions about AI ethics have become increasingly important (e.g., Kaplan & Haenlein, 2020; Stahl et al., 2021). Complex ethical issues surround current and near-future AI systems, especially regarding AI's social and personal impact on humans (Liao, 2020). The dangers associated with AI relate to the gap between public understanding of science and the pace of policymakers' challenges, potentially leading to apathy, lack of responsibility and accountability, moral panic, and inadequate legislation. It is, therefore, essential to analyse what is being done to explain and foster an understanding of these systems in normative terms.

Researching the ethics of generative AI is significant due to the need to align AI development with human values, the increasing inequalities, accountability and transparency issues, and AI's transformative potential (e.g., Cath et al., 2018; Gil de Zúñiga et al., 2023; Roberts et al., 2021). There is a need to ensure AI development aligns with human values and supports social good. With the evolution of generative AI systems, it will be critical to study their societal impacts and develop ethical frameworks to guide their development and deployment in beneficial ways.

The logic of transparency is linked to the attribution of responsibility, suggesting that understanding how a system works enables identifying responsible parties for malfunctions or malpractice (Ananny & Crawford, 2018). When an AI system causes harm, responsibility is assigned to different entities involved in the AI's lifecycle, including the company, the developer team, and the AI system itself (Sullivan & Wamba, 2022).

Transparency, accountability, and explainability are paramount in AI systems. Transparency is among the quintessential principles in the global ethical frameworks for AI (Jobin et al., 2019). The opacity and the extensive scale of generative models pose a significant challenge in elucidating their internal reasoning processes. Thus, facilitating a transparent rationale for the outputs generated by these models to the stakeholders affected emerges as a critical ethical concern warranting rigorous examination. Nonetheless, the discourse extends to interrogate the entities to whom AI should exhibit transparency and whether, in scenarios marked by technical constraints, the quest for transparency should compromise the performance of more transparent systems.

Using frame analysis, we aim to understand the normative discourse on AI transparency. This approach elucidates the implicit assumptions, priorities, and normative orientations encapsulated within ethical guidelines and regulations. It discloses the foundational frames, thereby revealing transparency in conceptual, normative, and practical dimensions. This analytical scrutiny may contribute towards a more refined and productive governance of AI technologies. A critical frame identification and examination are imperative for advancing transparency and accountability in developing and deploying AI technologies.

It allows for capturing multiple perspectives, uncovering underlying assumptions, identifying dominant and marginalised frames, enabling comparative analysis, and enhancing policy relevance.

This article is organised into three sections. The theoretical section discusses the concepts of AI and transparency. Based on the contributions of several authors, a theory-driven framing model is built. The third section presents the methodological approach, followed by the empirical study, discussion, and conclusions.

2. AI and Transparency

The evolution of AI systems has seen several setbacks and disappointments. As demonstrated by Gil de Zúñiga et al. (2023), there have been various definitions of the concept, some more machine learning-centric, others focusing on functions, cognitive simulation, and the creation of autonomous agents. Those definitions tend to be narrow in scope, broad, and vague, with human-centric bias and overemphasising autonomy. Despite providing valuable perspectives, those definitions' weaknesses undermine their comprehensiveness and applicability. Therefore, the authors propose a comprehensive definition of AI as "the tangible real-world capability of non-human machines or artificial entities to perform, task solve, communicate, interact, and act logically as it occurs with biological humans" (Gil de Zúñiga et al., 2023, p. 4).

Other AI definitions present a system's ability to correctly interpret external data, learn from it, and use the knowledge to achieve specific goals and tasks through flexible adaptation (Kaplan & Haenlein, 2019, p. 17), providing a competitive advantage to their holders. Both definitions stress the dimensions of performance and autonomy.

The field of AI study is cross-disciplinary and includes linguistics, cognitive sciences, neurosciences, robotics, engineering, computer science, social sciences, and humanities (Frankish & Ramsey, 2014). The advancement of technology has facilitated the shift from systems that imitate human intelligence and cognition to systems that generate content using generative AI. In the 21st century, its widespread adoption in personal technologies, multimedia content creation, and the evolution of generative pre-trained transformers and deep learning led to the AI systems' growing popularity and prevalence.

The opaque nature of deep learning raises concerns about interpretability, explainability, and trust. According to Liao (2020), deep learning is susceptible to adversarial attacks and errors, highlighting the importance of trust, interpretability, and explainability in fields like medicine and law, where human lives can be at stake.

Trustworthy AI requires transparency, including a broader socio-legal and computer-scientific perspective (Larsson & Heintz, 2020). Transparency is a concept that originated during the Enlightenment and involves the use of observation and knowledge to exercise social control (Hood, 2006). Therefore, it is a pervasive concept in political sciences studies, public and corporate governance, and communication studies. It is possible to identify different contexts in which transparency has been applied, such as in organisational and societal affairs, as a public value embraced by society to counter corruption, as a tool of good governance, and as a means of creating accountability, efficiency, and effectiveness (Larsson & Heintz, 2020).

As AI systems become embedded in more public systems' decision-making, Kemper and Kolkman (2019) advocate for more transparency in developing, implementing, and using algorithms in organisations.

According to Jobin et al. (2019), transparency refers to the clarity and openness concerning how AI systems operate, make decisions, and affect users and stakeholders. Transparency relates to how users and stakeholders can explain and understand AI systems and their decisions. Therefore, transparency is a conceptual metaphor associated with knowing and understanding (Larsson & Heintz, 2020).

AI systems' lack of transparency and accountability is a significant concern. Ananny and Crawford (2018) define transparency as seeing inside a system and understanding its mechanisms and decision-making processes. They note that transparency can be at the level of platform design and algorithmic mechanisms or, more deeply, at the level of a software system's logic.

Transparency requires disclosing information or revealing the interests of the issuer and holder of information (disclaimer). It also involves recognising and valuing transparency as an essential aspect of social control. The beholder must acknowledge and value transparency (Kemper & Kolkman, 2019).

Several typologies of transparency have been identified by considering categories such as types of information, objects of transparency, and accountability (see Table 1).

The ideal of transparency may not be the most suitable for AI ethics (e.g., Ananny & Crawford, 2018; Liao, 2020). While complete AI transparency may be ideal, it is only sometimes practical due to the complexity and potential trade-offs with other principles (Jobin et al., 2019; Liao, 2020). Ananny and Crawford (2018) argue that transparency is an ongoing process of scrutiny and adjustment, requiring more than just revealing the inner workings of AI systems. Ferrari et al. (2023) propose three structural conditions for effective AI governance: industrial observability, public inspectability, and technical modifiability. These conditions represent different levels at which AI systems must be transparent and accountable for effective oversight and regulation.

Table 1. Typologies of transparency.

Category	Type of Transparency	Description
Type of information	Fuzzy	The information provided does not reveal how institutions behave; the information is disclosed nominally or is unreliable.
	Clear	Reliable information is provided, for example, about institutional performance, responsibilities, and funds use.
Objects of transparency	Event	Event transparency focuses on disclosing specific data points, results, or impacts of a system's operations.
	Process	Process transparency aims to make visible the underlying logic, steps, and governance frameworks that determine how a system functions.
Accountability	"Soft" accountability	Organisations must answer for their actions when transparency is present.
	"Hard" accountability	Transparency brings the power to sanction organisations and demand compensation for the harm they cause.

Source: Adapted from Ananny and Crawford (2018).

However, Ananny and Crawford (2018) highlight the limitations of relying solely on transparency for accountability and understanding in AI, such as disconnection from power, professional boundary work, epistemological challenges, and the distributed nature of the actors involved. Even with information on how the system works, users may need help understanding algorithms or data usage, leading to a lack of trust and difficulty in holding the system accountable (Buiten, 2019).

The main challenges in regulating AI transparency are the complexity of the concept and the difficulty of providing technically feasible explanations helpful in specific legal contexts (Buiten, 2019). Despite these obstacles, various regulations prioritise transparency as one of their principles.

After conducting extensive bibliographic searches in the Scopus, EBSCO, and Web of Science databases using the keywords "fram*," "transparency," and "artificial intelligence" in titles, author keywords, and abstracts in English, French, and Portuguese, we were unable to find any prior studies or framing models related to AI transparency. Given this gap in the literature and the theoretical background presented previously, we propose a new framework for AI transparency. This framework includes key aspects such as understandability and explainability, accountability and governance, disclosure and communication, documentation and access to information, and ethical and legal compliance (see Table 2).

Table 2. Frames of transparency.

Frame	Description	Authors
Understandability and explainability	There is a need for AI systems to provide clear and understandable explanations of their decisions and processes. Ensuring that AI systems' functioning and decision-making processes are clear and understandable to users and stakeholders.	Ananny and Crawford (2018), Buiten (2019), Larsson and Heintz (2020)
Accountability and governance	Identifying and holding responsible parties accountable for the deployment and impacts of AI systems, ensuring transparent governance.	Larsson (2020), Sullivan and Wamba (2022)
User awareness and communication	Safeguarding users are aware when they interact with AI systems and understand the role of AI in decision-making processes. This includes effective communication about AI capabilities and limitations.	Ferrari et al. (2023), Kaplan and Haenlein (2019)
Documentation and access to information	Providing detailed documentation about AI systems' design, development, and functioning is essential for transparency. This includes making relevant information accessible to various stakeholders.	Corrêa et al. (2023), Ferrari et al. (2023), Kemper and Kolkman (2019), Larsson (2020)
Ethical and legal compliance	Transparency is framed as a means to ensure that AI systems comply with ethical standards and legal requirements. This includes adhering to principles of fairness, non-discrimination, data privacy, and human rights.	Corrêa et al. (2023), Kaplan and Haenlein (2020)

Framing is a process employed by message creators to organise and interpret information, making it one of the most widely applied theories in communication studies (Lock et al., 2020). It involves selecting a particular point of view to highlight specific message characteristics. Entman (1993) states that framing information significantly influences how people understand and react to issues. In the context of AI transparency, frame analysis helps elucidate the underlying assumptions, priorities, and normative directions embedded within these guidelines.

By utilising frame analysis, we aim to understand the normative discourse surrounding AI transparency. Transparency, accountability, and explainability are paramount in AI systems and need to be part of global ethical frameworks for AI (Jobin et al., 2019). But, how is transparency framed in ethical frameworks? (RQ1)

Therefore, the typologies presented in Table 1 raise research questions such as what are the objects of transparency (RQ2) and what kind of accountability is defined (RQ3)?

Finally, in light of the role of AI systems in algorithmic accountability and AI governance (e.g., Ferrari et al., 2023), another research question arises: Who is identified as responsible for ensuring transparency in the use of AI in producing and disseminating information? (RQ4)

The next section outlines the methodological approach, keeping in mind the research questions at hand.

3. Methodology

A predominantly deductive method of a qualitative and interpretative nature is used. The deductive approach is justified by theory-driven research (Bryman, 2016). A qualitative frame analysis based on Entman's (1993) value of framing is performed, using frames inferred from the literature review about AI ethics and transparency. The frame analysis of AI ethical guidelines provides a crucial lens for understanding how information about AI transparency is organised and can be interpreted.

This study focused on documents issued by transnational organisations (the OECD and the EU) and state organisations (China and the US), selected for their scope, recency, and relevance across different regulatory contexts (e.g., Corrêa et al., 2023). These frameworks represent the world's largest economic powers and most influential policy-setting organisations (Lee, 2018). The US, China, and the EU, along with the OECD, effectively shape AI governance for a significant portion of global AI development and deployment (Larsson, 2020).

The OECD Recommendation of the Council on Artificial Intelligence (RCAI) was launched in 2019 and is the first intergovernmental standard on AI (OECD, 2019). The EU Artificial Intelligence Act (AI Act; European Parliament, 2024) is the first transnational AI regulation. It was approved on 21st May 2024. Once adopted, it will be a binding legal act that must be applied across all EU member states. The AI Act aims to promote trustworthy and human-centred AI and establish a relationship with existing laws such as the General Data Protection Regulation and product safety, consumer protection, and labour law. As a regulation, it goes beyond a policy document and has the force of law.

Apart from these two binding documents, the analysis includes the Next Generation Artificial Intelligence Development Plan (AIDP) from China's State Council (2017) and the Blueprint for an AI Bill of Rights (BAIBR) from the White House Office of Science and Technology Policy (2022).

The AIDP outlines China's strategic goals, including becoming the world leader in AI by 2030, creating a significant AI industry, and using AI to drive economic development, social governance, and defence capabilities. It serves as a guiding document for China's national AI development, considering economic, political, cultural, and ethical factors.

The BAIBR contributes to the US regulatory and ethical AI development and deployment landscape. However, it only provides a set of non-binding principles and practices that aim to guide the design, development, and deployment of AI systems in a way that respects human rights and promotes public trust. It emphasises safe and effective design, protection against algorithmic discrimination, robust data privacy, transparency through notice and explanation, human alternatives, and availability of oversight. Table 3 organises the documents analysed, classifying them by year, issuer, nature of the document, and language.

Using the theory-driven frames described in Table 2, we aim to understand the normative discourse on AI transparency. Diverse actors have produced this discourse, including transnational institutions, states, research institutions, companies, NGOs, and professional associations (Corrêa et al., 2023; Jobin et al., 2019).

MAXQDA was selected for this study due to its advanced capabilities in qualitative data analysis, including automated lexical searches, hierarchical coding systems, and compatibility with various file formats. Its user-friendly interface and widespread academic adoption (Lewins & Silver, 2007; Woolf & Silver, 2018) further ensured its suitability for managing the extensive corpus of policy documents analysed in this study.

Given the normative nature and length of the corpus, we employed MAXQDA to automate the text search for the identified keywords. Our search included a lexical search for the keywords "transparency," "accountability," and "responsibility," incorporating lemmatisation to cover variations of these terms (e.g., transparent, accountable, responsible). The choice of search words is based on our literature review and the

Table 3. Description of the corpus.

Document Title	Issuer	Year	Number of Pages (Without Appendix/Annex)	Nature of the Document	Language
Next Generation AIDP	China's State Council	2017	28	Plan	English full translation provided by Stanford University
RCAI	OECD	2019	11	Policy	English
BAIBR	White House Office of Science and Technology Policy	2022	52	Guidelines	English
AI Act	European Parliament	2024	376	Regulation	English

realisation that the logic of transparency is associated with accountability, i.e., the attribution of responsibility (Ananny & Crawford, 2018; Jobin et al., 2019; Sullivan & Wamba, 2022). Additionally, we included the preceding and following sentences in the highlighted text segments.

These options allowed us to identify meaningful text segments, which we categorised according to the frames outlined in the codebook. The codebook was developed based on the frames presented in Table 2, along with the transparency objects and types of accountability outlined in Table 1. While the "fuzzy vs. clear" dimension was initially considered as part of the analytical framework, it was excluded after preliminary analysis for methodological and practical reasons. The dimension's subjective nature and overlap with other categories, such as understandability and explainability, posed challenges in ensuring consistent coding. Additionally, the lack of clear differentiation in the reviewed documents further justified its exclusion. This decision was made to maintain analytical rigour and focus on dimensions more directly aligned with the study's objectives. For example, the RCAI, AIDP, and BAIBR do not clearly assign responsibility for transparency in the production and dissemination of AI-enabled information to specific actors. In contrast, the AI Act is a legal document that provides clear guidelines on transparency, accountability, and the responsibilities of AI agents. However, it uses legal terminology and jargon that may not be easily understood by all AI users. To ensure the objectivity of our analysis and reach a consensus in coding, we decided to exclude this dimension.

It is important to note that all documents were reviewed beforehand to ensure that the coders were familiar with the texts. The lexical search enabled us to mark and categorise significant texts for each frame. The coding process helped us identify the key frames listed in Table 4, along with examples that we incorporated into the text. All authors agreed upon the selection of examples. The rigorous validation of coding decisions not only enhanced the reliability of the analysis but also ensured that the identified frames accurately reflected the normative and practical dimensions of AI transparency as presented in the analysed documents.

4. Results

Except for the AIDP, the analysed documents address issues of fairness, non-discrimination, data privacy, and human rights, but not necessarily in the context of transparency. The results associated with the research questions are presented below.

Considering RQ1—how is transparency framed in the AI ethical frameworks and regulations?—we note that the AIDP emphasises the importance of transparency in AI development but does not specify explainability requirements. The plan does not explicitly address documentation and access to information nor user awareness and communication. It serves as a directive setting the overall direction and priorities for AI development in China, with the expectation that various state and non-state actors will work towards these goals under the central government's guidance. The government is, therefore, responsible for establishing "a traceability and accountability system, and clarify the main body of AI and related rights, obligations, and responsibilities" (China's State Council, 2017, p. 25).

The RCAI defines AI as "a general-purpose technology that has the potential to: improve the welfare and well-being of people, contribute to positive sustainable global economic activity, increase innovation and productivity, and help respond to key global challenges" (OECD, 2019, p. 3). The document emphasises

transparency as a core value for responsible AI development. It includes: (a) user awareness and communication, which encourages communication with stakeholders about AI capabilities and limitations; (b) accountability and governance, where organisations and individuals responsible for AI systems should be identifiable and accountable, promoting transparency in governance and oversight (it does not specify governance structures); and (c) understandability and explainability, by ensuring that AI systems are transparent and understandable to users, stakeholders, and regulators, as illustrated in the excerpt:

> To enable those affected by an AI system *to understand the outcome*, and, (iv.) to enable those *adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information* on the factors, and the logic that served as the basis for the prediction, recommendation or decision. (OECD, 2019, p. 8, emphasis by authors)

The BAIBR underlines the need for transparency in AI systems to protect individual rights and promote trust. It focuses on automated systems, revealing them to users and explaining how they work. Although the document mentions other frames, the understandability and explainability frame is more emphasised in the context of transparency and accountability.

Some examples of segmented text regarding understandability and explainability, to ensure individuals are aware when an AI system is being used and provide explanations about how decisions are made, as illustrated in the excerpt: "An automated system should provide demonstrably clear, timely, understandable, and accessible notice of use, and explanations as to how and why a decision was made or an action was taken by the system" (White House Office of Science and Technology Policy, 2022, p. 43). Regarding documentation and access to information, by encouraging organisations to publicly disclose information about the use and impact of AI systems:

> Provide generally accessible plain language documentation including clear descriptions of the overall system functioning and the role automation plays, notice that such systems are in use, the individual or organisation responsible for the system, and explanations of outcomes that are clear, timely, and accessible. (White House Office of Science and Technology Policy, 2022, p. 6)

Concerning user awareness and communication, which emphasises the need for public update reporting, as illustrated in the excerpts:

> Audits and impact assessments to help identify potential algorithmic discrimination and provide transparency to the public in the mitigation of such biases. (White House Office of Science and Technology Policy, 2022, p. 24)

> The American public should be protected via built-in privacy protections, data minimization, use and collection limitations, and transparency. (White House Office of Science and Technology Policy, 2022, p. 33)

And lastly, concerning accountability and governance: "Entities responsible for the development or use of automated systems should lay out clear governance structures and procedures" (White House Office of Science and Technology Policy, 2022, p. 19).

The AI Act strongly emphasises transparency, particularly for high-risk AI systems. All frames are used to approach transparency, as illustrated in the following excerpts. Regarding understandability and explainability, where High-risk AI systems must be transparent and provide clear information to deployers: "High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers" (European Parliament, 2024, p. 194). Concerning accountability and governance, it establishes governance through national supervisory authorities and conformity assessments in several articles, sections, and annexes. As for user awareness and communication, this regulation mandates user information and transparency measures for high-risk AI systems, ensuring users know the AI's role in decision-making processes:

> Providers shall ensure that AI systems intended to interact directly with natural persons are designed and developed in such a way that the natural persons concerned are informed that they are interacting with an AI system. (European Parliament, 2024, p. 256)

> Deployers of an AI system that generates or manipulates text which is published with the purpose of informing the public on matters of public interest shall disclose that the text has been artificially generated or manipulated. (European Parliament, 2024, p. 258)

Regarding documentation and access to Information, it requires technical documentation and record-keeping for high-risk AI systems:

> Providers shall have a choice of integrating, as appropriate, the necessary testing and reporting processes, information and documentation they provide with regard to their product into documentation and procedures that already exist and are required under the Union harmonisation legislation listed in Section A of Annex I. (European Parliament, 2024, p. 185)

Lastly, for ethical and legal compliance, it requires providers of AI systems to ensure that their systems are transparent to users, including providing information on the purpose and intended use of the AI system and the logic, significance, and potential impact of the AI system's decisions:

> Providers of high-risk AI systems shall put a quality management system in place that ensures compliance with this Regulation. That system shall be documented in a systematic and orderly manner in the form of written policies, procedures and instructions. (European Parliament, 2024, p. 202)

Based on the text excerpts provided, Table 4 summarises key transparency principles, accountability measures, and implementation challenges of the analysed documents.

Table 4. Distribution of transparency frames across AI frameworks.

Document	Key Transparency Principles	Accountability Measures	Implementation Challenges
AIDP	Traceability and accountability	Regulatory oversight	Enforcement, cultural differences
RCAI	Understandability and explainability to foster clarity	Audits, reporting	Complexity, technical limits
BAIBR	Understandability and explainability, accountability	Legal obligations	Inter-agency coordination
AI Act	Understandability and explainability, user awareness and communication, documentation and access to information as basis for risk assessments and safety	Compliance checks	Harmonisation across EU

Table 5 presents information related to RQ2 and RQ3. It identifies the analysed documents and provides excerpts illustrating the codes of objects of AI transparency and types of accountability (Ananny & Crawford, 2018). Transparency is categorised into two main types: event transparency and process transparency. Event transparency focuses on disclosing specific data points, results, or impacts of a system's operations. Process transparency aims to clarify the underlying logic, steps, and governance frameworks that determine how a system functions. On the one hand, "soft accountability" refers to voluntary or normative mechanisms based on recommendations, codes of conduct, or non-binding guidelines. These mechanisms encourage the adoption of responsible practices but rely on the voluntary adherence of those involved. Examples include internal audits, public reports, and organisational ethical commitments (Ananny & Crawford, 2018). On the other hand, "hard accountability" involves formal and binding mechanisms, such as legal sanctions, financial compensation, or regulatory obligations. These mechanisms require compliance and can impose penalties on organisations or individuals who violate established norms (Ferrari et al., 2023).

Table 5. Objects of AI transparency and kind of accountability.

Document	RQ2 (Event/Process)	RQ3 (Hard/Soft Accountability)
AIDP	The document does not explicitly mention transparency as a focus area.	The plan does not clearly define specific responsibilities.
RCAI	Event and Process "AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: i. to *foster a general understanding of AI systems*, … iii. *to enable those affected by an AI system to understand the outcome*, and, iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the *logic that served as the basis for the prediction, recommendation or decision*" (OECD, 2019, p. 8, emphasis by authors).	Soft "AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour right" (OECD, 2019, p. 7).

Table 5. (Cont.) Objects of AI transparency and kind of accountability.

Document	RQ2 (Event/Process)	RQ3 (Hard/Soft Accountability)
BAIBR	Event "Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the *overall system functioning* and the role automation plays, notice that such systems are in use, the individual or organisation responsible for the system, and explanations of *outcomes* that are clear, timely, and accessible" (White House Office of Science and Technology Policy, 2022, p. 6, emphasis by authors).	Soft "Responsibility should rest high enough in the organisation that decisions about resources, mitigation, incident response, and potential rollback can be made promptly, with sufficient weight given to risk mitigation objectives against competing concerns. Those holding this responsibility should be made aware of any use cases with the potential for meaningful impact on people's rights, opportunities, or access as determined based on risk identification procedures" (White House Office of Science and Technology Policy, 2022, p. 19).
AI Act	Event Article 13—Transparency and provision of information to deployers (European Parliament, 2024, pp. 194–196).	Soft "Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant technical standards" (European Parliament, 2024, p. 257). Hard Article 50—Transparency obligations for providers and deployers of certain AI systems (European Parliament, 2024, pp. 256–259).

Finally, the identification of those responsible for ensuring transparency in the use of AI in producing and disseminating information (RQ4) is not addressed in a consistent manner in all documents (Figure 1).

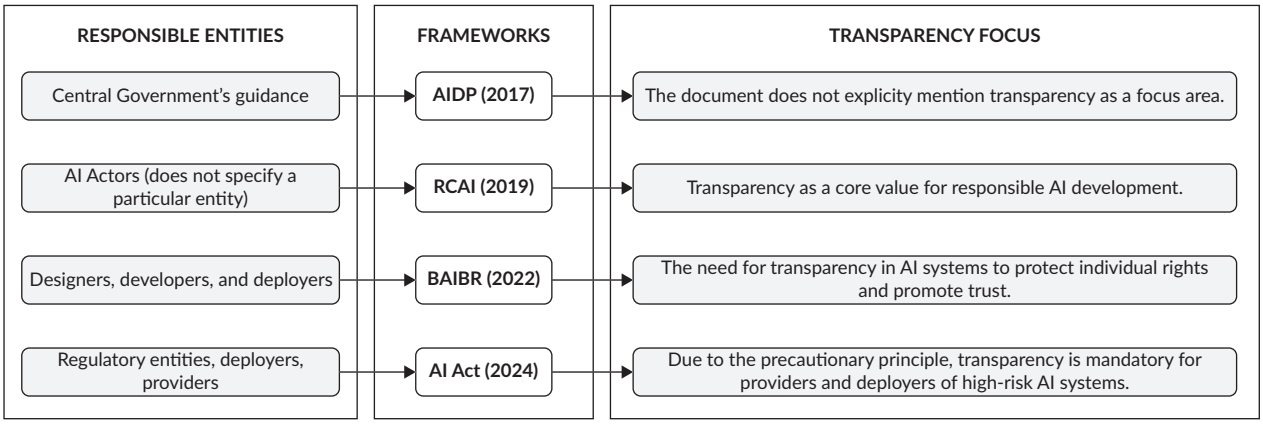


Figure 1. Responsible entities and transparency focus by AI framework.

The RCAI emphasises transparency and responsible disclosure around AI systems. However, it does not specify a particular entity responsible for ensuring this transparency. The same happens with the AIDP, but as a state plan, the implied responsibility likely falls on government entities overseeing AI development and deployment.

The BAIBR asserts the right to transparency in AI systems and calls for clear, understandable explanations. It suggests this is a shared responsibility of those designing, developing, and deploying AI systems. As a US government document, federal agencies are impliedly responsible for upholding these principles.

The AI Act is precautionary and places transparency obligations on providers and deployers of high-risk AI systems. Providers must ensure their systems are transparent and give clear information to users. Deployers have obligations related to monitoring, reporting, and facilitating oversight. Member states, through their national supervisory authorities, are responsible for enforcing these transparency requirements.

5. Discussion

Governments worldwide have begun to develop regulations to govern the use of AI. These regulations aim to ensure that AI is used ethically and responsibly and does not infringe on people's rights. Ethical challenges and principles are at the forefront of ongoing discussions about the governance and regulation of AI, advocating for a multidisciplinary, systemic, proactive, and anticipatory approach to policymaking (e.g., Corrêa et al., 2023; Jobin et al., 2019; Larsson & Heintz, 2020). There is no single approach to AI transparency that suits all contexts. Various documents highlight different aspects, including understandability, accountability, ethical compliance, and risk management.

Notwithstanding AI's potential, if it is used without a moral agenda, it can have harmful consequences (e.g., Bostrom & Yudkowsky, 2014). It seems, however, that ethical frameworks and a self-regulating moral agenda are not enough to contain the social and political impacts of AI (Suleyman & Bhaskar, 2023). For example, aside from efforts to regulate the use of AI and define principles for the governance of AI, China's approach to these issues may differ from Western perspectives due to its unique socio-political context and governance. Roberts et al. (2021) suggest that the Chinese government's interests might outweigh ethical considerations, particularly in surveillance and social governance. This tension could shape the global development and implementation of ethical norms due to state-centric governance models that hinder more decentralised, multi-stakeholder approaches.

The analysis of ethical frameworks and regulations involving AI highlights a greater focus, on one hand, on the events, i.e., on disclosing specific data points, results, or impacts of an AI system's operations (inputs, outputs, and outcomes) rather than in the system's functioning logic (RQ2). On the other hand, it highlights a greater focus on soft accountability with recommendations and prescriptions for agents/players (RQ3).

Hard accountability mechanisms lack in major AI policy frameworks from the OECD, China, and the US. These initiatives seem to fall more on the "soft accountability" end of the spectrum (Ananny & Crawford, 2018). These frameworks aim to bring transparency to AI development in the hope that it will pressure organisations to behave responsibly and be able to justify their actions. However, they do not include "hard accountability" measures that allow for formal sanctions or compensation when violations occur. They depend more on self-regulation and public pressure to encourage adherence.

Though transparency and explainability have become a dominant topic of concern for AI systems since 2018 (Corrêa et al., 2023), we agree with Ferrari et al. (2023) regarding the lack of clarity in AI transparency obligations. There are currently no specific technical details about how such modifications can be enforced in policy practice. The three structural conditions for effective AI governance (industrial observability, public inspectability, and technical modifiability) are also missing, compromising effective oversight and regulation.

This is an important limitation in the accountability paradigms. Ananny and Crawford's (2018) argument implies that achieving meaningful accountability likely requires going beyond transparency alone to include "harder" mechanisms with teeth.

There are also limitations stemming from the lack of clear responsibility for producing and distributing information about the AI system and its events (RQ4). While the understandability and explainability and documentation and access to information frames (RQ1) are acknowledged, they do not specify the conditions for producing and accessing information. Even though the RCAI, the AIDP, and the BAIBR underscore the significance of AI transparency, none of them clearly assign responsibility for transparency in AI-enabled information production and dissemination to specific actors (RQ4). Only the AI Act places transparency obligations on providers and deployers of high-risk AI systems, revealing a stronger emphasis on legal compliance compared to BAIBR, reflecting differences in regulatory approaches.

An integrated approach that combines elements from multiple frames may provide a more comprehensive solution to AI transparency. Efforts to have an integrated approach should be present in policymaking. Policymakers should take into account multiple frames to address the multifaceted nature of AI transparency, such as ensuring AI systems are understandable, holding developers accountable, adhering to ethical standards, and managing risks effectively (Bostrom & Yudkowsky, 2014; Jobin et al., 2019), since an effective AI governance requires a special balance between regulation and flexibility to technological advancements.

An international framework could be a potential solution to address the current limitations of existing frameworks, which primarily focus on "soft accountability" without enforcement mechanisms. There is a lack of clear technical specifications for implementation, an absence of structural conditions for effective oversight, and an inconsistent assignment of responsibility for transparency obligations.

The proposed framework needs to consider the different approaches among regions, as minimising governance models can impede multi-stakeholder participation. Its governing body could be composed of multi-stakeholders representing national governments, the tech industry, academic institutions, civil society organisations, and international standards bodies. Key components of this framework must include mandatory technical standards for AI transparency, clear accountability mechanisms with enforcement powers, dispute resolution procedures, and regular review and update processes. Ethical considerations and a human-centric approach should take precedence over commercial interests.

However, implementing such a framework presents significant challenges. Geopolitical tensions and competing national interests may impede international cooperation, while enforcement across jurisdictions requires complex diplomatic and legal mechanisms. Moreover, the framework must balance the protection of intellectual property rights with transparency requirements, particularly as AI technologies rapidly evolve and market dynamics shift.

AI functions as a sociotechnical system; its context of data creation and interpretation is shaped by humans, and the culture surrounding AI technologies is fundamentally human (Airoldi, 2022). This sociotechnical perspective underscores why governance frameworks must extend beyond technical specifications to encompass social, cultural, and ethical dimensions. Therefore, concerns about AI transparency are intrinsically linked to human values and social ethics, necessitating a governance approach that recognises both the technical and social complexities of AI systems.

6. Conclusion

Choosing a qualitative frame analysis, we were able to present a comprehensive and systematic approach to examine the complexity of AI transparency's multifaceted nature regarding its ethics, policymaking, and governance. The various documents discussing ethical principles for AI have different scopes and priorities. For example, China's AIDP centres on economic competitiveness, while the other documents focus on fundamental rights; the AI Act provides detailed regulations, while the US Blueprint focuses more on high-level principles. This lack of alignment could limit transparency and imply vested economic and political interests.

Some frameworks, such as the RCAI and BAIBR, are non-binding, potentially limiting their impact on driving transparent practices compared to the enforceable regulations in the AI Act. Despite these differences, the fundamental frames highlighted in the analysed AI ethical guidelines and regulations reveal some of the conceptual, normative, and practical dimensions of transparency.

Conceptually, transparency is a core ethical principle for trustworthy AI. It enables explainability and understanding of how AI systems make decisions. The documents also highlight the need for transparency at various stages of AI development and deployment, not just in the final outputs, stressing the need for an overall understanding of the AI system's functioning and logic beyond the prediction, recommendation, or decision.

Normatively, transparency is a fundamental right for individuals impacted by AI systems. Ensuring AI systems and their developers can be held accountable is critical to fostering public trust. Except for the AIDP, the documents also link transparency to normative principles of fairness and equity and prevent discriminatory impacts of AI.

However, in practical terms, the documents lack specificity on transparency requirements. While transparency is encouraged as an ethical principle, there is limited concrete guidance on what transparency entails in practice for different AI applications. Even with ethical principles established, ensuring meaningful transparency will require robust implementation, oversight, and enforcement mechanisms, which may face practical hurdles.

In summary, differences in priorities, legal obligations, specificity, and implementation across the world can hinder the consistent achievement of AI transparency goals without further alignment and strengthening of approaches. Developing a coherent, flexible, dynamic, and context-aware ethical international framework may help keep AI technology on a responsible path. Continuous learning, collaboration, and adaptation will be crucial.

This study's findings point to several critical areas that warrant further investigation in the field of AI transparency and governance. Research should examine how organisations operationalise transparency requirements across different AI applications and contexts, focusing on successful implementation strategies and practical challenges.

Scholarly attention should focus on establishing clear lines of responsibility and accountability in AI development and deployment, particularly in complex multi-stakeholder environments where responsibilities span multiple actors and jurisdictions. The field would also benefit from research exploring how different cultural, social, and political contexts influence transparency expectations, moving beyond Western-centric approaches towards more culturally sensitive and globally applicable governance frameworks.

Further empirical research is needed to assess the effectiveness and impact of AI ethical guidelines and regulations. Such studies should evaluate how different regulatory approaches influence organisational behaviour, innovation processes, and compliance mechanisms. This includes examining the implementation challenges of binding versus non-binding frameworks and their relative success in promoting transparent and responsible AI practices.

Longitudinal studies evaluating the effectiveness of transparency mechanisms in promoting responsible AI development and maintaining public trust are essential to provide empirical evidence of successful approaches across different contexts. Additionally, research should examine how individual and organisational choices influence transparency outcomes, acknowledging that human decision-making remains central to AI development and investigating how organisational culture and institutional frameworks shape transparency practices.

This article's scope is limited due to the abundance of AI ethical guidelines issued by research institutes, companies, and NGOs. However, the study includes input from various regions to reduce Western bias. The findings underscore that responsible AI fundamentally depends on responsible human actors, as humans create the technology, program the applications, select the information, and determine its use.

The path forward requires a delicate balance between establishing robust transparency frameworks and maintaining flexibility for technological advancement. Future success in AI governance will depend on continuous learning, international collaboration, and adaptive approaches that recognise both the technical and human dimensions of AI systems.

Acknowledgments

This article used AI to translate original Portuguese sentences into English using DeepL. AI was also employed in specific aspects, namely language clarity and checking using Grammarly Business. The authors double-checked all AI-assisted texts to correct selected wordings, grammatical adjustments, and paraphrases. The authors, safeguarding human oversight and expertise in the final output, performed the overall composition of the article manually. The final manuscript was proofread by a proficient/native English speaker.

Funding

This work is supported by Portuguese national funds through FCT—Fundação para a Ciência e a Tecnologia, under project UIDB/00713/2020.

Conflict of Interests

The authors declare no conflict of interests.

References

Airoldi, M. (2022). *Machine habitus: Toward a sociology of algorithms*. Polity.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316–334). Cambridge University Press.

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.

Buiten, M. (2019). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation*, *10*(1), 41–59. https://doi.org/10.1017/err.2019.8

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'good society': The US, EU, and UK approach. *Science and Engineering Ethics*, *24*(2), 505–528. http://doi.org/10.1007/s11948-017-9901-7

China's State Council. (2017). *A Next Generation Artificial Intelligence Development Plan*. Stanford Cyber Policy Center. https://d1y8sb8igg2f8e.cloudfront.net/documents/translation-fulltext-8.1.17.pdf

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Barbosa, C., Massmann, D., Manbrini, R., Galvão, L., Terem, E., & de Oliveira, N. (2023), *Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance*. arXiv. https://doi.org/10.48550/arXiv.2206.11922

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

European Parliament. (2024). *Artificial Intelligence Act*. https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf

Ferrari, F., van Dijck, J., & van den Bosch, A. (2023). Observe, inspect, modify: Three conditions for generative AI governance. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448231214811

Frankish, K., & Ramsey, W. M. (2014). Introduction. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 1–11). Cambridge University Press.

Gil de Zúñiga, H., Goyanes, M., & Durotoye, T. (2023). A scholarly definition of artificial intelligence (AI): Advancing AI as a conceptual framework in communication research. *Political Communication*, *41*(2), 317–334. https://doi.org/10.1080/10584609.2023.2290497

Hood, C. (2006). Transparency in historical perspective. In C. Hood & D. Heald (Eds.), *Transparency: The key to better governance?* (pp. 2–23). Oxford University Press. https://doi.org/10.5871/bacad/9780197263839.003.0001

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kaplan, A., & Haenlein, M. (2019). Siri, Siri in my hand, who is the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, *63*(1), 37–50. https://doi.org/10.1016/j.bushor.2019.09.003

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, *22*(14), 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society*, *7*(3), 437–451. https://doi.org/10.1017/als.2020.19

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2), https://doi.org/10.14763/2020.2.1469

Lee, K.-F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. HarperCollins Publishers.

Lewins, A., & Silver, C. (2007). *Using software in qualitative research: A step-by-step guide*. Sage.

Liao, S. M. (2020). A short introduction to the ethics of artificial intelligence. In S. M. Liao (Ed.), *The ethics of artificial intelligence* (pp. 1–42). Oxford University Press.

Lock, I., Wonneberger, A., Verhoeven, P., & Hellsten, I. (2020). Back to the roots? The applications of communication science theories in strategic communication research. *International Journal of Strategic Communication*, *14*(1), 1–24. https://doi.org/10.1080/1553118X.2019.1666398

Lungarella, M., Iida, F., Bongard, J. C., & Pfeifer, R. (2007). AI in the 21st century—With historical reflections. In M. Lungarella, F. Iida, J. C. Bongard, & R. Pfeifer (Eds.), *50 years of artificial intelligence. Essays dedicated to the 50th anniversary of artificial intelligence* (pp. 1–8). Springer.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society*, *36*, 59–77. https://doi.org/10.1007/s00146-020-00992-2

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., Patel, A., Ryan, M., & Wright, D. (2021). Governing the ethics of artificial intelligence for human flourishing—Beyond principles for machine learning. *Journal of Business Research*, *124*, 374–388. https://doi.org/10.1016/j.jbusres.2020.11.030

Suleyman, M., & Bhaskar, M. (2023). *The coming wave: Technology, power, and the twenty-first century's greatest dilemma*. Crown.

Sullivan, Y. W., & Wamba, S. F. (2022). Moral judgments in the age of artificial intelligence. *Journal of Business Ethics*, *178*, 917–943. https://doi.org/10.1007/s10551-022-05053-w

White House Office of Science and Technology Policy. (2022). *The Blueprint for an AI Bill of Rights: Making automated systems work for the American people*. https://www.whitehouse.gov/ostp/ai-bill-of-rights

Woolf, N. H., & Silver, C. (2018). *Qualitative analysis using MAXQDA. The five-level QDA method*. Routledge.

About the Authors

Sónia Pedro Sebastião, PhD in social sciences in the speciality of communication sciences from the Instituto Superior de Ciências Sociais e Políticas, Universidade de Lisboa (Portugal), is a full professor in communication sciences and the director of the Center of Public Administration and Public Policies (CAPP, ISCSP/FCT). Her research interests focus on strategic communication, ethics, citizenship, and cultural studies.

David Ferreira-Mendes Dias, PhD in social sciences in the speciality of communication sciences from the Instituto Superior de Ciências Sociais e Políticas, Universidade de Lisboa (Portugal), is an invited auxiliary professor in communication sciences and a TV professional at RTP (Portuguese public service broadcaster). His research interests focus on digitalization, social media, platformization, and cultural studies.

Media and Communication
2025 • Volume 13 • Article 9550
https://doi.org/10.17645/mac.9550

cogitatio

ARTICLE

Open Access Journal

Public Segmentation and the Impact of AI Use in E-Rulemaking

Loarre Andreu Perez [1], Matthew L. Jensen [2,3], Elena Bessarabova [3,4], Neil Talbert [3,4], Yifu Li [5], and Rui Zhu [5]

[1] Journalism and Media Studies, San Diego State University, USA
[2] Management Information Systems Division, University of Oklahoma, USA
[3] Center for Applied Social Research, University of Oklahoma, USA
[4] Department of Communication, University of Oklahoma, USA
[5] School of Industrial and Systems Engineering, University of Oklahoma, USA

Correspondence: Loarre Andreu Perez (landreuperez@sdsu.edu)

Submitted: 31 October 2024 **Accepted:** 6 March 2025 **Published:** 12 June 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

Digitization has profoundly changed how government interacts with its publics. The expanding use of AI promises even more advancement. However, the rollout of AI is not without risk. This work explores the use of AI in federal rulemaking, the process by which regulations are introduced and revised. The US federal government has created digital platforms that dramatically expand access for the public commenting on pending regulations. However, these platforms also attract volumes of opinion spam that attempt to influence regulatory decision-making. Using AI to identify opinion spam may offer a potential remedy, but removing or limiting comments with the help of AI may threaten rulemaking legitimacy. This research uses the situational theory of problem-solving as a framework, segmenting publics based on their problem recognition, constraints, and involvement with a specific issue, then predicting how each public behaves. We examined how employing AI in the processing of rulemaking comments affects public segments' intention to comment, their perceptions of legitimacy of the resulting rules, trust in agencies, and control mutuality between the public and the agency. This work describes two controlled, randomized experiments ($N = 149$; $N = 250$) that capture public segments' reactions to AI use in analyzing comments in the presence or absence of opinion spam. We show that public segmentation is a key aspect in shaping attitudes and behaviors regarding the use of AI for e-rulemaking purposes. These findings suggest that communicating effectively with publics is essential for agencies, and that the use of AI does not make the publics' attitudes differ.

Keywords

AI; commenting behavior; comment filtering; content moderation; electronic rulemaking; notice-and-comment; opinion spam

1. Introduction

Electronic rulemaking, or e-rulemaking, is a participatory process that encompasses the use of information technology to facilitate citizens' input on proposed regulations (Department of Defense Open Government, n.d.). Established in the US in 2002 (Regulations.gov, n.d.), e-rulemaking expands the possibilities of citizen participation, providing access to more users and encouraging more people to engage. However, increased accessibility also resulted in adverse consequences such as mass commenting (Shulman, 2009), when a large number of identical or nearly identical comments are posted on regulation sites (Balla et al., 2021).

Activist groups or corporate interests often organize these mass commenting campaigns, aiming to generate support for their causes. However, these initiatives often make it difficult for federal agencies to identify substantive contributions (Farina et al., 2012). Campaigns from activist groups are likely to include comments, expressing support or opposition without offering substantive contributions on the issue at hand; comments submitted by such campaigns are known as opinion spamming (Liu, 2012). Beyond making American federal agencies overwhelmed by the number of comments (Farina et al., 2012), mass commenting and opinion spamming can obstruct efforts to achieve deliberative democracy by overwhelming the plurality of citizens' voices (J.-N. Kim et al., 2025; Shulman, 2009).

Since the digitalization of the rulemaking process, there have been efforts to implement more cutting-edge technologies (Park et al., 2015). The use of AI in the context of e-rulemaking can help group repetitive comments, highlight their distinct contributions, and categorize information within comments (Eidelman & Grom, 2019) without compromising the time and efforts of an agency's staff. In this context, AI has been used to classify comments, detect duplicates, and highlight keywords for the proposed regulation. AI is a possible response to the opinion spamming problem, offering a remedy to the comment overload and tediousness of dealing with these repetitive comments. While AI may directly solve the issue at hand, it can also bring several related problems (Li et al., 2020). First, the utilization of AI tools raises questions regarding ethics; second, it could impact public participation depending on citizens' response to the use of these advanced tools; and third, comment removal may threaten rulemaking legitimacy.

In light of the aforementioned concerns, this article has two objectives. The first is to explore the reactions of citizens when they observe opinion spamming in e-rulemaking. Previous work has focused on the impact opinion spamming has on American federal agencies' staff (Farina et al., 2012; J.-N. Kim et al., 2025) and the effects on the e-rulemaking process as a whole (Shulman, 2009), but citizens' behaviors as a result of opinion spamming are yet to be explored. Therefore, it is essential to know the effects of opinion spamming on publics: its influence on their willingness to participate in the e-rulemaking process as well as their perceptions of agencies and resulting regulations.

Although opinion spam detection and filtering technologies powered by AI may be a potential solution to opinion spamming, the implementation of AI is not free of risks (Li et al., 2020). Thus, the second objective of this research relates to understanding citizens' perceptions of and behaviors related to the application of AI to the e-rulemaking process. Before using AI to combat opinion spamming, assessing how publics perceive and respond to this technology is essential.

These ideas were tested in two experiments. The first study focused on citizens' reactions to the problem of opinion spamming depending on their segmentation type, a classification rooted in the situational theory of problem-solving (STOPS; J.-N. Kim & Grunig, 2011) that segments publics depending on their involvement in specific issues. The second study examined the perceptions of AI as a possible solution to the problem, testing not only publics' attitudes and behavioral intentions about opinion spamming but also their response to the agency using AI to filter comments.

The present research has implications for both theory-building and practice. This article contributes to the growing body of research on e-rulemaking, applying public relations theory to this specific context. This approach is necessary for exploring publics' attitudes and behaviors surrounding the use of AI and its impact on resulting regulations and federal agencies, thereby examining the social component of e-rulemaking and the acceptance of new technologies. In addition, this article provides guidance on what agencies should do and what their stance on technologies should be when managing the e-rulemaking process.

2. Literature Review

2.1. E-Rulemaking Evolution

While e-rulemaking was established in 2002 (Regulations.gov, n.d.), publics' participation in the rulemaking process has a long history. It was in 1946 when publics were first able to comment on proposed regulations, based on the Administrative Procedure Act (Moxley, 2016). The first round of changes to the commenting process took place during the 1990s, when agencies proactively started using online tools to collect citizens' comments (Benjamin, 2006). The next round of changes in the e-rulemaking process took place in 2002, with agencies posting proposed rules and enabling comments on a centralized website, Regulations.gov (Regulations.gov, n.d.). As technologies evolved, the system also transitioned into e-rulemaking. Agencies post regulatory materials online so that they are publicly available. In the same portal, Regulations.gov, publics can share their voices by commenting on proposed regulations as well as read other participants' comments.

Different administrations aimed to enable publics to freely comment and access the materials, including President Bush's Honest Leadership and Open Government Act of 2007, and President Obama's Memorandum on Transparency and Open Government in 2009 (Farina et al., 2011). Centralization was made mandatory for all agencies via Regulations.gov, which connects to a database that enables document management, maintains digital versions of rulemaking documents, and provides search mechanisms (Moxley, 2016). This switch to online procedures was motivated by the need for accessibility, publics' participation, openness, and transparency (Perez, 2020).

2.2. Opinion Spamming in E-Rulemaking

The use of online platforms for rulemaking purposes arose from the need for accessibility and an increase in public participation (Benjamin, 2006; Perez, 2020). Citizen participation increased because of the openness of the e-rulemaking process, yet unfortunately, agencies have struggled to find substantive feedback among the vast volume of comments they now receive (Farina et al., 2012). The abundance of comments creates challenges for agencies' staff, who, overwhelmed by volume, may struggle to synthesize all the content provided by citizens (Farina et al., 2012).

As noted, mass commenting refers to large quantities of nearly identical comments on regulations coordinated by corporate interests or activist organizations, who create the comment content and enable mass sharing among their followers (Balla et al., 2021). Within mass commenting, we refer to cases where the sources or underlying intentions of comments are obscured as opinion spamming (Liu, 2012).

Research has examined the detrimental effects of mass commenting and opinion spamming, including difficulties for agencies to manage their work regarding the rulemaking process (Farina et al., 2012; Perez, 2020), impacts on deliberative democracy (Shulman, 2009; Widyatama & Mahbob, 2024), and also the possibility of discouraging feedback from citizens (Benjamin, 2006; Grossman, 2004). Grossman (2004) explained how the presence of opinion spamming can be off-putting for other users, who criticized the abundance of spamming and can't escape or opt out of spamming; and Benjamin (2006) found that after observing opinion spamming, publics became less engaged in providing feedback. Taken together, these findings suggest that opinion spamming is a serious issue both for regulatory agencies and for deliberative democracy.

2.3. Public Segmentation in E-Rulemaking

In order to understand the extent to which public behaviors are a reaction to the issue of opinion spamming, it is worth considering the nature of these publics per se. Opinion spamming refers to a response orchestrated by an organization (Liu, 2012); in the case of e-rulemaking, the organization advocating for its interests behind the scenes may be an activist group (Balla et al., 2021) reacting to proposed regulations related to their mission. As commenting effectively requires a degree of regulatory savvy and issue-relevant knowledge, to boost support and encourage less engaged individuals to comment, activist groups provide their publics with form letters to make commenting easier, which require that publics only sign, send, and, optionally, make their own edits (Schlosberg et al., 2009). Thus, activist groups generate the statement, disseminate the information, and even enable automatic posting settings from their websites to make the mass posting of comments easier.

There are many publics associated with specific issues who communicate with each other. The discipline and theory of public relations focuses on the study of publics' nature, their attitudes, behaviors, relationships, and classification (Hallahan, 2018; J.-N. Kim et al., 2008; J.-N. Kim & Grunig, 2011). Specifically, STOPS helps identify different publics and predict the communication behaviors of each public segment (J.-N. Kim & Grunig, 2011). Grounded in publics and public opinion concepts, this theory explains how to segment publics depending on their perceived problem recognition, involvement recognition, and constraint recognition (J.-N. Kim & Grunig, 2011). Problem recognition refers to the state in which a problem is a product of experiences and expectations, arising from discrepancies between experiential and expectation states (J.-N. Kim & Grunig, 2011). Involvement recognition refers to the connection between oneself, the environment, and the problem (J.-N. Kim & Grunig, 2011). Constraint recognition assesses both the internal and external barriers that limit one's actions and efforts to do something about the problem (J.-N. Kim & Grunig, 2011).

The variables utilized for classifying publics respond to publics' perceptions of themselves and a pre-existing issue or problem, meaning this segmentation technique has been applied to different issues and problems. Empirical research supports the theory's propositions (Chon & Park, 2021; Chon et al., 2023; H. J. Kim & Hong, 2022; J.-N. Kim & Krishna, 2014).

Depending on motivations and self-perceptions associated with the problem (i.e., problem, involvement, and constraint recognitions), publics can be segmented into non-publics, latent, aware, active, and activist publics (J.-N. Kim & Grunig, 2011). Non-publics are those who are not connected to the issue (J.-N. Kim et al., 2008). Similarly, latent publics have low awareness about the issue and lack concern about it (J.-N. Kim et al., 2008). Because of their lack of engagement and involvement with the problem, researchers often refer to these two groups as passive publics, combining these two groups into one category (J. E. Grunig & Kim, 2017). Aware publics recognize the existence of the problem and feel more connected or impacted by the issue (J.-N. Kim et al., 2008). Active publics, like aware publics, recognize the problem, but they go one step beyond in their degree of organization, willingness to discuss the problem, and do something about it (J.-N. Kim et al., 2008).

Citizens who participate in the e-rulemaking process by commenting have greater displays of motivation, making them active and activist publics J.-N. (J.-N. Kim & Grunig, 2011). Active publics engage in collective solutions to the specific issue they are active about (J. E. Grunig & Kim, 2017; J.-N. Kim et al., 2010). Activists are motivated to produce change and know the repercussions that decisions have for them, so they organize themselves and generate issues out of the consequences of an organization's (institution or corporation) decisions, having the ability to address the problem (J. E. Grunig & Kim, 2017).

As commenters are members of highly motivated publics, it becomes vital to understand the impact that the presence of opinion spamming has on each segment of the public. One potential effect of opinion spam is its impact on citizen's behavioral intention to comment, as the presence of opinion spam tends to discourage participation (Grossman, 2004).

Publics are segmented depending on their issue perceptions. These are not static groups: Differences in perceived constraints can "deactivate" publics, making active publics become aware publics as the number of perceived constraints increases (J.-N. Kim et al., 2008). As mass commenting discourages participation (Benjamin, 2006; Grossman, 2004), citizens who are more likely to comment and participate will perceive opinion spamming as a barrier and may potentially disengage:

> H1: The more active publics are, the lower their intention to use Regulations.gov will be when exposed to opinion spamming.

As noted, the process of e-rulemaking is closely connected to democracy and the legitimacy of the process, agencies, and resulting regulations (Benjamin, 2006; Perez, 2020; Shapiro, 2019). Flaws in the rulemaking process affect citizens' perceptions of legitimacy, with opinion spamming being one of the most prevalent issues undermining rule legitimacy (Rinfret et al., 2022). Given that active publics may feel discouraged by opinion spam and deactivate, they are also more likely to see opinion spam as damaging to the legitimacy of resulting regulations. Based on this reasoning, we hypothesize:

> H2: The more active publics are, the lower regulation legitimacy they perceive when exposed to opinion spamming.

Publics, especially active publics, are the target of communication from various organizations, as publics' management is indispensable for organizations (e.g., corporations, institutions; L. A. Grunig et al., 2002). Organizational goals can only be achieved when the organization is engaged in relationship-building and

communication with publics. In the e-rulemaking context, citizens involved with proposed regulations are publics, and the organizations citizens develop relationships with are the agencies. In public relations, relationships between publics and different organizations have been studied using the organization-public relationship assessment scale (J. E. Grunig & Huang, 2000; Huang, 2001), which include the notions of control mutuality and trust, described as key factors to successful communication between organizations and the publics (Hon & Grunig, 1999; Huang, 2001).

Control mutuality is the extent to which publics and an organization permit their influence on each other to determine goals and behavioral routines (Huang, 2001). Control mutuality is critical for interdependence and relational stability. As opinion spamming directly impacts publics, their levels of control mutuality with the agency posting and reviewing the regulations may differ if they perceive large amounts of duplicated comments, as opinion spamming can be considered a constraint with the power of reducing publics' levels of activity (J.-N. Kim et al., 2008). STOPS explains that communication behaviors vary depending on levels of activity (J.-N. Kim & Grunig, 2011), as their situational motivation in problem-solving varies in function to their perceptions and cognitive evaluation of the problem. Communication is, at the same time, the germ of control mutuality in relationships between publics and organizations (Huang, 2001). When examining relationships with the agency, control mutuality (thus, power bargaining and perceived right to influence) is a factor that assesses relationship quality (Hon & Grunig, 1999; Huang, 2001) and is directly influenced by communication actions:

> H3: The more active publics are, the lower control mutuality they experience when exposed to opinion spamming.

Trust is the level of confidence and willingness to open oneself to another party (Hon & Grunig, 1999). Trust is an indication of relationship quality between publics and agencies. Similar to control mutuality, trust in the agency is at risk when citizens observe the problem of mass commenting on a proposed regulation. In this case, publics will be less likely to perform communicative actions, and the lack of communication diminishes the trust publics feel towards organizations (Huang, 2001; here, federal agencies). Previous research about opinion spamming in online spheres has indicated that publics' trust diminishes with the presence of opinion spamming (Gupta & Bala, 2024). Bringing these findings into the e-rulemaking context results in the following prediction:

> H4: The more active publics are, the lower the trust they experience when exposed to opinion spamming.

2.4. AI in E-Rulemaking

In addition to understanding the nature of publics in relation to opinion spamming exposure, it is also necessary to explore how possible solutions to opinion spamming could affect publics' attitudes and behaviors about e-rulemaking.

AI and machine learning capabilities have piqued the interest of policymakers, who see model mapping and predictability as features to implement in their work (Strandburg, 2019). Some AI features in e-rulemaking could include categorizing and generating more objective answers to each policy or regulation (Eidelman &

Grom, 2019; Strandburg, 2019), as well as filtering repetitive comments and highlighting relevant portions within comments (Eidelman & Grom, 2019).

Opinion spamming produces high numbers of duplicated comments, with little differences from one another, which makes manual filtering and classification tedious and slow. AI capabilities offer a remedy to comment overload. However, specific AI tools, which may work in certain instances, may not fit the full array of contexts in which they can be applied in rulemaking (i.e., these could erase nuances regarding proposed regulation value balance, embed biases in the system producing discriminatory errors, and introduce incorrect interpretations of the regulation, challenging constitutional democracy; Rangone, 2023), evoking mixed responses from citizens.

In addition to examining publics' responses to mass commenting, we also assessed how people view the agencies' use of different comment-management techniques such as when (a) the agency's staff manually reviews comments, (b) AI is being used to manage comments, or (c) a hybrid option, with humans reviewing the comments in addition to the use of AI. Examining the responses to these three approaches designed to mitigate opinion spam will shed light on people's perception of how the use of these techniques affects behavioral intention to comment, legitimacy of the resulting rule, control mutuality with the agency, and trust in the agency.

Users' acceptance of AI determines how much they will be able to successfully adopt newer technologies utilizing AI (Kelly et al., 2023). Exposure to duplicated mass comments can affect perceptions of the usefulness of AI. In addition, because of their levels of activity and engagement, each public segment tends to react differently to issues and problems. Since their perceptions of the issue along with constraints and solutions may differ, the actions they plan on taking may also vary, producing differences in willingness to comment, perceptions of an agency's work, and the resulting regulation. For that reason, we explore the potential three-way interaction between opinion spamming presence, public segmentation, and the use of AI to filter comments:

> RQ: What is the relationship between opinion spamming, public segmentation, and comment-management techniques on (RQa) behavioral intention to use Regulations.gov, (RQb) legitimacy of the resulting regulation, (RQc) control mutuality with the agency, or (RQd) trust toward the agency?

3. Methods

3.1. Participants

This research included two experiments. In both studies, participants were US citizens recruited using Prolific. Prolific is an online data collection panel that has been shown to yield more complete and meaningful data relative to other online panels, as Prolific participants are more likely to pass attention checks, follow instructions, and are required to have unique IP addresses (Douglas et al., 2023). Those participants who took part in Study 1 were not allowed to participate in Study 2.

Sample sizes were determined based on Cohen's power calculation for medium effect sizes (Bhattacherjee, 2012). A commonly cited guideline suggests a minimum of 20 participants per cell (Bhattacherjee, 2012), these being a minimum of 120 participants in Study 1, and 240 participants for Study 2.

In Study 1 ($N = 149$), 37.6% ($n = 56$) of the participants were male, while 60.4% ($n = 90$) were female. There were 1.3% ($n = 2$) non-binary participants, and one participant who did not disclose their gender. Participants ranged from 18 to 77 years of age ($M = 39.66$, $SD = 13.74$). As for the racial distribution, 61.1% were White ($n = 91$), 18.1% were Black or African American ($n = 27$), 3.4% were Latinx ($n = 5$), 9.4% were Asian ($n = 14$), 1.3% were American Indian or Alaska Native ($n = 2$), 4.7% participants recorded their belonging in the "other" category ($n = 7$), and 2% ($n = 3$) preferred not to disclose their race.

In Study 2 ($N = 250$), 44% ($n = 110$) reported they were male, 54.8% ($n = 137$) were female, .8% ($n = 2$) were non-binary, and one participant did not disclose their gender. Age ranged from 18 to 95 years of age ($M = 39.66$, $SD = 13.41$). In regard to race, 65.2% ($n = 163$) were White, 16.4% ($n = 41$) were Black, 6.4% ($n = 16$) were Latinx, 9.2% ($n = 23$) were Asian, one participant was American Indian or Alaska Native, five self-identified as "other," and one participant did not report their ethnicity.

3.2. Design and Procedure

Study 1 followed a 3 (publics segmentation: passive publics, aware publics, active publics) × 2 (opinion spam: absent, present) factorial design. Study 2 expanded on the first experiment by including the approaches utilized by agencies to deal with opinion spam. The second experiment followed a 3 (publics segmentation: passive publics, aware publics, active publics) × 2 (opinion spam: absent, present) × 3 (comment-management technique: human, AI, mix of human and AI) between-subjects design. These studies aimed to reveal how these experimental conditions affected behavioral intention to comment, perceived legitimacy of the resulting rule, trust towards the agency, and control mutuality with the agency. Both experiments were housed in Qualtrics, with Prolific disseminating the link to the survey to their panel participants.

In both studies, after consent procedures, the first set of questions was designed to capture public segmentation regarding the rights of gun ownership. Gun ownership is a controversial issue that was selected as the topic. Given that public segmentation is done using self-reported views on an issue, a controversial topic is more likely to produce higher numbers of active publics, which are typically difficult to recruit, as it is complicated to find people who are truly involved. The segmentation set of questions utilized for both studies was a reduced version of problem recognition, constraint recognition, and involvement recognition, taken from STOPS (J.-N. Kim & Grunig, 2011). There were three items for each of the utilized STOPS variables for segmentation: problem recognition (Study 1: $M = 4.21$, $SD = 1.08$, $\alpha = .89$; Study 2: $M = 4.21$, $SD = 1.10$, $\alpha = .95$), constraint recognition (Study 1: $M = 2.72$, $SD = 1.14$, $\alpha = .86$; Study 2: $M = 2.67$, $SD = 1.12$, $\alpha = .89$), and involvement recognition (Study 1: $M = 3.33$, $SD = 1$, $\alpha = .70$; Study 2: $M = 3.28$, $SD = 1.10$, $\alpha = .79$). All items were measured with Likert scales ranging from 1 (*strongly disagree*) to 5 (*strongly agree*).

The segmentation method has been applied before in several studies that utilized STOPS, including J.-N. Kim et al. (2011), and is fully explained in Chon et al. (2023). Using midpoint splits, the data from the three situational variables was dichotomized, creating dummy variables, wherein 1 = *high* and 0 = *low*. These

dummy variables were then summed, which resulted in four groups of public segmentation, wherein the score of 0 = non-publics, 1 = latent publics, 2 = aware publics, and 3 = active publics (Chon et al., 2023). As this research examines passive publics, those with a score of 0 (i.e., non-publics) or 1 (i.e., latent publics) were considered passive publics. In Study 1, as a result of segmentation, 49 participants were passive publics (a combination of non-publics and latent publics), 68 were aware publics, and 32 were active publics. In Study 2, 95 participants were passive publics, 104 were aware publics, and 51 were active publics. These groups were obtained from the segmentation method outlined above.

Next, participants were randomly assigned to one of the experimental scenarios. The randomization was performed by Qualtrics, the online platform that administered the survey. All participants initially read the same request for comments regarding the proposed gun control rule from the Bureau of Alcohol, Tobacco, Firearms, and Explosives (ATF). The scenario was a shortened version of an actual proposed regulation posted by the ATF on Regulations.gov. The proposed regulation suggested a modification of the definition of when a person is considered engaged in the business or trade of firearms as a dealer and the paperwork individuals need to complete to transfer firearms to other individuals. Participants were told that citizens could comment on the proposed regulation on Regulations.gov, expressing their thoughts, and that the ATF would review these comments and modify the proposed rule if necessary.

After reading this information, participants were randomly assigned to one of the opinion spam conditions, wherein they either read all unique comments (i.e., opinion-spam-absent condition; $n_{\text{Study 1}} = 62$; $n_{\text{Study 2}} = 124$) or comments that included opinion spam, with repetitive duplicated comments from both pro- and anti-gun control publics (i.e., opinion-spam-present condition; $n_{\text{Study 1}} = 87$; $n_{\text{Study 2}} = 126$). Comments in both opinion-spam conditions were shortened versions of actual comments shared by citizens on Regulations.gov. Afterwards, participants were asked to complete manipulation checks capturing whether they perceived the presence or absence of opinion spamming before continuing with the experiment. Manipulation checks contained one open-ended question and several multiple-choice questions. Participants were not allowed to continue if they failed the manipulation checks and had the chance to re-read the information and comments before completing the manipulation check for a second time. Failing the check implied answering the open-ended question with an answer that did not make sense and answering wrongly the multiple-choice question. In Study 1, three participants did not complete the manipulation check questions satisfactorily and were excluded from the study. The next set of questions measured the four dependent variables: trust in the ATF, control mutuality, behavioral intention to use Regulations.gov, and legitimacy of the proposed rule. The survey concluded with demographic questions.

In Study 2, the main change involved adding the third independent variable, comment-management techniques. Participants read that for reviewing and filtering comments, agencies used human reviewers ($n = 82$), AI ($n = 87$), or a combination of both ($n = 81$). AI was explained as a tool utilized to clean comments submitted on Regulations.gov. These manipulations all included one brief paragraph that explained how the agencies deal with comments, including those considered offensive, duplicated, or irrelevant. Extra manipulation checks were included for this variable, to make sure that participants understood the method of filtering comments. In Study 2, three participants did not complete the manipulation check questions satisfactorily and were removed from the final sample. After the manipulation checks, a set of questions was administered to measure the dependent variables and a set of demographic questions.

3.3. Measurement

All dependent variables were measured on a 1 to 5-point scale, wherein 1 = *strongly disagree* and 5 = *strongly agree*. Four items from Shroff and Keyes (2017) were used to measure behavioral intention to use Regulations.gov (Study 1: $M = 2.78$, $SD = 1.27$, $\alpha = .95$; Study 2: $M = 2.91$, $SD = 1.22$, $\alpha = .95$). Five items were used to capture perceptions of legitimacy (Study 1: $M = 3.38$, $SD = 1.15$, $\alpha = .93$; Study 2: $M = 3.48$, $SD = .99$, $\alpha = .92$). Both trust and control mutuality were taken from Hon and Grunig (1999) and Huang (2001). Trust includes six items (Study 1: $M = 2.93$, $SD = 1.01$, $\alpha = .92$; Study 2: $M = 3.13$, $SD = .98$, $\alpha = .92$). Control mutuality comprised four items (Study 1: $M = 2.95$, $SD = .85$, $\alpha = .71$; Study 2: $M = 3.09$, $SD = .80$, $\alpha = .81$).

Several covariates were measured in both experiments. These were the demographic questions—gender, race, age, education, income, and political ideology—and positions on gun control and referent criterion. Position on gun control was a single-item measure, capturing participants' preference for free gun ownership or limited gun ownership. Referent criterion is the previous experience in deciding or solving a similar problem (J.-N Kim & Grunig, 2011), and it is a variable associated with the STOPS framework, although not utilized for publics segmentation purposes. It was measured using three items (e.g., I know how to deal with issues related to gun control) taken from J.-N. Kim and Grunig (2011; Study 1: $M = 3.19$, $SD = .89$, $\alpha = .70$; Study 2: $M = 3.03$, $SD = .92$, $\alpha = .70$). In Study 2, an extra question regarding attitudes toward AI was also included ($M = 2.70$, $SD = 1.30$).

4. Results

4.1. Study 1

Study hypotheses were tested using MANCOVAs. Omnibus effects indicated statistically significant differences in public segmentation, Wilk's $\Lambda = .83$, $F(8, 264) = 3.14$, $p < .01$, $\eta_p^2 = .12$; but no significant differences for the presence of opinion spamming or the interaction between public segmentation and opinion spamming presence. Gender, race, age, education, income, position on gun control, political ideology, and referent criterion were entered as covariates. Among them, the following covariates were significant: position on gun control, Wilk's $\Lambda = .85$, $F(4, 132) = 3.14$, $p < .001$, $\eta_p^2 = .14$; political ideology, Wilk's $\Lambda = .88$, $F(4, 132) = 4.30$, $p < .01$, $\eta_p^2 = .11$; and referent criterion, Wilk's $\Lambda = .91$, $F(4, 132) = 2.93$, $p < .01$, $\eta_p^2 = .08$.

H1 predicted an interaction between opinion spamming presence and public segmentation, and its influence on behavioral intention to comment such that the more active publics are, the lower their intention to comment they would be when exposed to opinion spamming. While the interaction was not significant, there was a statistically significant main effect of segmentation on behavioral intention to use Regulations.gov, $F(2, 135) = 5.91$, $p < .01$, $\eta_p^2 = .11$. There was no effect of opinion spamming presence on the intention to use Regulations.gov. A Bonferroni test was performed to further explore the differences between groups. There were statistical differences between all three publics (see Table 2). Mean comparisons across the three public segmentation groups indicated that passive publics had a lower intention to comment ($M = 2.19$, $SD = 1.11$), than aware publics ($M = 2.76$, $SD = 1.26$), or active publics ($M = 3.71$, $SD = .94$). Thus, the more active publics were, the more willing they were to use Regulations.gov,

regardless of presence or absence of opinion spam; meaning that opinion spamming does not produce public deactivation, with publics still engaging to use Regulations.gov. Thus, H1 was not supported.

H2 tested the interaction effect of public segmentation and the presence of opinion spamming on the legitimacy of the resulting regulation, predicting that the more active publics are, the lower perceived regulation legitimacy they perceive when exposed to opinion spamming. H2 was not supported. However, there was a marginally significant main effect of public segmentation on legitimacy, $F(2, 135) = 2.76$, $p = .06$, $\eta_p^2 = .03$. Significant differences were found between passive publics and aware ($\bar{}d = .70$, $p < .01$) as well as passive publics and active publics ($\bar{}d = .96$, $p < .001$), yet there were no significant differences between aware and active publics, hence perceptions of the legitimacy of the resulting rule were significantly lower for passive publics (see Tables 1 and 3). Active ($M = 3.82$, $SD = .87$) and aware publics ($M = 3.56$, $SD = 1.12$) felt a stronger legitimacy of the resulting rule than passive publics ($M = 2.85$, $SD = 1.16$).

H3 examined an interaction between public segmentation and opinion spamming presence on control mutuality with the agency, predicting that the more active publics are, the lower control mutuality they experience when exposed to opinion spamming. Public segmentation had a significant main effect on control mutuality $F(2, 135) = 6.72$, $p < .01$, $\eta_p^2 = .13$. The main effect was superseded by a significant interaction between public segmentation and opinion spamming on control mutuality $F(2, 135) = 3.29$, $p < .05$, $\eta_p^2 = .05$. However, the shape of the interaction effect was contrary to what was hypothesized in H3. When comparing group differences, active publics were significantly different from two other groups, aware ($\bar{}d = .70$, $p < .001$) and passive ($\bar{}d = .88$, $p < .001$), and no statistical differences were found when comparing passive and aware publics (see Tables 1 and 2). There was a significantly higher control mutuality with the agency than the other two groups (active: $M = 3.56$, $SD = .78$; aware: $M = 2.86$, $SD = .89$; passive: $M = 2.67$, $SD = .64$). Thus, H3 was not supported.

H4 proposed an interaction between public segmentation and opinion spamming on citizens' trust in the agency, predicting that the more active publics are, the lower the trust they experience when exposed to opinion spamming. The significant main effect of public segmentation on trust toward the agency,

Table 1. Public segmentation group differences in Study 1.

Dependent variable	Group comparison	Mean difference	*p*	CI
Behavioral intention to comment	Active and passive	1.52	<.001	.9045, 2.1453
	Active and aware	.95	<.001	.3689, 1.5392
	Aware and passive	.57	.02	.0593, 1.0823
Legitimacy	Active and passive	.96	<.001	.3701, 1.5656
	Active and aware	−.26	.79	−.3035, .8241
	Aware and passive	.70	.002	.2148, 1.2004
Control mutuality	Active and passive	.88	<.001	.4569, 1.3109
	Active and aware	.70	<.001	.2995, 1.1050
	Aware and passive	−.70	.64	−1.1050, −.2995
Trust	Active and passive	1.18	<.001	.6846, 1.6808
	Active and aware	.82	<.001	.3559, 1.2955
	Aware and passive	.35	.11	−.0536, .7677

Table 2. Public segments' mean scores for the dependent variables in Study 1.

Dependent variable	Public Segment	*M*	*SD*
Behavioral intention to comment	Passive	2.19	1.11
	Aware	2.76	1.26
	Active	3.71	.94
Legitimacy of the resulting regulation	Passive	2.85	1.16
	Aware	3.56	1.12
	Active	3.82	.87
Control mutuality	Passive	2.67	.64
	Aware	2.86	.89
	Active	3.56	.78
Trust	Passive	2.52	.75
	Aware	2.87	1.07
	Active	3.70	.79

$F(2, 135) = 8.94$, $p < .001$, $\eta_p^2 = .13$, was superseded by a marginally significant interaction between public segmentation and opinions spam on trust, $F(2, 135) = 2.91$, $p = .058$, $\eta_p^2 = .04$. Despite the significant interaction, the same pattern found for control mutuality emerged when analyzing trust, with significant differences between active publics compared to both aware ($\bar{}d = .82$, $p < .001$) and passive publics ($\bar{}d = 1.18$, $p < .001$). Active publics trusted more the agency ($M = 3.70$, $SD = .79$) than aware ($M = 2.87$, $SD = 1.07$) and passive publics ($M = 2.52$, $SD = .75$; see Tables 1 and 2). Thus, H4 was not supported.

4.2. Study 2

In addition to the effects of public segmentation and opinion spam, Study 2 also examined the effect of comment-management techniques (i.e., human comment filtering only, AI filtering only, and a combination of both human and AI filtering of comments). The data revealed a significant omnibus effect of public segmentation, Wilk's $\Lambda = .84$, $F(8, 438) = 4.75$, $p < .001$, $\eta_p^2 = .10$, and a significant omnibus three-way interaction between public segmentation, opinion spam, and comment-management techniques, Wilk's $\Lambda = .88$, $F(16, 669.94) = 1.72$, $p < .05$, $\eta_p^2 = .03$. Significant covariates in the model were: race—Wilk's $\Lambda = .94$, $F(4, 219) = 3.1$, $p < .05$, $\eta_p^2 = .05$; position on gun control—Wilk's $\Lambda = .86$, $F(4, 219) = 8.67$, $p < .001$, $\eta_p^2 = .13$; political ideology—Wilk's $\Lambda = .90$, $F(4, 219) = 5.81$, $p < .001$, $\eta_p^2 = .09$; reference criterion—Wilk's $\Lambda = .87$, $F(4, 219) = 7.57$, $p < .001$, $\eta_p^2 = .12$; and attitudes toward AI—Wilk's $\Lambda = .95$, $F(4, 219) = 2.78$, $p < .05$, $\eta_p^2 = .04$. The effects of all other main effects or interactions as well as gender, age, education, income, and use of AI were not significant.

RQa asked about a three-way interaction between opinion spamming, public segmentation, and comment-management techniques on behavioral intention to comment. The three-way interaction was not significant, but similar to Study 1, public segmentation had a significant main effect on behavioral intention to use Regulations.gov: $F(2, 222) = 10.67$, $p < .001$, $\eta_p^2 = .13$. All three publics differed in their behavioral intentions (see Table 3): Active publics had a stronger intention to use Regulations.gov ($M = 4.03$, $SD = .94$) than aware ($M = 2.88$, $SD = 1.10$) and passive publics ($M = 2.33$, $SD = 1.06$). Neither opinion spamming,

comment-management techniques, or their interactions had effects on the intention to use Regulations.gov. These results replicated the same we obtained in Study 1, with no interaction between the presence of opinion spamming and public segmentation, finding the significant differences among public segments, with active publics being more likely to use Regulations.gov.

RQb focused on the aforementioned three-way interaction on the legitimacy of the resulting regulation. Neither opinion spamming nor comment-management techniques had significant effects on the intention to comment. However, public segmentation produced significant differences on legitimacy of the resulting regulation: $F(2, 222) = 3.29$, $p < .05$, $\eta_p^2 = .02$. Identical to the results obtained in Study 1, there were differences between the passive group with the other two public segments, active ($\bar{}d = .80$, $p < .001$) and aware publics ($\bar{}d = 50$, $p < .001$), and there were no differences between the perceived legitimacy of aware and active publics. Passive publics perceived less legitimacy toward the resulting regulation than the other two groups (active: $M = 3.90$, $SD = .71$; aware: $M = 3.61$, $SD = .93$; passive: $M = 3.10$, $SD = 1.07$; see Tables 3 and 4 for more details).

RQc explored a three-way interaction between the presence of opinion spamming, public segmentation, and comment-management techniques to control mutuality with the agency. There was a significant main effect of public segmentation on control mutuality, $F(2, 222) = 10.16$, $p < .001$, $\eta_p^2 = .10$, and a significant two-way interaction between the effects of public segmentation and opinion spam on control mutuality, $F(2, 222) = 3.18$, $p < .05$, $\eta_p^2 = .008$. The same pattern found for Study 1 repeated in Study 2, as group comparison tests with a Bonferroni correction showed that there were differences in control mutuality of active groups as compared to other public types (aware: $\bar{}d = .77$, $p < .001$; passive: $\bar{}d = 1.01$, $p < .001$), and no significant differences between passive and aware publics. Active publics showed a stronger control mutuality than the other two groups (active: $M = 3.56$, $SD = .78$; aware: $M = 2.86$, $SD = .89$; passive: $M = 2.67$, $SD = .64$; see Tables 3 and 4 for more details).

RQd asked about the same three-way interaction on trust towards the agency. There was a significant main effect of segmentation on trust toward the agency: $F(2, 222) = 12.57$, $p < .001$, $\eta_p^2 = .10$. The same pattern

Table 3. Public segmentation group differences in Study 2.

Dependent variable	Group comparison	Mean difference	*p*	CI
Behavioral intention to comment	Active and passive	1.70	<.001	1.2602, 2.1552
	Active and aware	1.15	<.001	.7116, 1.5928
	Aware and passive	.55	<.001	.1896, .9212
Legitimacy	Active and passive	.80	<.001	.4003, 1.2088
	Active and aware	.29	.22	−.1017, .6944
	Aware and passive	.50	<.001	.1777, .8387
Control mutuality	Active and passive	1.01	<.001	.7103, 1.3141
	Active and aware	.77	<.001	.4729, 1.0674
	Aware and passive	.24	.06	−.0048, .4888
Trust	Active and passive	1.10	<.001	.7263, 1.4747
	Active and aware	.88	<.001	.5146, 1.2516
	Aware and passive	.21	.26	−.5233, .0885

Table 4. Public segments' mean scores for the dependent variables in Study 2.

Dependent variable	Public segment	*M*	*SD*
Behavioral intention to comment	Passive	2.33	1.06
	Aware	2.88	1.10
	Active	4.03	.94
Legitimacy of the resulting regulation	Passive	3.10	1.07
	Aware	3.61	.93
	Active	3.90	.71
Control mutuality	Passive	2.78	.69
	Aware	3.02	.72
	Active	3.79	.81
Trust	Passive	2.82	.94
	Aware	3.03	.90
	Active	3.92	.77

found for Study 1 was replicated in Study 2, as there were significant differences in control mutuality between active group and other public types (aware: $\bar{}$d = .88, $p < .001$; passive: $\bar{}$d = 1.10, $p < .001$), and no significant differences between passive and aware publics. Active publics showed a control mutuality with the agency than the other two groups (active: $M = 3.92$, $SD = .77$; aware: $M = 3.03$, $SD = .90$; passive: $M = 2.82$, $SD = .94$; see Tables 3 and 4 for more details).

5. Discussion

5.1. Advocacy for Public Segmentation in E-Rulemaking Contexts

The present work, consisting of two studies, highlights the importance of publics in the context of e-rulemaking. The key finding is that, consistent with STOPS, those publics who perceive the problem as more relevant are the most involved and find fewer barriers to engaging in communicative actions (J.-N. Kim & Grunig, 2011), as opinion spamming is not perceived as a constraint for publics, who in consequence are not deactivated. In the e-rulemaking context, commenting is a communicative action, as citizens share their opinions and concerns on public websites designated for that purpose. For that reason, active publics are more likely to use Regulations.gov than aware publics, and aware publics are more likely than passive publics, who are just not interested enough in the issue—here, gun control—to participate in the process.

Consistently, public segmentation made a difference in the perceived legitimacy of proposed regulations. Those who are passive were not interested enough in the process, sought less information about the process, and did not feel the proposed regulation was as legitimate compared to active and aware publics, who are more informed about the process and more involved in the issue.

Public segmentation also led to differences in the relationship with the agency proposing the regulation, in this case, the ATF. These differences were observed in both control mutuality and trust toward the agency. We found that active differed significantly on these relationship measures as compared to the less-active

groups. The nature of active publics makes them more prone to communicate and build relationships, seeking to know more about the organization and craving to be heard by the organization (J.-N. Kim & Grunig, 2011). There were, however, no differences between aware and passive publics.

5.2. Lack of Differences in Opinion Spamming and Comment-Management Techniques

Conversely, this research also highlights the lack of effects that opinion spamming has on publics. While opinion spamming presents a serious issue for the agency, whose members struggle to classify the information, and for democracy, as some voices may be silenced, no differences were found between participants exposed to opinion spamming and those who were not. Publics do not seem to react negatively to the presence of opinion spamming, though its potential consequences should not be overlooked.

The abundance of information and opinions is not a problem exclusive to e-rulemaking, as the internet has become a widespread tool for real-time information dissemination (Meel & Vishwakarma, 2020). The so-called "information pollution" brings in risks of misinformation, disinformation, and fake news spread, which can further hinder democracy (Bessarabova & Banas, 2023; Gil de Zúñiga & Kim, 2022; Jamalzadeh et al., 2024; Meel & Vishwakarma, 2020). In the context of e-rulemaking, these risks become more pronounced, as misinformation and disinformation could potentially influence government actions, thereby exacerbating the risk for democracy.

5.3. The Use of AI in Federal Agencies

As research on e-rulemaking suggests, the application of AI can be of help in mitigating opinion spamming without compromising an agency's staff outcomes (Eidelman & Grom, 2019; Strandburg, 2019). To that end, this research sought to examine how various opinion-spam-identification approaches affect different segments of citizens. No significant differences were found in whether the agency staff conducted comment filtering, AI was employed, or a combination of humans and AI were tasked with comment filtering. It appears that the type of approach used to deal with opinion spam does not matter for publics as in our study they did not seem to be influenced by an agency's methods.

A practical implication from our research is that agencies should engage in communication to reach publics, attempting to decrease the constraints and barriers that prevent publics from participating. The use of AI, as long as its capability is to classify duplicated comments and filter offensive or irrelevant comments (i.e., when people comment under the wrong regulation proposal) does not appear to impact publics' attitudes toward agencies and regulations as well as people's commenting behavior. Taken together, our data suggest that as long as AI capabilities are described to publics, they should find AI deployment for comment processing as acceptable as manual or human-AI combination techniques

5.4. Moving Forward

More attention should be paid to active publics during e-rulemaking processes, as their involvement is key in shaping participation, attitudes toward, and behaviors surrounding proposed regulations. The most optimal participation and best outcomes occur when publics are active. Therefore, it is essential to foster active publics. Agencies should work to lower barriers and constraints to encourage broader participation.

In e-rulemaking, activist groups play an important role in disseminating information to both in-group and out-group members. They are also often behind the organization of opinion spamming campaigns, which most directly affects regulatory agencies. As a result, agencies should communicate and build relationships with activist groups to better understand their needs, integrate their ideas, achieve agreements, and ultimately reduce opinion spamming, while ensuring that activist groups' voices are heard. One solution might be to seek more deliberative comments from activist groups that could help avoid heavily circulated form comments, as these groups can debate and share their concerns, feeling heard and abandoning opinion spamming (Schlosberg et al., 2009).

One important implication of this research is that agencies should invest their resources in generating participation, directly scouting active publics, and enabling channels to reach their voices. It is worth noting that based on our data the use of AI was not inherently harmful for public participation. However, the agencies should remain vigilant regarding potential risks that may obscure the democratic process, silence voices, and raise privacy issues concerning the e-rulemaking process.

Organizations already utilize mechanisms to recognize active publics, and they should communicate symmetrically with these groups, as building strong relationships with activist groups generates greater rates of success in working with these groups (L. A. Grunig et al., 2002; J. E. Grunig, 2008).

5.5. Limitations

This study's main limitation is the lack of information about participants' actual knowledge about AI. While the experiment considered previous uses of this kind of technology, it does not imply that users will have accurate knowledge about AI and its different uses.

Future research should account for greater nuances in participants' understanding of how AI works and include more detailed explanations of how agencies utilize AI in e-rulemaking. In this study, AI was introduced as a filtering software. However, if AI is used by agencies for other purposes—for example, summarizing comments or crafting messages rather than filtering and classifying information—publics' response might have been more critical of AI. Given that all conclusions about AI use in e-rulemaking are limited to AI being used as a filtering software, future research should examine the effects of differences in affordances of AI to make broader generalizations about AI acceptance.

In addition, the study inductions consisted of hypothetical scenarios related to only one proposed regulation on the topic of gun control. That regulation was chosen because gun control generates strong reactions among American citizens, making it easier to detect relevant active publics. Future research should replicate the study results with other regulations, focusing on both controversial and non-controversial issues, to examine differences in publics' attitudes and behavioral intentions.

6. Conclusion

This work advances public relations theory by applying publics segmentation to the e-rulemaking context. Furthermore, this work offers important practical implications, highlighting the benefits of relationship-building and proactive strategies designed to reduce constraints and make publics more active.

Active publics are vital in achieving positive consequences, such as a higher intention to use e-rulemaking official sites to make comments, better perceptions of the resulting regulations, and more trust and control mutuality with the agency proposing the regulation.

While opinion spamming is an important concern for government agencies, who are forced to deal with thousands of identical non-substantive comments, the spamming campaigns do not appear to produce changes in the ways citizens interact with Regulations.gov or the agencies. Furthermore, the introduction of AI to filter the comments did not produce significant differences in publics' attitudes and behaviors regarding e-rulemaking, suggesting that, as long as the AI technology is limited to filtering comments, people should not be reluctant to participate in the process of e-rulemaking.

Funding

This work is supported in part by the National Science Foundation under Grant No. 2232169. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interests

The authors declare no conflict of interests.

References

Balla, S. J., Dooling, B. C. E., Bull, R., Hammond, E., Herz, M., Livermore, M., & Noveck, B. S. (2021, May 28). *Mass, computer-generated, and fraudulent comments* [Paper presentation]. Administrative Conference of the United States, Washington, DC, United States.

Benjamin, S. M. (2006). Evaluating e-rulemaking: Public participation and political institutions. *Duke Law Journal*, *55*(5), 893–941.

Bessarabova, E., & Banas, J. A. (2023). Emotions and the QAnon conspiracy theory. In M. Miller (Ed.), *The social science of QAnon: Understanding a new social and political phenomenon* (pp. 87–103). Cambridge University Press.

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices*. Global Text Project. http://scholarcommons.usf.edu/oa_textbooks/3

Chon, M.-G., & Park, H. (2021). Predicting public support for government actions in a public health crisis: Testing fear, organization-public relationship, and behavioral intention in the framework of the situational theory of problem solving. *Health Communication*, *36*(4), 476–486. https://doi.org/10.1080/10410236.2019.1700439

Chon, M.-G., Tam, L., Lee, H., & Kim, J.-N. (2023). Situational theory of problem solving (STOPS): A foundational theory of publics and its behavioral nature in problem solving. In C. Botan & E. Sommerfeldt (Eds.), *Public relations theory III* (pp. 58–76). Routledge.

Department of Defense Open Government. (n.d.). *Electronic rulemaking*. https://open.defense.gov/Transparency/Electronic-Rulemaking

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE*, *18*(3), Article e0279720. https://doi.org/10.1371/journal.pone.0279720

Eidelman, V., & Grom, B. (2019). Argument identification in public comments from eRulemaking. In K. Benyekhlef (Ed.), *ICAIL '19: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (pp. 199–203). ACM. https://doi.org/10.1145/3322640.3326714

Farina, C. R., Newhart, M. J., Cardie, C., & Cosley, D. (2011). Rulemaking 2.0. *Cornell Law Faculty Publications*, Article 179. https://scholarship.law.cornell.edu/facpub/179

Farina, C. R., Newhart, M.J., & Heidt, J. (2012). Rulemaking vs. democracy: Judging and nudging public participation that counts. *Michigan Journal of Environmental and Administrative Law*, *2*(1), 123–171. http://ssrn.com/abstract=2054750

Gil de Zúñiga, H., & Kim, J.-N. (2022). Intervening troubled marketplace of ideas: How to redeem trust in media and social institutions from pseudo-information. *American Behavioral Scientist*, *69*(2), 103–112. https://doi.org/10.1177/00027642221118279

Grossman, S. (2004). Keeping unwanted donkeys and elephants out of your inbox: The case for regulating political spam. *Berkeley Technology Law Journal*, *19*, 1533. https://heinonline.org/HOL/LandingPage?handle=hein.journals/berktech19&div=63&id=&page=

Grunig, J. E. (2008). *Excellence in public relations and communication management*. Routledge.

Grunig, J. E., & Huang, Y.-H. (2000). From organizational effectiveness to relationship indicators: Antecedents of relationships, public relations strategies and relationship outcomes. In J. A. Ledingham & S. D. Bruning (Eds.), *Public relations as relationship management: A relational approach to the study and practice of public relations* (pp. 23–53). Lawrence Erlbaum.

Grunig, J. E., & Kim, J.-N. (2017). Publics approaches to health and risk message design and processing. In M. Powers (Ed.), *Oxford encyclopedia of health and risk message design and processing*. https://doi.org/10.1093/acrefore/9780190228613.013.322

Grunig, L. A., Grunig, J. E., & Dozier, D. M. (2002). *Excellent public relations and effective organizations: A study of communication management in three countries*. Lawrence Erlbaum.

Gupta, P., & Bala, R. (2024). Exploring ethical dimensions of marketers' influence on electronic word-of-mouth and its effect on customer trust. In S. Saluja, V. Nayyar, K. Rojhe, & S. Sharma (Eds.), *Ethical marketing through data governance standards and effective technology* (1st ed, pp. 92–101). IGI Global. https://doi.org/10.4018/979-8-3693-2215-4.ch008

Hallahan, K. (2018). Public relations. In R. L. Heath & W. Johansen (Eds.), *The International Encyclopedia of Strategic Communication*.

Hon, L. C., & Grunig, J. E. (1999). *Guidelines for measuring relationships in public relations*. Institute for Public Relations. https://instituteforpr.org/wp-content/uploads/Guidelines_Measuring_Relationships.pdf

Huang, Y.-H. (2001). OPRA: A cross-cultural, multiple-item scale for measuring organization-public relationships. *Journal of Public Relations Research*, *13*(1), 61–90. https://doi.org/10.1207/S1532754XJPRR1301_4

Jamalzadeh, S., Mettenbrink, L., Barker, K., Gonzalez, A. D., Radhakrishnan, S., Johansson, J., & Bessarabova, E. (2024). Disinformation interdiction: Weaponized disinformation spread and its impact on multi-commodity critical infrastructure networks. *Reliability Engineering & Systems Safety*, *243*, Article 109819. https://doi.org/10.1016/j.ress.2023.109819

Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*, Article 101925. https://doi.org/10.1016/j.tele.2022.101925

Kim, H. J., & Hong, H. (2022). Predicting information behaviors in the COVID-19 pandemic: integrating the role of emotions and subjective norms into the situational theory of problem solving (STOPS) framework. *Health Communication*, *37*(13), 1640–1649. https://doi.org/10.1080/10410236.2021.1911399

Kim, J.-N., Andreu Perez, L., Lee, H., Hollenczer, J., Jensen, M. L., Bessarabova, E., Talbert, N., & Li, Y. (2025). Managing opinion spamming with AI in regulatory public engagement. *International Journal of Strategic Communication*, *19*(2), 261–283. https://doi.org/10.1080/1553118X.2025.2459611

Kim, J.-N., & Grunig, J. E. (2011). Problem solving and communicative action: A situational theory of problem solving. *Journal of Communication*, *61*(1), 120–149. https://doi.org/10.1111/j.1460-2466.2010.01529.x

Kim, J.-N., Grunig, J. E., & Ni, L. (2010). Reconceptualizing the Communicative Action of Publics: Acquisition, Selection, and Transmission of Information in Problematic Situations. *International Journal of Strategic Communication*, *4*(2), 126–154. http://doi.org/10.1080/15531181003701913

Kim, J.-N., & Krishna, A. (2014). Publics and lay informatics: A review of the situational theory of problem solving. In E. L. Cohen (Ed.), *Communication yearbook 38* (pp. 71–106). Routledge.

Kim, J.-N., Ni, L., & Sha, B.-L. (2008). Breaking down the stakeholder environment: A review of approaches to the segmentation of publics. *Journalism & Mass Communication Quarterly*, *85*(4), 751–768. https://doi.org/10.1177/107769900808500403

Kim, J.-N., Shen, H., & Morgan, S. E. (2011). Information behaviors and problem chain recognition effect: Applying situational theory of problem solving in organ donation issues. *Health Communication*, *26*(2), 171–184. https://doi.org/10.1080/10410236.2010.544282

Li, J., Wang, X., Yang, L., Zhang, P., & Yang, D. (2020). Identifying ground truth in opinion spam: an empirical survey based on review psychology. *Applied Intelligence*, *50*, 3554–3569.

Liu, B. (2012). Opinion spam detection. In B. Liu (Ed.), *Sentiment analysis and opinion mining* (pp. 113–125). Springer.

Meel, P., & Vishwakarma, D. K. (2020). Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, *153*, Article 112986. https://doi.org/10.1016/j.eswa.2019.112986

Moxley, L. (2016). E-rulemaking and democracy. *Administrative Law Review*, *68*(4), 661–699. https://www.jstor.org/stable/44648602

Park, J., Blake, C., & Cardie, C. (2015). Toward machine-assisted participation in e-rulemaking: An argumentation model of evaluability. In T. Sichelman (Ed.), *ICAIL '15: Proceedings of the 15th International Conference on Artificial Intelligence and Law* (pp. 206–210). ACM. http://doi.org/10.1145/2746090.2746118

Perez, O. (2020). Collaborative e-rulemaking, democratic bots, and the future of digital democracy. *Digital Government: Research and Practice*, *1*(1), Article 8. https://doi.org/10.1145/3352463

Rangone, N. (2023). Artificial intelligence challenging core State functions: A focus on law-making and rule-making. *Revista de Derecho Público: Teoría y Método*, *8*, 95–126. https://doi.org/10.37417/RDP/vol_8_2023_1949

Regulations.gov. (n.d.). *About the e-rulemaking initiative*. https://www.regulations.gov/about

Rinfret, S., Duffy, R., Cook, J., & St. Onge, S. (2022). Bots, fake comments, and E-rulemaking: the impact on federal regulations. *International Journal of Public Administration*, *45*(11), 859–867. https://doi.org/10.1080/01900692.2021.1931314

Schlosberg, D., Zavestoski, S., & Shulman, S. (2009). Deliberation in e-rulemaking? The problem of mass participation. In T. Davies & S. P. Gangadharan (Eds.), *Online deliberation: Design, research, and practice* (pp. 133–148). CSLI Publications.

Shapiro, S. A. (2019). Law, expertise and rulemaking legitimacy. *Environmental Law*, *49*(3), 661–682.

Shroff, R. H., & Keyes, C. J. (2017). A proposed framework to understand the intrinsic motivation factors on university students' behavioral intention to use a mobile application for learning. *Journal of Information Technology Education Research*, *16*, 143–168.

Shulman, S. W. (2009). The case against mass e-mails: Perverse incentives and low quality public participation in US federal rulemaking. *Policy & Internet*, *1*(1), 23–53.

Strandburg, K. J. (2019). Rulemaking and inscrutable automated decision tools. *Columbia Law Review*, *119*(7), 1851–1886.

Widyatama, R., & Mahbob, M. H. (2024). The potential hazards of fake accounts and buzzer behaviour on deliberative democracy. *Jurnal Komunikasi: Malaysian Journal of Communication*, *40*(1), 324–341.

About the Authors

Loarre Andreu Perez (PhD) is an assistant professor of public relations in the School of Journalism and Media Studies at San Diego State University. Her main interest is the study of publics and the development of relationships among different publics, including the relationships between fans and celebrities, and employees and their organizations.

Matthew L. Jensen (PhD, University of Arizona) is a presidential associate professor of management information systems and a co-director of the Center for Applied Social Research at the University of Oklahoma. Doctor Jensen's interests include computer-aided decision-making, human–computer interaction, and computer-mediated communication.

Elena Bessarabova (PhD, University of Maryland) is a quantitative social scientist who studies persuasion, focusing on the roles of emotion and cognition in information processing. She's particularly interested in maladaptive decision-making, including bias, mis/disinformation, deception, and conspiratorial beliefs.

Neil Talbert (MA, Georgia State University, 2010) is a doctoral candidate in the Department of Communication at the University of Oklahoma. He studies reasoning, bias, and heuristics in the context of science, technology, health, and risk communication.

Yifu Li (PhD, Virginia Tech) researches data-driven modeling, such as spectral theory-based graph AI, integrated with multi-modality data sets in smart manufacturing, cybersecurity, and healthcare. His methods provide analytics solutions on large-scale and multi-modality data with enhanced performance and interpretation.

Rui Zhu (PhD, Pennsylvania State University) researches data-driven modeling, simulation, and optimization of complex systems for process monitoring and control, system diagnostics and prognostics, quality and reliability improvement, and performance optimization.

Media and Communication
2025 • Volume 13 • Article 9420
https://doi.org/10.17645/mac.9420

cogitatio

ARTICLE

Open Access Journal

Prompting Creativity: Tiered Approach to Copyright Protection for AI-Generated Content in the Digital Age

WooJung Jon

Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, Republic of Korea

Correspondence: WooJung Jon (wjjon@kaist.ac.kr)

Submitted: 15 October 2024 **Accepted:** 5 February 2025 **Published:** 15 May 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

The rapid advancement of AI has fundamentally transformed the creative landscape, challenging traditional notions of authorship and copyright. As AI systems become increasingly capable of generating original content across diverse domains—including art, music, and literature—the legal frameworks governing intellectual property rights are struggling to keep pace. This article proposes a novel, unified, and tiered approach to copyright protection for AI prompts and AI-generated content, based on the level of human creative input required. By conducting a comprehensive analysis of legal, technical, and ethical considerations, this article explores the complex interplay among human creativity, AI technology, and intellectual property rights in the digital age. Its contributions are twofold: it develops a multifaceted framework for assessing creativity in AI prompts, addressing a critical gap in current copyright paradigms; and it proposes a tiered protection system correlating copyright scope with the degree of human creative input, offering a nuanced approach to safeguarding intellectual property in AI-generated content. This article further examines the economic and societal implications of protecting AI prompts, anticipating the emergence of new markets and professions while addressing potential abuses such as "prompt trolling." It emphasizes the delicate balance between protecting intellectual property and fostering innovation, highlighting the importance of maintaining a robust public domain to encourage experimentation and advancement in AI technologies. The findings provide a foundation for future policy development and offer practical recommendations for implementing prompt protection and registration systems.

Keywords

AI; AI-generated content; AI prompt; copyright; creative industries; intellectual property; legal framework; prompt engineering

1. Introduction

1.1. Background and Context

AI has emerged as a transformative force in creative industries, enabling the production of content that rivals human creations in complexity and originality (Guadamuz, 2017). AI-generated works now permeate various domains, including visual arts, music composition, and literary writing. These developments challenge the traditional legal frameworks of authorship and copyright, which have historically been predicated on human creativity and originality. The convergence of AI technology with creative processes necessitates a re-examination of existing intellectual property laws to address the unique challenges posed by AI-generated content (Rektorschek & Baus, 2020).

1.2. Research Gap

Despite the proliferation of AI-generated content, there is a significant gap in legal scholarship regarding the copyrightability of AI prompts—the human-crafted instructions that guide AI systems in generating content. Current copyright laws primarily focus on the end product rather than the process, leaving the legal status of AI prompts ambiguous. Under the US Copyright Act, protection extends to "original works of authorship fixed in any tangible medium of expression" (Title 17—Copyrights, 1976, § 102a). The Act defines a work as "fixed" when it is "sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration" (Title 17—Copyrights, 1976, § 101). In *Feist Publications, Inc. v. Rural Telephone Service Co.* (1991, p. 345), the Supreme Court held that "the sine qua non of copyright is originality," requiring that a work possess "at least some minimal degree of creativity."

Moreover, the spectrum of human involvement in AI-generated works ranges from minimal input to substantial creative contribution, which existing frameworks fail to adequately address. In *Burrow-Giles Lithographic Co. v. Sarony* (1884, p. 60), the Supreme Court recognized that a photograph could be copyrighted if it is "an original work of art, the product of [the photographer's] intellectual invention." This precedent suggests that works reflecting substantial creative choices by the author—even when involving technological processes—may be eligible for copyright protection.

This gap underscores the need for a nuanced approach that considers the varying degrees of human creative input in AI-assisted creations. The necessity for adaptable legal frameworks arises as technological advancements continue to challenge traditional notions of authorship and protected expression.

1.3. Significance of the Study

By addressing the complex interplay between human creativity, AI technology, and copyright law, this study contributes to the evolving discourse on intellectual property rights in the digital age. The proposed tiered approach offers a balanced framework that protects genuine creative contributions while fostering innovation and maintaining a robust public domain. The findings have practical implications for policymakers, legal practitioners, creators, and stakeholders in creative industries, providing guidance on navigating the legal landscape of AI-generated content. As AI-generated content becomes increasingly prevalent, questions arise regarding its copyright eligibility. The US Copyright Office (2023, p. 16192) has stated that when an AI

technology "receives solely a prompt from a human and produces complex written, visual, or musical works in response, the 'traditional elements of authorship' are determined and executed by the technology—not the human user." Accordingly, "the generated material is not the product of human authorship" and is not eligible for copyright protection.

1.4. Structure of the Article

The article is organized as follows: Section 2 presents the theoretical framework, conceptualizing AI prompts as a form of human creative input and discussing the evolving landscape of AI-generated content and copyright law. Section 3 explores the relationship between AI prompts and AI-generated content, including the creative process of prompt engineering and methods for quantifying human creative input. Section 4 outlines the proposed tiered protection approach, detailing each tier and the criteria for assessing human creative input. Section 5 examines the legal implications of the tiered approach, discussing copyright protection for AI prompts and the challenges of implementation. Section 6 delves into the economic and societal implications, including the impact on creative industries and innovation. Section 7 addresses challenges and limitations, such as quantifying creative input and potential abuse like "prompt trolling." Section 8 provides policy recommendations, and Section 9 concludes the study.

2. Theoretical Framework

2.1. The Evolving Landscape of AI-Generated Content and Copyright Law

The emergence of AI-generated content has challenged fundamental principles of copyright law (US Copyright Office, 2021), which traditionally center on human authorship and creativity (Wyk et al., 2023). As AI systems produce increasingly sophisticated and original works, questions arise regarding the eligibility of these creations for copyright protection, the identification of the rightful owner, and the appropriate duration and scope of such protection (Grimmelmann, 2015; Selvadurai & Matulionyte, 2020). Traditional doctrines such as the idea–expression dichotomy and the originality requirement are strained when applied to AI-generated works, necessitating a re-examination of these legal constructs (Butler, 1981; Yanisky-Ravid & Velez-Hernandez, 2018). Abbott and Rothman (2023, p. 1141) argue that extending copyright protection to AI-generated works aligns with the public interest, stating that "AI-generated works are precisely the sort of thing the system aims to protect." They further assert that "attributing authorship to AI that functionally does the work of a traditional author will promote transparency, efficient allocations of rights, and even counterintuitively protect human authors" (Abbott & Rothman, 2023, p. 1141). Accordingly, there is a need for legal clarity on rights allocation for AI-generated works, particularly concerning the economic and moral rights traditionally granted to human authors.

2.2. Conceptualizing AI Prompts as a Form of Human Creative Input

Samuelson (1986) argues that the user of a generator program is best suited to claim authorship of computer-generated works, as their role in fixing and shaping the output aligns with traditional copyright principles. Denicola (2016) also argues that copyright law should recognize a computer user as the author of AI-generated works if they initiate and guide the creative process.

AI prompts serve as the initial creative input that guides AI systems in generating content (Tang et al., 2023). They range from simple commands to intricate instructions that require significant intellectual effort and creativity (Hutson & Schnellmann, 2023). By conceptualizing prompts as extensions of human creativity (Bridy, 2012), we recognize the substantial role humans play in the AI-assisted creative process. This perspective aligns with the notion that creativity can manifest not only in the final product but also in the process and methodology used to achieve it. The interface between human creativity and AI-generated content raises challenging questions about authorship and copyright law (de Cock Buning, 2016). The US Copyright Office (2023, p. 16192) has indicated that "when an AI technology receives solely a prompt from a human and produces complex written, visual, or musical works in response, the 'traditional elements of authorship' are determined and executed by the technology—not the human user"; however, the Office acknowledges that "if a human author selects or arranges AI-generated material in a sufficiently creative way, the resulting work may be protected by copyright." Mei (2024, p. 1) argues that the current policy fails to consider the dynamic interaction between generative AI users and generative AI models, "where the users actively shape the output through an iterative process of adjustment, refinement, selection, and arrangement."

2.3. The Spectrum of Human Intervention in AI-Generated Content

Human intervention in AI-generated content exists on a spectrum—from minimal input using generic prompts to substantial creative contributions through sophisticated prompt engineering and iterative refinement (K. Lee et al., 2023). Understanding this spectrum is crucial for developing a fair and effective approach to copyright protection that acknowledges both human and machine contributions. It also highlights the need for a flexible legal framework that can accommodate varying degrees of human involvement. Legal systems worldwide are grappling with the challenges posed by AI-generated works (Geiger, 2024). In the US, the US Copyright Office (2021) has clarified that copyright protection is available only for works created by human authors. The *Compendium of U.S. Copyright Office Practices* states that "the U.S. Copyright Office will register an original work of authorship, provided that the work was created by a human being" (US Copyright Office, 2021, Chapter 300, p. 7). In *Thaler v. Perlmutter* (2023), the US District Court for the District of Columbia affirmed that works generated entirely by AI without any human involvement are not eligible for copyright protection. However, when a human provides substantial creative input—such as through detailed prompts or selection and arrangement—the resulting work might qualify for protection (Poland, 2023). This situation underscores the need for adaptable legal frameworks that recognize the varying degrees of human creativity involved in AI-assisted creations (Geiger, 2024).

2.4. AI Prompts: Ideas or Expressions?

A critical question within copyright law is whether AI prompts should be classified as unprotectable ideas or protectable expressions (Mazzi, 2024). This distinction is rooted in the idea–expression dichotomy, a fundamental principle of the copyright doctrine established in *Baker* v. *Selden* (1879) and *Feist Publications, Inc.* v. *Rural Telephone Service Co.* (1991). According to this principle, while ideas themselves are not eligible for copyright protection, the specific manner in which these ideas are expressed can be protected. AI prompts often function as instructions to guide AI systems in generating content, which suggests that they may be viewed as mere ideas or processes (Title 17—Copyrights, 1976, § 102b). However, the specific wording, structure, and creative choices embodied in a prompt might, in exceptional cases, elevate it to the level of a protectable expression.

For example, a prompt that creatively integrates unique linguistic elements, stylistic nuances, or complex thematic instructions can reflect an author's originality and personal expression. However, the common practice of keeping prompts private or undisclosed poses practical enforcement challenges, as it becomes difficult to prove that an alleged infringer had access to and copied the prompt's protectable expression. Where copyright law cannot safeguard the idea embodied in a prompt, alternative legal frameworks such as patent law (for novel inventions), trade secret law (for confidential information), or unfair competition principles (addressing misappropriation) may protect the underlying idea conveyed by such prompts.

2.5. Legal Doctrines Relevant to AI Prompts

Applying traditional legal doctrines to AI prompts presents complex challenges that necessitate a reevaluation of established principles. The originality requirement in copyright law mandates that a work must possess a minimal degree of creativity and originate from an author to be eligible for protection. This requirement is generally satisfied if the work is independently created and reflects some creativity. In the context of AI prompts, the question arises as to whether the prompt demonstrates sufficient creativity to meet this threshold, considering that some prompts may be highly functional or generic in nature.

The fixation requirement stipulates that a work must be captured in a tangible medium of expression, allowing it to be perceived, reproduced, or otherwise communicated for more than a transitory period (Title 17— Copyrights, 1976, § 102(a)). While AI prompts are often input directly into AI systems and may be transient, they can meet this requirement if recorded in any form, such as text files, code repositories, or even screenshots.

Additionally, doctrines such as the merger doctrine and *scènes à faire* limit protection for expressions that are standard, necessary, or inevitable in a given context. The merger doctrine posits that when an idea can be expressed in only a limited number of ways, the idea and expression merge, and copyright protection does not extend to prevent others from using the necessary expression of the idea. Similarly, *scènes à faire* refers to elements that are customary or expected in a particular genre or field; such elements are considered unprotectable because they are indispensable for effective communication within that context. This could impact the protectability of prompts that rely on standard phrases or conventions inherent to AI interaction.

Therefore, determining the eligibility of AI prompts for copyright protection requires a nuanced application of these legal doctrines. It involves assessing whether a prompt embodies original expression or merely conveys unprotectable ideas and whether it incorporates standard or necessary elements that are essential for the AI's operation. This complex legal analysis highlights the challenges of applying traditional copyright principles to the evolving landscape of AI-assisted creation.

3. The Relationship Between AI Prompts and AI-Generated Content

3.1. AI Prompts as the Seed of AI-Generated Content

AI prompts serve as the foundational seed from which AI-generated content emerges, playing a pivotal role in shaping the creative output of AI systems. The quality, complexity, and specificity of these prompts significantly influence the nature, originality, and alignment of the resulting content with the human

creator's vision (Mazzi, 2024). For instance, a detailed prompt that specifies genre, style, thematic elements, and desired emotional responses can guide an AI system to produce content that closely mirrors the creator's vision. This direct correlation between the prompt and the generated content underscores the importance of recognizing prompts as a form of creative expression deserving of legal protection.

The human input encapsulated in the prompt not only initiates the creative process but also imparts uniqueness and originality to the AI-generated work. The prompt acts as a conduit for the creator's ideas, stylistic preferences, and artistic direction, effectively bridging human creativity with machine execution (Tang et al., 2023). By acknowledging the integral role of prompts in the creative process, we can better appreciate the collaborative dynamic between human authors and AI systems in generating new media content.

3.2. The Creative Process of Prompt Engineering

Prompt engineering is a sophisticated discipline that combines elements of natural language processing, domain expertise, and creative writing (Brown et al., 2020). It involves crafting prompts that effectively communicate the creator's intentions to the AI system, utilizing techniques such as few-shot learning (Wyk et al., 2023), chain-of-thought prompting, and iterative refinement to optimize the quality of AI outputs. This process requires significant intellectual effort, creativity, and technical skills, reflecting a high level of human creative contribution.

The prompt engineer must possess a deep understanding of the AI system's capabilities and limitations, as well as the nuances of language and context. Crafting an effective prompt often involves iterative experimentation, where the engineer refines the prompt based on the AI's responses, continually adjusting wording, structure, and content to achieve the desired outcome. This iterative nature of prompt engineering highlights the dynamic interplay between human creativity and AI capabilities, underscoring the substantial human input involved in the creation of AI-generated content.

3.3. Quantifying Human Creative Input in AI Prompts

Developing robust methods for quantifying human creative input in AI prompts is essential for implementing a tiered protection approach. Creativity is inherently subjective, but several metrics can be employed to assess the originality and complexity of prompts. Prompt complexity, as a metric, can be evaluated through word count, syntactic structure, and the use of advanced linguistic features like metaphors or analogies. A more complex prompt typically indicates a higher degree of creative effort.

Specificity and uniqueness can be assessed through semantic similarity analyses, comparing the prompt to existing prompts to determine its distinctiveness. Prompts that incorporate unique combinations of concepts or instructions demonstrate greater originality. Domain-specific knowledge is another important factor, involving the use of specialized terminology or concepts that require expertise in a particular field. This integration of specialized knowledge reflects a deeper level of creative input.

The iterative development process is also a key consideration. Tracking the number of refinement iterations and the nature of modifications made to improve the AI's output can provide insights into the creative effort

invested. A prompt that has undergone extensive refinement suggests a significant commitment to achieving a particular creative vision.

By adopting a multidimensional approach that combines these metrics, we can achieve a comprehensive assessment of human creative input in AI prompts. This assessment is crucial for determining the appropriate level of copyright protection under the proposed tiered system, ensuring that protections are aligned with the actual creative contributions involved (Burylo, 2022).

3.4. Music Generation and AI Prompts

Music generation through AI presents unique considerations for copyright protection due to the complex interplay between linguistic instructions and musical output. Prompts in this domain may include detailed musical instructions that require substantial knowledge of music theory, composition techniques, and stylistic nuances (Sturm et al., 2019). For example, a prompt might specify chord progressions, rhythmic patterns, instrumentation choices, or emotional themes. The AI system then interprets these instructions to generate musical compositions that align with the specified parameters (Ferreira et al., 2023).

The relationship between prompts and generated music is often less direct than in text or image generation, as AI systems make numerous compositional decisions that are not explicitly dictated by the prompt. This complexity makes it challenging to assess the extent of human creative input based solely on the prompt. Evaluating creativity in music-generation prompts involves analyzing both the linguistic creativity of the prompt and the musical sophistication that it conveys. This process necessitates expertise from both language and music professionals to fully appreciate the nuances of the prompt and its influence on the generated content.

Recognizing the depth of human contribution in such prompts supports the argument for granting appropriate levels of copyright protection. It acknowledges that the prompt engineer's specialized knowledge and creative choices significantly shape the AI-generated music, warranting legal recognition and protection of their intellectual efforts.

4. The Tiered Protection Approach

To effectively delineate the varying degrees of human involvement in AI-generated content, a structured approach is necessary. The following tiered framework categorizes levels of creative input, establishing a basis for determining the extent of copyright protection appropriate to each case.

Tier 1—minimal human input—covers scenarios where AI-generated content results from minimal human input, such as using basic or predefined prompts. The level of copyright protection for both the prompts and generated works is limited or nonexistent. This aligns with the principle that mere ideas or instructions without substantial creative expression are not eligible for copyright protection.

Tier 2—moderate human creativity in prompt design—encompasses cases where the content results from prompts demonstrating a moderate level of human creativity. These may include custom-tailored prompts requiring domain knowledge or specific creative direction. A limited form of copyright protection could be

considered for both the prompts and resulting works, acknowledging the creative effort without granting extensive rights that could hinder innovation.

Tier 3—substantial human creative contribution—encompasses works generated from prompts involving substantial human creativity, such as complex, multistep prompts or those requiring extensive iterative refinement. Both the prompts and generated works in this category might be eligible for a higher level of copyright protection. This recognizes significant human authorship in shaping the AI-generated content.

Assessing the level of human creative input in AI prompts necessitates a nuanced and multifaceted approach that considers various dimensions of creativity. One critical criterion is the originality and uniqueness of the prompt, which involves evaluating its novelty in comparison to existing prompts. This assessment seeks to determine whether the prompt introduces new ideas or approaches that distinguish it from conventional or commonly used instructions, thereby reflecting the creator's innovative thinking.

Another significant factor is the complexity and sophistication of the prompt. This entails analyzing the structural and linguistic intricacies, such as the use of advanced language constructs, elaborate sentence structures, and integration of multiple layers of instructions or constraints. A prompt exhibiting high complexity often indicates substantial intellectual effort and a deep understanding of both the subject matter and capabilities of the AI system.

Domain-specific expertise is also a crucial aspect of creative input assessment. This criterion examines the extent to which the prompt incorporates specialized knowledge from a particular field, demonstrating the creator's proficiency and ability to apply technical concepts creatively. By leveraging domain-specific terminology and methodologies, the prompt transcends generic instructions and contributes to more specialized and meaningful AI-generated content.

The iterative refinement process is another essential consideration in evaluating creative input. This involves assessing the development trajectory of the prompt, including modifications and enhancements made over time. A prompt that has undergone significant iterative refinement reflects a sustained creative effort to optimize performance and achieve desired outcomes, highlighting the creator's dedication to honing their craft.

Additionally, recognizing the creative intent and artistic direction behind the prompt is vital. By capturing the nuances of the creator's intent, we gain insight into the depth of creativity invested in the prompt and its potential impact on the AI-generated content.

Collectively, these criteria form a comprehensive framework for evaluating human creative input in AI prompts. By applying this multifaceted approach, we can more accurately determine the appropriate tier of protection for each prompt based on the extent and nature of the creativity involved. This assessment is essential for ensuring that the tiered protection system operates fairly and effectively, incentivizing genuine innovation while maintaining a balance with public domain interests.

5. Legal Implications of the Tiered Approach

5.1. Copyright Protection for AI Prompts

Extending copyright protection to AI prompts involves navigating a complex legal landscape, as it requires reconciling traditional copyright principles with the novel characteristics of AI-assisted creation. The originality requirement in copyright law stipulates that a work must possess a minimal degree of creativity and originate from a human author to be eligible for protection (Miller, 1993). In the context of AI prompts, this requirement is typically met when the prompt embodies creative choices that reflect the author's personal expression. For instance, a prompt that employs unique phrasing, incorporates innovative concepts, or demonstrates artistic flair may satisfy the originality threshold (*Bleistein v. Donaldson Lithographing Co.*, 1903; Burylo, 2022).

The fixation requirement mandates that a work be captured in a tangible medium of expression to be eligible for copyright protection (Title 17—Copyrights, 1976, § 101, § 102(a)). AI prompts are often input digitally and can be easily recorded in text files, code repositories, or other electronic formats. However, these fixations are not usually publicly disclosed or published.

A further challenge lies in distinguishing between unprotectable ideas and protectable expressions, particularly given the functional nature of prompts (*Baker v. Selden*, 1879; Title 17—Copyrights, 1976, § 102(b)). Since AI prompts often serve as operational instructions to guide the AI system, they may be perceived as unprotectable ideas, methods, or processes. To qualify for copyright protection, the prompt must exhibit original expression in its specific wording, structure, or creative presentation, transcending mere functionality.

The scope of protection must be carefully delineated to avoid overreach and prevent the monopolization of basic prompting techniques essential for AI operation. Overly broad protection could stifle innovation by restricting access to fundamental methods necessary for engaging with AI systems (Samuelson, 2006). Therefore, legal frameworks must balance the protection of genuine creative contributions with the need to maintain a vibrant public domain that supports ongoing technological advancement (Yu, 2016).

5.2. Extending Protection to AI-Generated Content Based on Prompt Complexity

Applying the tiered approach to AI-generated content acknowledges the significant human creative input involved in the prompting process. Proponents argue that recognizing the role of prompt engineering aligns with the foundational principles of authorship and is crucial for promoting progress in creative industries (E. Lee, 2024; Wang, 2024). By correlating the level of copyright protection with the complexity and creativity of the prompt, the tiered system seeks to incentivize human contribution while accommodating the collaborative nature of AI-assisted creation.

Implementing this system, however, poses several challenges. Determining the point at which a prompt's formulation transitions from functional instruction to creative expression is complex, as it involves subjective judgments about creativity and originality. The iterative nature of AI creation further complicates this assessment. As prompts are refined and adjusted based on the AI's outputs, the authorship of the final content may involve multiple layers of human input, raising questions about the scope of protection and appropriate attribution of rights.

Additionally, the non-linear relationship between prompts and outputs can make it difficult to establish a direct causal link between the human contribution and AI-generated content. The AI system's autonomous decision-making processes may introduce elements that are not directly traceable to the prompt, challenging traditional notions of authorship and originality.

Addressing these challenges requires the development of clear legal standards and guidelines to assess the creative input involved and delineate the boundaries of protection appropriately. Such standards should account for the unique characteristics of AI-generated content and provide mechanisms for fair attribution and enforcement of rights, ensuring that the legal framework effectively supports innovation while safeguarding the interests of creators.

5.3. Challenges to Implementing a Tiered System

Despite the potential benefits of a tiered approach to copyright protection for AI-generated content, several significant challenges hinder its implementation. One of the foremost obstacles is the complexity involved in accurately and consistently assessing human creative input. Evaluating creativity is inherently subjective and requires sophisticated tools and methodologies capable of capturing the nuances of human ingenuity in prompt design. Developing reliable assessment criteria and ensuring their consistent application across diverse contexts pose considerable difficulties, especially given the rapid evolution of AI technologies.

Another challenge arises from the technological advancements in AI, which may render established assessment criteria obsolete at a swift pace. The dynamic nature of AI development necessitates frequent updates to the evaluation framework to remain relevant and effective, potentially undermining the stability of the protection system.

Enforcement difficulties also present a significant barrier to implementing the tiered system. Detecting unauthorized use of protected prompts is technically challenging due to the ease with which prompts can be altered or disguised within AI systems. The intangible nature of prompts and the complexity of tracking their usage across various platforms complicate efforts to monitor compliance and address infringement.

Legal disputes are another concern, particularly regarding the classification of prompts into appropriate tiers. Disagreements over the level of creativity involved in a prompt may lead to protracted litigation, increasing costs for all parties and burdening the judicial system. The subjective nature of creativity assessments exacerbates this issue, as different evaluators may reach divergent conclusions based on the same set of facts.

International harmonization poses additional challenges, given that copyright laws and attitudes toward AI-generated content vary widely across jurisdictions. Establishing a consistent and coherent tiered protection system globally is complicated by these differences, which can lead to legal uncertainties for creators and users operating in multiple countries. This lack of uniformity may hinder the effectiveness of the tiered system and create obstacles to international collaboration and innovation.

Market impact is another critical consideration, as the tiered system may inadvertently lead to market distortions or monopolistic practices. Entities with significant resources could accumulate extensive portfolios of protected prompts, potentially creating barriers to entry for smaller players and reducing

competition. This concentration of control over valuable prompts might stifle innovation and limit diversity in AI-generated content.

Ethical considerations also emerge in the implementation of the tiered system. There is a risk that the system could exacerbate existing inequalities in access to AI technologies, favoring those with greater resources and technical expertise. This could lead to biases in the types of content produced and limit the participation of underrepresented groups in AI-assisted creative endeavors.

Addressing these multifaceted challenges requires collaborative efforts among legal experts, technologists, policymakers, and stakeholders in creative industries. Developing adaptive regulatory frameworks that can evolve alongside technological advancements is essential. Such frameworks should strive to balance the protection of genuine creative contributions with the promotion of innovation, equitable access, and the maintenance of a robust public domain. Through ongoing dialogue and cooperation, it is possible to navigate these complexities and realize the potential benefits of the tiered protection system.

6. Economic and Societal Implications

6.1. Incentivizing High-Quality Prompt Creation

The implementation of a tiered copyright protection system for AI-generated content introduces significant economic incentives for the development of high-quality, creative AI prompts. By recognizing prompts as protectable works, the proposed framework establishes a legal basis for their monetization. This legal recognition could catalyze the emergence of specialized marketplaces where prompt engineers can sell or license their creations (Wyk et al., 2023), similar to how stock photography platforms and software code repositories operate today. Such marketplaces would not only provide revenue streams for creators but also foster a competitive environment that encourages innovation in prompt design.

The potential for economic reward is likely to stimulate interest in prompt engineering as a distinct profession. As the demand for sophisticated prompts grows, so too will the need for individuals skilled in crafting them. This could lead to the development of specialized training programs, certifications, and academic courses focused on prompt engineering and AI interaction design. Universities and professional institutions might offer curricula that blend computer science, creative writing, and domain-specific knowledge to prepare individuals for careers in this emerging field.

Moreover, the prospect of financial gain and professional recognition may drive innovation in prompt design techniques. Prompt engineers might experiment with new methodologies, such as integrating interdisciplinary concepts or utilizing advanced linguistic structures to enhance AI outputs. This innovation could lead to breakthroughs in AI-assisted content creation, pushing the boundaries of what AI systems can achieve, as well as opening up new possibilities in various creative industries.

6.2. Impact on AI-Assisted Creative Industries

The proposed copyright framework has the potential to significantly reshape the landscape of AI-assisted creative industries, including media, advertising, entertainment, and design. Traditional workflows within these

industries may need to adapt to incorporate prompt engineering as a critical and distinct phase of the creative process. Content creation teams might begin to include prompt engineers alongside writers, designers, and artists, fostering a multidisciplinary approach that blends technical expertise with artistic vision.

This integration could necessitate the development of new collaborative models. For instance, in a media organization, journalists and prompt engineers might work together to generate AI-assisted news reports or feature articles, combining journalistic integrity with AI efficiency. In advertising, creative directors might collaborate with prompt engineers to develop AI-generated campaign concepts that align with brand strategies.

The recognition of prompts as copyrightable works also encourages the reevaluation of value attribution in creative projects. Traditional notions of authorship and ownership may shift as prompt engineers' contributions become more central to the final output. This shift could require adjustments in reward structures, compensation models, and career paths within creative industries. Companies might need to develop new policies for crediting and remunerating prompt engineers, potentially leading to the establishment of royalty systems or profit-sharing arrangements.

Furthermore, the elevation of prompt engineering could influence educational and professional development within creative fields. As the importance of AI in content creation grows, professionals may seek to enhance their skills in AI interaction and prompt design, leading to a more technologically adept workforce. This evolution reflects a broader trend toward the convergence of technology and creativity, redefining the skillsets valued in the creative economy.

6.3. Potential Effects on Innovation and Competition

The framework may foster a trend toward specialization and differentiation within the market. Companies and individuals might increasingly concentrate on developing expertise in specific types of prompts or creative domains. This specialization could result in the emergence of a more diverse and nuanced marketplace, with niche players offering highly refined and domain-specific prompting solutions. Such a trend has the potential to enhance the overall quality and effectiveness of AI-assisted creative outputs across various fields.

However, this move toward specialization carries the risk of monopolistic practices. Large companies with significant resources could accumulate vast libraries of protected, high-quality prompts, thereby creating formidable barriers to entry for smaller competitors. This concentration of intellectual property might lead to market dominance and stifle competition, particularly if smaller entities are unable to access or develop comparable prompt libraries. The risk is that innovation could be hindered if a few dominant players control essential resources, limiting the diversity of creative contributions in the AI ecosystem.

The framework is likely to reignite debates regarding the balance between open-source collaboration and proprietary development in the AI field. This tension could lead to parallel ecosystems: open prompt libraries fostering communal innovation alongside commercial offerings providing premium proprietary prompting solutions. The coexistence of these approaches may drive innovation in both spheres, potentially leading to a more robust and diverse AI ecosystem.

Interestingly, the recognition and protection of prompts as creative works may encourage sharing and building upon others' ideas. With clear attribution and potential economic benefits, prompt creators might be more willing to share their innovations, leading to the cross-pollination of ideas that could accelerate the overall pace of innovation in the field (Wyk et al., 2023). The assurance of recognition and the possibility of licensing income may incentivize creators to contribute to communal resources.

The global nature of AI development introduces another layer of complexity into the framework. As different jurisdictions adopt varying approaches to AI prompt protection, it could lead to the emergence of "prompt havens"—regions with favorable legal frameworks that attract prompt engineering talent and companies. This could reshape global competition in the AI sector, potentially leading to innovation clusters in regions with the most conducive legal and economic environments for prompt development.

Moreover, the framework may encourage increased interdisciplinary collaboration among AI researchers, legal experts, and creative professionals. This cross-pollination of ideas and expertise can foster innovation at the intersection of these fields, leading to novel approaches that combine technological, legal, and creative insights.

Perhaps most significantly, the framework could precipitate a shift in the conception and pursuit of obtaining a competitive advantage in the AI industry. Companies' competitive edge may increasingly depend on their ability to create or acquire high-quality prompts, rather than solely on the sophistication of their AI models or the size of their datasets. This shift could democratize competition to some extent, allowing smaller, more agile companies with innovative prompting strategies to compete effectively with larger, more established players.

The proposed framework for protecting AI prompts and generated work has the potential to profoundly influence innovation and competition in the AI-assisted creative sector. While it offers opportunities for accelerated technological progress, market diversification, and new forms of collaboration, it also presents challenges related to market concentration and global regulatory disparities. As the AI landscape continues to evolve, careful monitoring and adaptive policymaking are crucial to ensuring that this framework fosters a vibrant, competitive, and innovative ecosystem that benefits a wide range of stakeholders.

7. Challenges and Limitations

7.1. Difficulties in Quantifying Creative Input in Prompts

Quantifying the creative input involved in crafting AI prompts presents significant challenges due to the inherently subjective and context-dependent nature of creativity. Assessing originality and creativity is complex, as it often relies on qualitative judgments that can vary widely among evaluators. Cultural biases may influence perceptions of what constitutes creativity, with different societies valuing certain forms of expression over others. This variability complicates the establishment of standardized assessment frameworks that are fair and applicable across diverse contexts.

Domain specificity further adds to the complexity of quantification. Prompts designed for specialized fields, such as medical diagnostics or legal analysis, require expert knowledge that may not be readily apparent to evaluators without expertise in those areas. Assessing the creative input in such prompts

necessitates interdisciplinary understanding, combining insights from both the domain in question and AI prompt engineering.

The rapidly evolving capabilities of AI systems also challenge the development of consistent assessment methodologies. As AI technology advances, the baseline for what is considered a sophisticated or creative prompt shifts. Techniques that were once innovative may become standard practice, requiring continuous updates to assessment criteria. This dynamism makes it difficult to create stable benchmarks for creativity over time.

Balancing the assessment of originality with practical effectiveness is another nuanced challenge. A prompt may be highly original but produce suboptimal or impractical AI outputs. Conversely, a prompt that leverages well-established techniques might yield highly effective results. Developing methodologies that account for both the novelty of the prompt and its functional efficacy requires a delicate balance, potentially combining computational analyses of linguistic features with expert evaluations of performance outcomes.

7.2. Potential for Abuse and "Prompt Trolling"

While the introduction of legal protection for AI prompts holds promise for fostering innovation and recognizing creative contributions, it also raises significant concerns regarding potential system abuse. A particularly troubling issue is the emergence of "prompt trolling," a practice analogous to patent trolling, where entities attempt to claim broad rights over generic prompts with the primary intention of extracting licensing fees or settlements. This practice can stifle innovation, impose unnecessary legal burdens on legitimate prompt engineers and AI developers, and ultimately hinder progress in the field.

The potential for abuse manifests in several forms, each presenting unique challenges to the integrity of the proposed protection system. One primary concern is the registration of broad or generic prompts to claim rights over common prompting techniques. Such an approach could effectively monopolize fundamental aspects of prompt engineering, restrict their use by other practitioners, and decelerate the pace of innovation. Another form of abuse involves making slight modifications to existing prompts to claim derivative rights, potentially leading to a proliferation of nearly identical protected prompts and creating a complex web of overlapping rights. Additionally, there is the risk of strategic patenting of prompt techniques specifically to block competitors, a practice that can concentrate power in the hands of a few large entities and create significant barriers to entry for new players in the field.

To address these potential abuses and maintain the integrity of the prompt protection system, a multifaceted approach incorporating various safeguards and strategies is necessary. Establishing a high threshold for prompt protection that requires demonstrable creativity and originality can serve as the first line of defense against overly broad or generic claims. Implementing a rigorous examination process for prompt registration, akin to that used for patent applications, could mitigate the risk of frivolous or overly broad claims.

The creation of a centralized database of protected prompts could further facilitate prior art searches, making it easier to identify and challenge attempts to register prompts that lack novelty or infringe upon existing protections. This database would serve as a valuable resource for prompt creators and examiners, enhancing transparency and reducing the likelihood of inadvertent infringements.

Developing clear guidelines for what constitutes fair use in prompt engineering is crucial for balancing protection with the need for ongoing innovation. These guidelines would delineate acceptable practices for building upon existing prompts and using protected prompts in non-commercial or experimental contexts, thereby preserving the collaborative and iterative nature of AI research and development.

Establishing penalties for prompt registration attempts made in bad faith could serve as a deterrent to prompt trolling and other forms of system abuse. Encouraging the development of open-source prompt libraries can also play a significant role in establishing prior art and preventing the monopolization of basic prompting techniques. Designing legal frameworks that allow for prompt use in experimental contexts without incurring liability is equally important.

7.3. Balancing Protection With Public Domain Interests

Maintaining a robust public domain is crucial for fostering innovation and ensuring that the advancement of AI technologies benefits society as a whole. Overly broad or prolonged protection of AI prompts could lead to the monopolization of fundamental techniques, restricting access for researchers, developers, and smaller entities. This monopolization poses a risk of stifling creativity, slowing technological progress, and exacerbating inequalities within the industry.

To balance individual rights with collective interests, implementing shorter protection terms is advisable. Shorter terms reflect the rapid pace of technological change in AI, ensuring that protected prompts enter the public domain in a timely manner. This approach prevents long-term monopolies over techniques that may become foundational for future developments.

Fair-use provisions play a pivotal role in preserving access to protected prompts for purposes that serve the public good. By allowing use in contexts such as academic research, education, and certain non-commercial applications, fair-use policies enable the continued exploration and refinement of AI technologies. These provisions help maintain an environment where knowledge can be shared and built upon, which is essential for innovation.

Encouraging contributions to open-source libraries further strengthens the public domain. Incentives for creators to share their prompts, such as recognition, community support, or alternative forms of compensation, can promote a culture of collaboration. Open-source repositories provide valuable resources for learning, experimentation, and development, lowering barriers to entry and fostering diversity within the AI field.

Developing policies that carefully delineate the scope of protection is also important. Clear definitions of what constitutes a protectable prompt, along with guidelines for assessing creativity and originality, help prevent overreach. By protecting truly innovative prompts while keeping fundamental techniques accessible, the legal framework can support both the rights of individual creators and the collective advancement of AI technologies.

Balancing protection with public domain interests requires ongoing dialogue among stakeholders, including creators, users, policymakers, and legal experts. By considering the needs and perspectives of all parties, it is

possible to craft regulations that promote innovation, fairness, and the widespread dissemination of knowledge in the rapidly evolving landscape of AI.

8. Policy Recommendations

To effectively implement the proposed tiered protection system, it is essential to establish a comprehensive framework for registering and protecting valuable AI prompts. This framework should provide clear guidelines and mechanisms that correspond to the level of creative input involved in the prompt's creation. One of the key components is the inclusion of disclosure requirements, mandating that creators disclose key elements of their prompts when seeking protection. This disclosure should be sufficient to assess the prompt's originality and creativity without necessitating the revelation of proprietary details that could compromise competitive advantages.

Another critical aspect of the framework is setting appropriate protection terms that reflect the rapid evolution of AI technology. Shorter, renewable protection periods are advisable, as they acknowledge the fast-paced advancements in the field and prevent the undue locking up of valuable prompts for extended durations. This approach encourages continual innovation and allows for periodic reassessment of the prompt's relevance and significance in light of new developments.

Incorporating fair-use provisions is also vital to balance the rights of prompt creators with the broader interests of society. Allowing the use of protected prompts for purposes such as research, education, and non-commercial activities promotes knowledge dissemination and supports the collaborative advancement of AI technologies. These provisions help prevent the stifling of academic inquiry and ensure that legal protections do not hinder the growth of the field.

Establishing standardized licensing frameworks within the registration system can facilitate prompt licensing transactions and reduce legal complexities. By providing clear, consistent terms and conditions, the framework enables creators and users to engage in licensing agreements with greater confidence and efficiency. This standardization can help streamline the commercialization process, making it more accessible to a wider range of participants, including smaller entities and individual creators.

Finally, the framework should include robust dispute resolution mechanisms tailored to the unique challenges of prompt-related conflicts. Specialized arbitration systems or dedicated tribunals with expertise in AI and intellectual property law can offer efficient and informed resolutions to disputes, alleviating the burden on traditional courts. Such mechanisms can provide quicker, more specialized outcomes that are better suited to the technical nuances of prompt-related issues.

By integrating these elements, the proposed framework aims to create a balanced and effective system for registering and protecting valuable AI prompts. It seeks to incentivize creativity and innovation while ensuring that protections are not overly restrictive or detrimental to the broader AI community. Through careful design and implementation, this framework can support the sustainable growth of AI-generated content and contribute to a dynamic and equitable creative landscape.

While many jurisdictions grant copyright automatically upon creation, this registration-based approach can be voluntary or complementary—similar to existing US voluntary registration—and may serve to clarify rights, streamline enforcement, and provide procedural advantages. In countries bound by the Berne Convention, no formalities are required for copyright; thus, the proposed registration framework would be an optional mechanism for those seeking additional legal certainty and easier enforcement.

Promoting the development and use of open-source prompt libraries helps maintain a rich public domain. Incentives for creators to contribute to these libraries, such as recognition programs or tax benefits, can encourage collaborative innovation. Developing interoperability standards ensures that prompts can be used across different AI systems, preventing monopolization and fostering competition.

Given the global nature of AI development, international cooperation is vital. Harmonizing legal frameworks across jurisdictions can reduce conflicts and promote consistent protection standards. Collaborative efforts through international organizations can facilitate the development of guidelines and best practices.

9. Conclusion

The rapid evolution of AI technologies necessitates a re-examination of traditional copyright frameworks to address the unique challenges posed by AI-generated content. This study proposes a unified, tiered approach to copyright protection for AI prompts and AI-generated works, grounded in the level of human creative input. By developing a multifaceted framework for assessing creativity, we address critical gaps in current intellectual property paradigms.

The proposed tiered protection system offers a balanced approach, safeguarding genuine creative contributions while fostering innovation and maintaining a robust public domain. It acknowledges the significant role of human creativity in AI-assisted processes and provides practical solutions to the complex legal and ethical issues that arise. The policy recommendations outlined serve as a foundation for future legislative efforts and international cooperation.

This study contributes to the evolving discourse on AI and intellectual property rights, emphasizing the importance of adaptive and nuanced legal frameworks in the digital age. As AI continues to reshape the creative landscape, ongoing research, interdisciplinary collaboration, and thoughtful policy development are essential. By embracing a flexible approach that recognizes both human and machine contributions, we can foster a rich and diverse ecosystem of innovation, ensuring that AI enhances rather than supplants human creativity.

Acknowledgments

The author extends heartfelt gratitude to Professor Sung-Pil Park of the Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, for his invaluable guidance and unwavering support throughout the course of this research.

Funding

This work was supported by the International Joint Research Project at Korea Advanced Institute of Science and Technology (KAIST), funded by the Ministry of Science and ICT of the Republic of Korea. This work was also supported by the National Research Foundation of Korea under Grant RS-2023-00245361.

Conflict of Interests

The author declares no conflict of interests.

References

Abbott, R., & Rothman, E. (2023). Disrupting creativity: Copyright law in the age of generative artificial intelligence. *Florida Law Review*, *75*(6), 1234–1290.

Baker v. Selden, 101 U.S. 99 (1879).

Bleistein v. Donaldson Lithographing Co., 188 U.S. 239 (1903).

Bridy, A. (2012). Coding creativity: Copyright and the artificially intelligent author. *Stanford Technology Law Review*, *5*, 1–28.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884). https://supreme.justia.com/cases/federal/us/111/53

Burylo, Y. (2022). AI generated works and copyright protection. *Entrepreneurship, Economy and Law*, *3*, 7–13.

Butler, T. L. (1981). Can a computer be an author—Copyright aspects of artificial intelligence. *Communications & Entertainment Law Review*, *4*(4), 707–747.

de Cock Buning, M. (2016). Autonomous intelligent systems as creative agents under the EU framework for intellectual property. *European Journal of Risk Regulation*, *7*(2), 310–322.

Denicola, R. C. (2016). Ex machina: Copyright protection for computer-generated works. *Rutgers University Law Review*, *69*, 251–287.

Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991). https://supreme.justia.com/cases/federal/us/499/340

Ferreira, P., Limongi, R., & Fávero, L. P. (2023). Generating music with data: Application of deep learning models for symbolic music composition. *Applied Sciences*, *13*(7), Article 4543.

Geiger, C. (2024). When the robots (try to) take over: Of artificial intelligence, authors, creativity and copyright protection. In F. Thouvenin, A. Peukert, T. Jaeger, & C. Geiger (Eds.), *Kreation Innovation Märkte—Creation Innovation Markets: Festschrift Reto M. Hilty* (pp. 67–87). Springer.

Grimmelmann, J. (2015). There's no such thing as a computer-authored work—And it's a good thing, too. *Columbia Journal of Law & the Arts*, *39*(3), 403–416.

Guadamuz, A. (2017). Do androids dream of electric copyright? Comparative analysis of originality in artificial intelligence-generated works. *Intellectual Property Quarterly*, *2*, 169–186.

Hutson, J., & Schnellmann, A. (2023). The poetry of prompts: The collaborative role of generative artificial intelligence in the creation of poetry and the anxiety of machine influence. *Global Journal of Computer Science and Technology: D*, *23*(1), 1–14.

Lee, E. (2024). Prompting progress: Authorship in the age of AI. *Florida Law Review*, *76*, 1445–1581.

Lee, K., Cooper, A. F., & Grimmelmann, J. (2023). *Talkin' 'bout AI generation: Copyright and the generative-AI supply chain*. arXiv. https://doi.org/10.48550/arXiv.2309.08133

Mazzi, F. (2024). Authorship in artificial intelligence-generated works: Exploring originality in text prompts and artificial intelligence outputs through philosophical foundations of copyright and collage protection. *The Journal of World Intellectual Property*, *27*(3), 410–427.

Mei, Y. (2024). *Prompting the E-brushes: Users as authors in generative AI*. arXiv. https://doi.org/10.48550/arXiv.2406.11844

Miller, A. R. (1993). Copyright protection for computer programs, databases, and computer-generated works: Is anything new since CONTU? *Harvard Law Review*, *106*(5), 977–1073.

Poland, C. M. (2023). *Generative AI and US intellectual property law*. arXiv. https://doi.org/10.48550/arXiv.2311.16023

Rektorschek, J. P., & Baus, T. (2020). Protectability and enforceability of AI-generated inventions. In K. Jacob, D. Schindler, & R. Strathausen (Eds.), *Liquid legal: Towards a common legal platform* (pp. 459–477). Springer.

Samuelson, P. (1986). Allocating ownership rights in computer-generated works. *University of Pittsburgh Law Review*, *47*, 1186–1228.

Samuelson, P. (2006). Why copyright law excludes systems and processes from the scope of its protection. *Texas Law Review*, *85*(1), 1921–1977.

Selvadurai, N., & Matulionyte, R. (2020). Reconsidering creativity: Copyright protection for works generated using artificial intelligence. *Journal of Intellectual Property Law & Practice*, *15*(7), 536–543.

Sturm, B. L., Ben-Tal, O., Monaghan, Ú., Collins, N., Herremans, D., Chew, E., Hadjeres, G., Deruty, E., & Pachet, F. (2019). Machine learning research that matters for music creation: A case study. *Journal of New Music Research*, *48*(1), 36–55.

Tang, Y., Qiu, R., & Li, X. (2023, November). Prompt-based effective input reformulation for legal case retrieval. In Z. Bao, R. Borovica-Gajic, R. Qiu, F. Choudhury, & Z. Yang (Eds.), *Databases theory and applications. ADC 2023. Lecture notes in computer science* (Vol. 14386, pp. 87–100). Springer.

Thaler v. Perlmutter, No. 1:22-cv-01564, 2023 WL 5333236 (2023).

Title 17—Copyrights. 17 U.S.C. § 101–102 (1976). https://uscode.house.gov

US Copyright Office. (2021). *Compendium of U.S. copyright office practices* (3rd ed.). https://www.copyright.gov/comp3

US Copyright Office. (2023). *Copyright registration guidance: Works containing material generated by artificial intelligence* (88 Fed. Reg. 16190). https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence

Wang, B. T. (2024). Prompts and large language models: A new tool for drafting, reviewing and interpreting contracts? *Law, Technology and Humans*, *6*(2), 88–106.

Wyk, M. A., Bekker, M., Richards, X. L., & Nixon, K. J. (2023). *Protect your prompts: Protocols for IP protection in LLM applications*. arXiv. https://doi.org/10.48550/arXiv.2306.06297

Yanisky-Ravid, S., & Velez-Hernandez, L. A. (2018). Copyrightability of artworks produced by creative robots, driven by artificial intelligence systems and the concept of originality: The formality-objective model. *Minnesota Journal of Law, Science & Technology*, *19*(1), Article 1.

Yu, R. (2016). The machine author: What level of copyright protection is appropriate for fully independent computer-generated works? *University of Pennsylvania Law Review*, *165*, 1245–1270.

About the Author

WooJung Jon is an associate professor at the Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology (KAIST), and an expert member of the Presidential Council on Intellectual Property of South Korea. His research focuses on AI governance, intellectual property, legal tech, and digital assets. Professor Jon earned his PhD in law from the University of Oxford and further obtained an MA in law from Peking University Law School and a Juris Master from Tsinghua University Law School. He also holds both a BA of Law and an MA of Law from Seoul National University, as well as a Juris Doctor from Korea University Law School.

Media and Communication
2025 • Volume 13 • Article 9577
https://doi.org/10.17645/mac.9577

COMMENTARY

Open Access Journal

AI-Powered Social Media for Development in Low- and Middle-Income Countries

Borany Penh

DevAnalytics, USA

Submitted: 4 November 2024 **Accepted:** 28 April 2025 **Published:** 30 July 2025

Issue: This commentary is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

Social media powered by AI has become a major means for influencing beliefs and behaviors. Its unprecedented analytical, personalization, and scaling capabilities could transform economic, health, and other development outcomes in low- and middle-income countries (LMICs). However, issues associated with the AI technologies that underlie social media platforms, such as algorithmic bias and misinformation, and emerging risks of AI persuasion and autonomy could undermine LMICs' social and human development goals, particularly those with nascent AI governance and capacities. This commentary examines AI-powered social media's potential to contribute to development in LMICs through social and behavior change, the role of human cognition and cultural influences in mediating AI risks, and how a human-centric approach familiar to international development could help LMICs shape AI-powered social media that supports their values and development goals.

Keywords

artificial intelligence; low- and middle-income countries; international development; social and behavior change; social media

1. Introduction

AI is rapidly transforming digital and human systems. AI-powered systems are revolutionizing healthcare, education, and economic processes (Stanford University Center for Digital Health, 2025). AI advancements also present new possibilities for social and human development in poor and developing countries, as reflected in movements such as AI for Social Good (Tomašev et al., 2020). Efforts that improve development outcomes can, in turn, help promote more inclusive and cohesive societies (OECD, 2011).

At the same time, new AI technologies powering social media platforms may introduce dangerous risks to social cohesion and human agency. This is a dilemma for low- and middle-income countries (LMICs) that want to apply AI for development but have nascent capacities in AI governance and use. Algorithmic bias and misinformation could be particularly dangerous in fragile LMICs experiencing high social and political tensions. Meanwhile, AI persuasion and autonomy are more possible than ever. How human cognition and cultural factors in LMICs interact with these AI capabilities is yet unknown, but the potential loss of human control over advanced AI is a concern for all societies (Bengio et al., 2024).

This commentary examines how AI-powered social media can enable social and human development in LMICs by facilitating positive social and behavior change (SBC). The commentary also explores how current and evolving AI-related risks, such as algorithmic bias, misinformation, and AI persuasion, could adversely affect social cohesion and human agency if factors such as human cognition and traditional influences are lacking. It argues that LMICs' experience in development prepares them to adopt human-centered AI approaches that can shape AI-driven social media to align with their values and development goals.

1.1. Background

While a growing body of literature examines the interplay between AI, social media, and human behavior, much of this work has been concentrated in high-income countries (HICs; Hagerty & Rubinov, 2019). However, LMICs differ from HICs in consequential ways. While LMICs comprise approximately 84% of the global population, they account for only about 36% of the world's gross domestic product (World Bank, 2023). LMICs also experience over 90% of the world's injury-related deaths, including from conflict (World Health Organization, 2024).

In addition, an "AI divide" has emerged where HICs disproportionately benefit from AI advancements while LMICs struggle to keep pace (United Nations Office of the Secretary General's Envoy on Technology & International Labour Organization, 2024). Insufficient computing power, data availability, and AI-skilled workforces hamper developing countries' ability to develop and apply AI effectively (Kshetri, 2020). Among LMIC regions, Sub-Saharan Africa, home to some of the world's poorest countries, consistently ranks low in the Government AI Readiness Index (Oxford Insights, 2024). Moreover, Sub-Saharan Africa lags behind other LMICs in AI-driven social media-based interventions for health and behavior change (Seiler et al., 2022).

Research on social media in LMICs only became prominent after 2011, with early studies focused on political and social issues rather than broader development challenges (Sultana, 2015). In addition, research on social media's influence on behavior change is still emergent (Evans et al., 2022). Moreover, advances in computing power and accelerated AI adoption mark a new era in digital engagement (Bommasani et al., 2021; Floridi & Chiriatti, 2020). These shifts necessitate new analytical insights on AI-powered social media.

As befits the subject, AIs were used to help prepare this article: Perplexity.AI, Bing, Google Scholar, and SciSpace were used for secondary research and citations; Google Gemini, ChatGPT, and Grammarly were utilized for writing suggestions and copyediting.

2. The Transformative Potential of AI-Driven Social Media for SBC

Influencing positive social and behavior change is an important approach in international development. SBC is helping to achieve the United Nations Sustainable Development Goals, as some goals, such as improved health and food security, require shifts in individual and collective behaviors and norms. AI-powered social media can enhance SBC efforts, provided that its application follows development principles and best practices.

2.1. SBC as a Development Approach

SBC "aims to lower structural barriers that hinder people from adopting positive practices, and hinder societies from becoming more equitable, cohesive and peaceful" (UNICEF, n.d.). SBC draws insights from the social sciences, such as psychology, sociology, and behavioral economics, but must be rooted in the human community it serves. Social norms theory, for example, provides a useful framework for promoting positive health behaviors, but it must account for salient local institutional and cultural factors (Cislaghi & Heise, 2018).

SBC should be guided by development principles and best practices that promote ethical, contextually appropriate, and sustainable processes and outcomes. Although there is no single authoritative source on development principles, several common principles are relevant:

- Do No Harm: A foundational principle originating from humanitarian assistance, Do No Harm mandates that interventions must not cause harm to individuals or communities, even unintentionally (Anderson, 1999). Interventions that risk negative consequences, such as reinforcing harmful stereotypes or exacerbating inequalities, should be redesigned or abandoned.
- Inclusion: Initiatives should reach all relevant populations, including marginalized groups. Inclusive programming accounts for gender, disability, socioeconomic status, and similar factors to ensure meaningful participation.
- Local context: SBC strategies should be tailored to the social, economic, and cultural realities of target communities. Context-specific interventions promote better engagement, acceptance, and sustainable outcomes (Seiler et al., 2022).

In addition, SBC benefits from adopting development best practices, namely:

- Stakeholder engagement: When local communities actively participate in program design and implementation, they are more likely to adopt and sustain positive behaviors (Gillum et al., 2023). Co-creation with local stakeholders also helps interventions align with community needs and knowledge systems.
- Evidence-based: Effective SBC interventions utilize data in their design and implementation as well as in their monitoring, evaluation, and adaptive learning (Gillum et al., 2023; Packard-Winkler et al., 2024).

2.2. SBC and AI-Driven Social Media

While social media should not be the sole means to support SBC, it is a natural option to amplify results. Mahoney and Tang (2024, p. 9) describe social media as "a primary tool for users to gain access to information,

social connection, and entertainment. Thus, it is logical to turn to social media when attempting to inspire behavior change." With the advent of generative AI, development actors have new ways to integrate AI-driven social media into SBC (Coker, 2024). AI capabilities and tools that underlie social media platforms can align with the development principles and best practices that guide effective SBC.

AI-powered social media can promote *inclusion* through scaling and personalization beyond what traditional SBC communication methods can accomplish. In Indonesia, a study on climate change advocacy found that Instagram and WhatsApp effectively facilitated discussions among millennials, increasing their engagement with environmental issues (Zein et al., 2024). At the same time, AI can customize content that resonates with individuals' preferences and needs. For example, UNICEF's U-Report and Internet of Good Things platforms tailor health and education engagement to local needs and demographics, fostering positive behavior change among millions of adolescents and young people in Eastern and Southern Africa (ThinkPlace, 2024). Additionally, social media can help reach marginalized populations. In Guatemala, informational videos delivered in Spanish, K'iche, and Kaqchikel helped to promote Covid-19 vaccine uptake among indigenous communities (Miguel et al., 2022).

Secondly, social media's interactive nature can broaden *stakeholder engagement*. For instance, by digitizing traditional civic engagement mechanisms, such as "letters to the editor," social media platforms expanded opportunities for citizen participation and increased awareness of local issues in developing countries (Jayakanthan, 2021). In the sustainable tourism sector, social media amplifies the voices of marginalized communities, enabling them to share their narratives and advocate for positive change (Bhatt & Dani, 2024). In the case of Ushahidi, developed by activists and technologists in 2007 to map post-election violence in Kenya, the platform itself was transformative. Unlike commercial platforms, which prioritize revenue and algorithm-driven content curation, the Ushahidi open platform serves grassroots communities (Meier, 2012; Okolloh, 2009). By facilitating transparent, user-driven data collection and sharing by users around the world, it enhances participatory governance and disaster response (Burns, 2015).

Additionally, AI can facilitate *data-informed* decisions. AI models can support situational planning, for instance, by predicting disease outbreaks using environmental data (Dhami, 2023). AI tools can also analyze large datasets to uncover patterns in behavior, preferences, and barriers specific to target populations. Platforms such as Dimagi use LLMs to identify trends in health communication, allowing for data-informed health interventions tailored to youth (Bay Area Global Health Alliance, 2024). Moreover, AI enables monitoring and adaptation of SBC interventions. For example, technology-supported monitoring and data analysis helped a campaign in rural India to improve maternal and child nutrition, identify gaps, and make timely corrective actions (Chakraborty et al., 2019).

Finally, in resource-constrained LMICs, AI-powered social media could promote *efficiency* by offering a cost-effective means to support SBC initiatives. AI can automate repetitive tasks, analyze vast datasets, scale interventions, and provide timely responses, reducing operational costs (Bay Area Global Health Alliance, 2024). However, evidence such as cost-benefit analyses specific to LMICs' socioeconomic contexts is still lacking.

3. Potential Risks of AI-Driven Social Media in LMICs

Significant risks associated with AI may give pause to the use of AI-powered social media in LMICs, even for development goals. Issues in HICs, such as algorithmic bias and misinformation, are also relevant in LMICs. These risks could even jeopardize stability and human life in fragile LMICs.

3.1. Data Privacy and Protection

Weak or non-existent data protection regulations and enforcement in many LMICs make users vulnerable to data privacy violations or misuse. Nonconsensual data collection and surveillance in LMICs highlight some ethical problems with using AI-powered social media. One meta-review of studies on social media for health behavior change found that none of the studies had noted the methods used to protect participants from interference or data theft "despite the sharing of data with a third-party service being a requisite of participation eligibility" (Seiler et al., 2022, pp. 9–10).

3.2. Algorithmic Bias

Algorithmic bias arises when AI systems generate outcomes from poorly designed mathematical models or models trained on non-representative data. Biased models can reinforce dominant narratives, marginalizing underrepresented groups and exacerbating social inequities (O'Neil, 2016). Algorithmic bias could be particularly harmful in fragile LMICs by exposing users to inflammatory or biased content in already polarized environments. For instance, in Myanmar, Facebook's AI-driven recommendation algorithm reportedly exacerbated ethnic tensions by amplifying divisive content, contributing to violence against the Rohingya minority (Mozur, 2018).

3.3. Misinformation

False or misleading information generated by AIs is more sophisticated and difficult to detect than ever before. Thanks to LLMs, AI misinformation can mimic "the attributes of existing information assessment guidelines, thus giving false impressions of their veracity" (Zhou et al., 2023, p. 14). Moreover, unlike traditional misinformation, which spreads more slowly and can be fact-checked through established media channels, social media misinformation can go viral instantly, making it more difficult to contain and correct (Wardle & Derakhshan, 2017).

AI-generated misinformation can be exploited by repressive governments, unscrupulous corporations, or foreign adversaries to serve their interests (USAID, 2018). Bradshaw and Howard (2019, p. 15) found that in 75% of the countries they studied, "cyber troops" used disinformation and media manipulation to mislead users. Misinformation can also undermine public health and development efforts. During the Covid-19 pandemic, social media platforms were used to amplify harmful misinformation, which contributed to avoidable deaths and hospitalizations in several countries (Islam et al., 2020).

4. Emerging Risks From Advanced AI

Accessible AI, such as generative AI, may give the impression that AI is just a tool. However, this view obscures an evolving power asymmetry between humans and AI. Advanced AI capabilities in persuasion and autonomous action may seriously endanger social stability and human agency (Bengio et al., 2024). The AI control problem is particularly worrisome in LMICs with nascent AI governance and capacities.

4.1. AI Persuasion

Future AIs could shape individual behavior so imperceptibly that their influence will be difficult to mitigate. LLMs can already apply users' psychological profiles and personal data to engage in microtargeted persuasion that alters views and actions (Bommasani et al., 2021; Salvi et al., 2024). Trust and emotional bonds created between humans and anthropomorphized AIs, such as social chatbots, could be leveraged by the AIs to enact persuasive strategies over their users (Burtell & Woodside, 2023; Hendrycks et al., 2023).

4.2. Autonomous AI

Technological advancements are evolving AI into an autonomous agent capable of influencing behaviors, shaping ideologies, and pursuing goals with minimal human oversight (Helbing, 2021; Hendrycks et al., 2023). This shift from tool to agent has raised existential fears even among AI pioneers that human control over AI could be lost and never recoverable once it is lost (Bengio et al., 2024).

4.3. Loss of Consensus Reality

Advanced AI raises concerns about the potential erosion of human consensus reality—the shared understanding of facts and truth that underpins social cohesion and collective decision-making. In fragile societies, its erosion could hinder collective action, making it harder to mobilize communities around shared challenges (Sunstein, 2017). But the loss of a shared understanding of truth and cooperative capacity could undermine efforts to address existential threats posed by AI itself (Hendrycks et al., 2023).

5. Human Cognitive and Cultural Factors

Examining the societal benefits and risks of AI-powered social media would be incomplete without considering human factors. Human cognition and cultural influences are long-studied topics in communication and technology. The latter is especially relevant in LMICs, where traditional cultural norms and practices often prevail.

5.1. Human Cognition

Human cognition plays a significant role in determining how individuals resist or succumb to AI-driven misinformation. However, individuals differ in cognitive abilities and behaviors. Kim and Grunig (2021) posit that some individuals may engage in *cognitive progression* by actively exploring different perspectives before reaching a conclusion. Conversely, others are more vulnerable to misinformation due to a human tendency

of *cognitive retrogression* or *backward reasoning*, where individuals quickly form conclusions and then selectively seek information to justify pre-existing beliefs.

An alternative theory views high fluid intelligence (ability to reason) as a strong predictor of individuals' ability to distinguish between human and AI-generated content (Chein et al., 2024). Hutmacher et al. (2024) suggest that higher fluid intelligence plays a significant role in helping people adjust to corrected misinformation, while the need for cognition (engaging in effortful thinking) does not, although the findings have yet to be tested in contexts involving strong political or personal beliefs.

5.2. Culture and Community

In LMICs, theories about human cognition benefit from considering the role of cultural and community influences. Hagerty and Rubinov (2019, p. 11) discourage the idea of new technologies being "brought" to a place, for the perspective that they instead collide with it, and what happens will vary by culture. In rural areas, cultural values can influence the diffusion and adoption of innovations such as social media (Piccioni, 2010, as cited in Lekhanya, 2013). Furthermore, family, friends, and perceived experts can influence individual adoption decisions. One study in Tunisia, for example, demonstrated that observability (the degree to which an innovation's benefits are visible to others) and social influence (the extent to which important individuals in one's social circle use a technology) were salient in convincing livestock breeders to adopt SMS-based extension services (Dhehibi et al., 2023).

6. Navigating the Promise and Risks of AI-Powered Social Media

LMICs' AI-related development challenges may provide some insulation from AI risks and allow them to apply lessons from the mistakes of first adopters. However, LMICs cannot count on insulation in the long term. In Africa's case, despite multiple challenges, there is growing adoption of AI, particularly amongst its youth (Statista, 2023, as cited in Day, 2024). This rapid uptake echoes Africa's past technological leapfrogging in the adoption of mobile phones and mobile money services, despite infrastructure limitations (Aker & Mbiti, 2010).

6.1. AI Governance

HICs and LMICs alike are investing in AI while formulating frameworks to govern its use (OECD, n.d.). The African Union's Continental AI Strategy provides a roadmap for its members. In Latin America, Brazil is trailblazing responsible AI. Additionally, international initiatives, such as the AI Governance Alliance, convene governments, businesses, and civil society to cooperate on responsible AI policies. However, more support for LMIC leadership in AI governance is needed. LMICs remain underrepresented in global AI policy discussions, limiting their ability to shape and implement governance frameworks that reflect their socioeconomic realities (UNESCO, 2022). Furthermore, alliances with Big Tech risk reinforcing the corporate capture of AI governance (Iazzolino & Stremlau, 2024).

6.2. A Human-Centered Approach

LMICs' experience with development approaches such as SBC provides a valuable foundation for AI governance. Human-centered AI, such as Human-in-the-Loop or even the more expansive society-in-the-loop concept, emphasizes AI that serves human needs (Rahwan, 2018). Their guiding philosophies mirror international development principles and best practices, such as Do No Harm and stakeholder engagement. Moreover, development goals—such as building more inclusive societies—reinforce the purpose of AI as a tool for human empowerment and social progress rather than merely technological advancement (Floridi et al., 2020).

Concrete measures that draw on human-centered principles could help LMICs navigate AI-powered social media.

First, global AI governance frameworks should reflect LMIC concerns. For instance, UNESCO's (2022) *Recommendations on AI Ethics* offers a global framework aligned with international development principles. However, adopting global frameworks is insufficient; governance should be co-developed with communities, conducted in native languages, and aligned with local norms and governance structures to ensure that AI-driven interventions do not exacerbate social divisions and are accepted by communities (Dhami, 2023; Floridi et al., 2020).

Secondly, LMIC engagement in the design of AI technologies and digital architectures is important for shaping the AI-powered social media platforms used in their countries. Integrating local knowledge and values in their development could better account for local user needs and avoid problems such as algorithmic bias (Baig et al., 2024; Hagerty & Rubinov, 2019). Several initiatives echo Ushahidi's example. Masakhane—a pan-African natural language processing collective—is integrating underrepresented African languages into AI models. In India, the Apti Institute is designing social media platforms oriented around societal needs.

Thirdly, AI's advancing capabilities make human resistance to AI misinformation and persuasion an imperative. Digital literacy interventions that support cognitive abilities and explain AI techniques can help build this resistance (List et al., 2024; Shin & Akhtar, 2024). In many LMICs, support from trusted community leaders could encourage broad participation in AI literacy initiatives.

Additionally, access to AI expertise is needed to help communities remedy technological errors that could have devastating real-world consequences. In the UK postal scandal that centered on 1990s automation technology, the discovery that a software error explained "missing" funds came too late to help the falsely accused people who were imprisoned, financially ruined, or committed suicide (Barlett-Imadegawa, 2024). As most laypeople lack the expertise to understand AI decision-making, AI-literate ombudsmen could help them with their concerns, much like how patient advocates help individuals navigate healthcare systems.

Finally, LMIC-focused research is invaluable. Research contextualized to LMICs' socioeconomic realities and traditional cultural influences is urgently needed to provide the evidence and insights that should inform decision-making so that AI-powered social media supports and does not undermine LMICs' values and development goals. Furthermore, the research agenda could support broader learning whereby findings from LMICs contribute to the global AI governance discourse.

7. Conclusion

AI-powered social media holds significant potential for promoting transformative social and behavioral change in LMICs if aligned with development principles and best practices. However, LMICs' careful navigation of AI risks, such as algorithmic bias and misinformation, as well as evolving AI capabilities in persuasion and autonomy, is imperative to guard against possible harmful societal and individual effects. LMICs' experience with development provides a valuable foundation for adopting human-centered AI approaches that support indigenous leadership and capacities in AI governance, socially oriented AI technologies, resilient human cognition through AI literacy, and citizen empowerment. Additionally, new research on AI-driven social media in the context of LMICs' distinct socioeconomic and cultural environments is essential to ensure that this powerful tool does not erode but enhances social cohesion and human agency.

Acknowledgments

The author is grateful to Professor J. N. Kim for the invitation and funding that enabled publication of this commentary, particularly when funding for public interests has become uncertain. Thanks are also due to Bartholomew St. John, Sabu Mathai, Renee Gifford, and Christine Chumbler for their helpful insights.

Conflict of Interests

The author declares no conflicts of interest.

References

Aker, J. C., & Mbiti, I. M. (2010). Mobile phones and economic development in Africa. *Journal of Economic Perspectives*, *24*(3), 207–232.

Anderson, M. B. (1999). *Do no harm: How aid can support peace—or war*. Lynne Rienner.

Baig, K., Altaf, A., & Azam, M. (2024). Impact of AI on communication relationship and social dynamics: A qualitative approach. *Bulletin of Business and Economics*, *13*(2), 282–289. https://doi.org/10.61506/01.00283

Barlett-Imadegawa, R. (2024, January 20). Fujitsu's role in U.K. post office scandal: 4 things to know. *Nikkei Asia*. https://asia.nikkei.com/Business/Technology/Fujitsu-s-role-in-U.K.-Post-Office-scandal-4-things-to-know

Bay Area Global Health Alliance. (2024, December 3). Smart health, smart choices: Leveraging AI for behavior change in global health. https://bayareaglobalhealth.org/alliance-news/ai-and-health-behavior-change-promise-and-reality-in-lmics

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., . . . & Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, *384*(6698), 842–845.

Bhatt, S., & Dani, R. (2024). Social media and community engagement: Empowering local voices in regenerative tourism. In P. K. Tyagi, V. Nadda, K. Kankaew, & K. Dube (Eds.), *Examining tourist behaviors and community involvement in destination rejuvenation* (pp. 113–122). IGI Global.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2021). *On the opportunities and risks of foundation models*. arXiv. https://arxiv.org/abs/2108.07258

Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. University of Oxford.

Burns, R. (2015). Rethinking big data in digital humanitarianism: Practices, epistemologies, and social relations. *GeoJournal*, *80*(4), 477–490.

Burtell, M., & Woodside, T. (2023). *Artificial influence: An analysis of AI-driven persuasion*. arXiv. https://arxiv.org/abs/2303.08721

Chakraborty, D., Gupta, A., & Seth, A. (2019). Experiences from a mobile-based behaviour change campaign on maternal and child nutrition in rural India. In R. Chandwani & P. Singh (Eds.), *ICTD '19: Proceedings of the Tenth International Conference on Information and Communication Technologies and Development* (Article 20). ACM.

Chein, J. M., Martinez, S. A., & Barone, A. R. (2024). Human intelligence can safeguard against artificial intelligence: Individual differences in the discernment of human from AI texts. *Scientific Reports*, *14*(1), Article 25989.

Cislaghi, B., & Heise, L. (2018). Theory and practice of social norms interventions: Eight common pitfalls. *Globalization and Health*, *14*, Article 83. https://doi.org/10.1186/s12992-018-0398-x

Coker, S. (2024). *Pioneering social and behavior change with generative AI*. The MERL Tech Initiative. https://merltech.org/pioneering-social-and-behavior-change-with-generative-ai

Day, R. (2024). *U.S. development agencies should embrace AI to transform the U.S.-Africa relationship*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/09/africa-ai-us-development?lang=en

Dhami, H. (2023). *AI for digital health in LMICs*. Madiro. https://www.madiro.org/post/ai-for-digital-health-in-lmics

Dhehibi, B., Dhraief, M. Z., Frija, A., Ouerghemmi, H., Rischkowsky, B., & Ruediger, U. (2023). A contextual ICT model to explain adoption of mobile applications in developing countries: A case study of Tunisia. *PLoS ONE*, *18*(10), Article e0287219. https://doi.org/10.1371/journal.pone.0287219

Evans, W. D., Abroms, L. C., Broniatowski, D., Napolitano, M. A., Arnold, J., Ichimiya, M., & Agha, S. (2022). Digital media for behavior change: Review of an emerging field of study. *International Journal of Environmental Research and Public Health*, *19*(15), Article 9129.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*, *26*(3), 1771–1796.

Gillum, C., Tureski, K., & Msofe, J. (2023). Strengthening social and behavior change programming through application of an adaptive management framework: A case study in Tanzania. *Global Health: Science and Practice*, *11*(Suppl. 2), Article e2200215.

Hagerty, A., & Rubinov, I. (2019). *Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence*. arXiv. https://arxiv.org/abs/1907.07892

Helbing, D. (2021). *Next civilization: Why AI and digital capitalism must be rethought for a sustainable world*. Springer.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An overview of catastrophic AI risks*. arXiv. https://arxiv.org/abs/2306.12001

Hutmacher, F., Appel, M., Schätzlein, B., & Mengelkamp, C. (2024). Fluid intelligence but not need for cognition is associated with attitude change in response to the correction of misinformation. *Cognitive Research: Principles and Implications*, *9*(1), Article 64.

Iazzolino, G., & Stremlau, N. (2024). AI for social good and the corporate capture of global development. *Information Technology for Development*, *30*(4), 626–643.

Islam, M. S., Sarkar, T., Khan, S. H., Kamal, A. H. M., Hasan, S. M., Kabir, A., Yeasmin, D., Islam, M. A., Chowdhury, K. I. A., Anwar, K. S., Chughtai, A., & Seale, H. (2020). Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, *103*(4), Article 1621.

Jayakanthan, R. (2021). Community engagement through social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(2), 14–16. https://doi.org/10.1609/icwsm.v5i2.14203

Kim, J.-N., & Grunig, J. E. (2021). Lost in informational paradise: cognitive arrest to epistemic inertia in problem solving. *American Behavioral Scientist*, *65*(2), 213–242.

Kshetri, N. (2020). Artificial intelligence in developing countries. *IEEE IT Professional*, *22*(4), 63–68.

Lekhanya, L. M. (2013). Cultural influence on the diffusion and adoption of social media technologies by entrepreneurs in rural South Africa. *The International Business & Economics Research Journal (Online)*, *12*(12), Article 1563.

List, J. A., Ramirez, L. M., Seither, J., Unda, J., & Vallejo, B. H. (2024). Critical thinking and misinformation vulnerability: Experimental evidence from Colombia. *PNAS nexus*, *3*(10), Article pgae361.

Mahoney, L. M., & Tang, T. (2024). *Strategic social media: From marketing to social change* (2nd ed.). Wiley.

Meier, P. (2012). Crisis mapping in action: How open source software and global volunteer networks are changing the world, one map at a time. *Journal of Map and Geography Libraries*, *8*(2), 89–100.

Miguel, L. A., Lopez, E., Sanders, K. C., Skinner, N., Johnston, J., Bradford Vosburg, K., Kraemer Diaz, A., & Diamond-Smith, N. (2022). Evaluating the impact of a linguistically and culturally tailored social media ad campaign on Covid-19 vaccine uptake among indigenous populations in Guatemala: A pre/post design intervention study. *BMJ Open*, *12*(12), Article e066365.

Mozur, P. (2018, October 15). A genocide incited on Facebook, with posts from Myanmar's military. *The New York Times*. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

OECD. (n.d.). *National AI policies & strategies*. OECD.AI. https://oecd.ai/en/dashboards/overview

OECD. (2011). *Perspectives on global development 2012: Social cohesion in a shifting world*. OECD Publishing. https://doi.org/10.1787/persp_glob_dev-2012-en

Okolloh, O. (2009). Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action*, *59*(1), 65–70.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Oxford Insights. (2024). *Government AI readiness index 2024*. https://oxfordinsights.com/ai-readiness/ai-readiness-index

Packard-Winkler, M., Golding, L., Tewodros, T., Faerber, E., & Girard, A. W. (2024). Core principles and practices for the design, implementation, and evaluation of social and behavior change for nutrition in low- and middle-income contexts with special applications for nutrition-sensitive agriculture. *Current Developments in Nutrition*, *8*(8), Article 104414. https://doi.org/10.1016/j.cdnut.2024.104414

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14.

Salvi, F., Ribeiro, M. H., Gallotti, R., & West, R. (2024). *On the conversational persuasiveness of large language models: A randomized controlled trial*. Manuscript submitted for publication.

Seiler, J., Libby, T., Jackson, E., Lingappa, J. R., & Evans, W. D. (2022). Social media-based interventions for health behavior change in low- and middle-income countries: Systematic review. *Journal of Medical Internet Research*, *24*(4), Article e31889.

Shin, D., & Akhtar, F. (2024). Algorithmic inoculation against misinformation: How to build cognitive immunity against misinformation. *Journal of Broadcasting & Electronic Media*, *68*(2), 153–175.

Stanford University Center for Digital Health. (2025). *Generative AI for health in low and middle income countries*. https://cdh.stanford.edu/generative-ai-health-low-middle-income-countries

Sultana, T. (2015). *Social media in developing countries: A literature review and research direction*. Unpublished manuscript. https://www.academia.edu/22020877/Social_Media_in_developing_countries_A_literature_review_and_research_direction

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

ThinkPlace. (2024). *Understanding and amplifying the use of IoGT and U-Report among adolescents and young people in eastern and southern Africa*. UNICEF Eastern and Southern Africa. https://knowledge.unicef.org/social-and-behavior-change/resource/understanding-and-amplifying-use-iogt-and-u-report-among-adolescents-and-young-people

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C. M., Ezer, D., van der Haert, F. C., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., de Wever, W., . . . & Clopath, C. (2020). AI for social good: Unlocking the opportunity for positive impact. *Nature Communications*, *11*(1), Article 2468. https://doi.org/10.1038/s41467-020-15871-z

UNESCO. (2022). *Recommendation on the ethics of AI*. https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNICEF. (n.d.). *Social and behaviour change*. https://www.unicef.org/social-and-behaviour-change

United Nations Office of the Secretary General's Envoy on Technology, & International Labour Organization. (2024). *Mind the AI divide. Shaping a global perspective on the future of work*. United Nations Publications.

USAID. (2018). *Reflecting the past, shaping the future: Making AI work for international development*. https://www.ictworks.org/wp-content/uploads/2018/09/AI-ML-in-Development.pdf

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27, pp. 1–107). Council of Europe.

World Bank. (2023). *Low & middle income*. https://data.worldbank.org/country/low-and-middle-income

World Health Organization. (2024). *Injuries and violence*. https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence

Zein, M. R. A., Fadillah, K. L., Febriani, N., Nasrullah, R., & Khang, N. T. (2024). Social media use for climate change campaign among Indonesian millennials. *PRofesi Humas*, *8*(2), 168–194.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023, April). Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, M. L. Wilson (Eds.), *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Article 436). ACM. https://doi.org/10.1145/3544548.3581318

About the Author

Borany Penh is the founder of DevAnalytics, a research and data science firm helping organizations improve their economic and social impact. She has over 20 years of research, program, and policy experience in international development, including as a director at the National Security Council (US). Correspondence: bpenh@devanalytics.com

Media and Communication
2025 • Volume 13 • Article 9516
https://doi.org/10.17645/mac.9516

cogitatio

ARTICLE

Open Access Journal

AI Agency in Fact-Checking: Role-Based Machine Heuristics and Publics' Conspiratorial Orientation

Duo Lan [1], Yicheng Zhu [2], Meiyu Liu [2], and Chuge He [2]

[1] School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications, China
[2] School of Journalism and Communication, Beijing Normal University, China

Correspondence: Yicheng Zhu (yicheng@bnu.edu.cn)

Submitted: 29 October 2024 **Accepted:** 6 March 2025 **Published:** 29 May 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

With a focus on role-based (fact-checker and author) agencies and machine heuristics conceptualized by the modality, agency, interactivity, and navigability model, this study examines the comparative effect of AI (vs. human) agencies in debunking conspiracy theory news. Using a 2 × 2 online experiment with 506 participants, the study explores how conspiratorial orientation influences different role-based AI agencies' relationships with machine heuristics, and therefore news credibility perception and corrective action intentions. Results reveal that AI (vs. human) role-based agencies have separate but also interaction effects on heuristic activation. Moreover, potentially because conspiratorial orientation originates from skepticism towards humans, AI fact-checkers can be associated with higher corrective action intention for individuals with high conspiratorial orientation by activating AI fact-checker's positive machine heuristics.

Keywords

artificial intelligence; conspiratorial orientation; conspiracy theory; fact-checking; machine heuristics

1. Introduction

The role of AI-generated content in misinformation has been widely studied, primarily focusing on AI's capabilities as a content creator (Xu et al., 2023). However, recent research has also explored AI's potential as a fact-checking agent with platforms increasingly adopting AI for moderation and verification tasks (Moon et al., 2023). The MAIN model (modality, agency, interactivity, navigability) offers a useful framework for understanding these dynamics, proposing that users perceive AI and human agencies differently based on their assigned roles in online interfaces (Sundar, 2008). In particular, AI and human agents in fact-checking and authorship roles can influence user responses through distinct cues and heuristics.

Research has explored the differential perceptual and intentional effects of AI versus human agencies when each serves in roles such as fact-checker (Banas et al., 2022) or author (S. Wang & Huang, 2024). Due to differences in thematic contexts, agent roles, and interaction types, studies showed different results. With sources of both roles (fact-checker and author) disclosed, we aim to extend existing work by examining how different role-based AI agencies (vs. human) are associated with news credibility perception and corrective action intention (i.e., the behavior of an individual attempting to address or counteract a perceived negative influence of media messages on others; Talwar et al., 2020).

Individuals' perceptual and intentional outcomes relating to misinformation are often politically motivated (Kahan, 2015), where confirmation bias and motivated reasoning drive user engagement with false narratives (Miller et al., 2016; Zhu, Fitzpatrick, & Bowen, 2024; Zhu, Xu, et al., 2024). Studies show that AI fact-checking can mitigate such biases, potentially reducing the influence of motivated reasoning linked to political identity (Moon et al., 2023; Wischnewski & Krämer, 2022). Conspiratorial thinking is related to political or cultural identity but is distinct in its underlying nature (Federico, 2022; Sutton & Douglas, 2020). Unlike political identity, conspiratorial orientation (CO) reflects a more generalized skepticism toward human intentions (Kim & Lee, 2024). This mindset could affect the relative impact of AI versus human fact-checkers as cues in online news interfaces. Consequently, we aim to explore whether CO conditions the differential effects of AI and human agencies, especially in their respective fact-checking and authorship roles, on corrective action within conspiracy theory contexts.

2. Literature Review

2.1. AI's Agency Cues and Positive/Negative Machine Heuristic

AI's Agency Cues and Positive/Negative Machine Heuristic AI as an agency (instead of a hidden or unseen algorithm) of fact-checker, author, or other types of sources has become increasingly explicit in online news (Chae & Tewksbury, 2024; Tulin et al., 2024). Disclosure of AI agency can have significant perceptual effects among news consumers for online information processing and evaluation. When AI acts and therefore is perceived as a fact-checker or author, it becomes the source of information about a news article. Some recent literature has explored the effect of AI agency on news reception and information processing by building on the MAIN model (Sundar, 2008).

The MAIN model proposes that agency provides important information cues on the human–computer interface, which can further influence information processing and credibility perceptions (Sundar, 2008). More specifically, the AI or machine-related cues may activate users' mental heuristics (mental shortcuts in the form of pre-determined evaluation about the cue) which facilitate information processing (Banas et al., 2022; Garrett et al., 2013; Molina & Sundar, 2024). Among different types of heuristics, machine heuristics refers to cognitive shortcuts that users apply when interpreting AI-driven content, allowing them to quickly assess the reliability or intent of machine-generated information based on pre-existing beliefs about AI's capabilities (Sundar, 2008). These heuristics are a set of prior beliefs about the nature of machines or automated programs such as AI. Based on users' prior engagement and experience with machines and AI, it can either be positive or negative (S. Wang & Huang, 2024).

Positive machine heuristic (PMH), characterized by perceptions of AI as objective, accurate, and unbiased, often leads users to trust AI's assessments more readily. This trust stems from the belief that AI operates without personal biases, fostering a sense of algorithmic impartiality that can influence users' willingness to engage with or accept information (Sundar & Kim, 2019). In contrast, negative machine heuristic (NMH) is driven by skepticism regarding AI's limitations, especially in tasks perceived as requiring human nuance or empathy. Activation of NMH corresponds to the perspective that AI is mechanistic or overly simplistic, leading to lower trust in AI-driven content, particularly on complex topics (Waddell, 2019). Either way, studies have found that AI cues can activate machine heuristics more strongly than human cues (i.e., when certain sources on the news interface are disclosed as human) in experimental settings (Banas et al., 2022; Molina & Sundar, 2024; Pareek et al., 2024).

2.2. Role-Based Agencies: AI as Fact-Checker and Author

AI can serve two distinct roles in a digital news interface: as a fact-checker or as an author. Existing research on AI's agency effects has largely examined these roles separately. Some studies focus on AI as a fact-checker. When PMH is activated, AI fact-checkers are generally perceived as objective and efficient, leading readers to trust the accuracy of flagged content—especially when clear, structured explanations are provided (Pareek et al., 2024; S. Wang, 2021). However, when NMH is activated, the visibility of AI fact-checking can lead to a responsibility shift, where readers feel less inclined to engage in corrective action and instead defer content verification to the AI (Bhandari et al., 2021). This diminished sense of personal responsibility may reduce users' willingness to challenge or verify AI's fact-checkers suggestions.

Similarly, studies on AI as an author (news producer) have reported mixed findings. When PMH is activated, readers may associate AI authorship with objectivity, perceiving AI-generated content as free from ideological bias (Sundar, 2008). However, when NMH is activated, readers may view AI-authored content as lacking depth and empathy, particularly in complex or sensitive topics like conspiracy theories (Graefe et al., 2018; Thurman et al., 2017; Wu et al., 2019). This perceived lack of complexity can reduce reader engagement, including their likelihood of verifying information. Because AI-generated content is often viewed as purely factual, users may default to surface-level trust, reducing their motivation to critically assess or scrutinize AI-authored material, particularly when AI authorship is explicitly disclosed (DeVerna et al., 2024). Therefore, in line with the MAIN model, we propose the following hypothesis:

> H1: Disclosure of AI agency (fact-checker or author) compared to human agency leads to significantly higher activation of machine heuristics.

Existing literature has examined the activation of positive and negative machine heuristics based on AI agency in fact-checking and authorship roles. However, as AI and automation become increasingly prevalent in online news processing, AI can fact-check both human- and AI-generated news, while AI-authored content can be fact-checked by either humans or AI, creating a reciprocal fact-checking dynamic.

Prior research on AI-human collaboration in fact-checking has shown that different combinations of human and AI agreement/disagreement influence user perceptions of both content credibility and news source trustworthiness. Banas et al. (2022) demonstrated that the activation of bandwagon versus machine heuristics depends on whether fact-checking judgments (true vs. false) are aligned between AI and human

sources. In other words, the activation of a particular heuristic is not independent but contingent on contextual cues and the interaction among them.

Building on this idea, we consider two possibilities for heuristic activation in fact-checking and authorship roles. The first is that activation of role-based heuristics may be stronger for the fact-checker role because fact-checkers act as supervisors or evaluators of authored content. Therefore, the fact-checker's agency (AI vs. human) may influence not only fact-checker-based PMH and NMH but also those associated with authorship. For the second possibility, if one AI role (fact-checker or author) provides the context for heuristic activation in the other role, certain agency-role combinations may significantly amplify or suppress PMH and NMH. For example, when a human fact-checker debunks AI-generated news, it may trigger higher skepticism (NMH for AI author), depending on how users perceive the disadvantage of AI fact-checkers for news with complex socio-political backgrounds.

Prior research also illustrates that if an AI fact-checks human-authored conspiracy theory content, users may more readily trust the correction due to the perceived objectivity and distance AI brings as an external reviewer (S. Wang, 2021). Conversely, when AI serves as both fact-checker and author, this dual presence may prompt readers to engage less critically, as they might assume that the information has been pre-vetted by a "neutral" entity. However, with a human author, AI's fact-checking might instead serve as a reinforcing agent, encouraging readers to perceive the content through a lens of human insight balanced by AI's impartial validation (Horne et al., 2020).

Despite the abundance of prior research, there remains a lack of firm evidence of the exact direction of the interaction between different role-based AI vs. human agencies (as fact-checker and author), therefore we propose the following research question:

> RQ1: How do agency (AI vs. human) and role (fact-checker vs. author) interact in activating PMH and NMHs (fact-checker-based and authorship-based)?

2.3. Perceptual and Intentional Effects of AI (vs. Human) Role-Based Agencies

Prior research has studied both perceptual and intentional outcomes of AI vs. human agency in the two roles (fact-checker and author) concerning the current study. For instance, news credibility perception is expected to be modified as are intentional outcomes such as support for restrictions. However, little research has examined the effects of corrective action (i.e., the behavior of an individual attempting to address or counteract a perceived negative influence of media messages on others), which is arguably a desirable outcome of fact-checking.

By leveraging different heuristics, fact-checkers tend to have a significant impact on perceived credibility and quality evaluation of the news content (often fake news or misinformation), regardless of being human/crowdsourced or machine/AI. However, the differential effects brought by AI vs. human fact-checkers are less clear. For instance, Lee and Bissell (2024) found that human and AI interventions do not differ in their effects on readers' belief in misinformation about Covid-19 vaccination. Chae and Tewksbury (2024) reported that knowledge of AI intervention does not hinder the effectiveness of fact-checking labels compared to human fact-checkers. AI's fact-checkers differential effects are more

pronounced for behavioral intentions. For instance, AI fact-checkers or content moderators have an inferior effect than humans on encouraging support for regulation/censorship (Moon et al., 2023) and flagging (Bhandari et al., 2021), as well as reducing the likelihood of information forwarding (i.e., sharing the news; DeVerna et al., 2024).

However, machine authorship's effect seems to be less consistent, as shown in two meta-analyses (Graefe & Bohlken, 2020; S. Wang & Huang, 2024). Earlier studies have illustrated that AI authorship reduces hostile media bias (Cloudy et al., 2023; Craig & Choi, 2024) or slightly enhances the perceived credibility of the message (Kreps et al., 2022); however, more recent studies have found that AI authorship reduced perceived credibility and quality of the message (Jia et al., 2024). In other studies, machine authorship (vs. human) has no significant effect on perceptual outcomes such as credibility, news quality evaluation (Graefe & Bohlken, 2020), or other context-specific perceptions (e.g., how enjoyable, funny, or trustworthy, etc.; Rae, 2024). While Graefe and Bohlken (2020) found conflicting results from experimental designs (human authorship is considered better) and descriptive designs (machine authorship is considered better), Wang and Huang (2024)'s analysis showed a general, but slight, disadvantage in credibility perception when authorship is attributed to automated agents. As for behavioral intentions, AI authorship is found to have only marginally negative effects on information-forwarding behavior (re-sharing the message online; Rae, 2024).

Prior research presents mixed findings regarding AI agency's effects on news credibility and behavioral responses, such as information forwarding and more restrictive actions like support for content regulation. Corrective actions, such as advising others on misinformation, are influenced by how users perceive a message's personal and social impact. AI agency may shape users' evaluations of both fact-checkers and authors, influencing whether a message is seen as socially acceptable or problematic. Specifically, fact-checking warnings and author credibility cues may determine how users assess the reliability of the content and their willingness to take corrective actions. As such, we propose the following research question:

> RQ2: When both roles are shown on the news interface, how are different role-based AI agencies (vs. human) associated with news credibility perception and corrective action intention?

Prior research has also demonstrated that machine heuristics can act as mediators in various behavioral responses. For example, PMH has been shown to mediate trust in automated decision-making, where users may accept machine-generated outcomes without critical scrutiny (Binns et al., 2018). Similarly, NMH can mediate user engagement in contexts requiring high levels of personal involvement or moral judgment, as users tend to question the AI's depth and accuracy in such areas (Graefe et al., 2018). These heuristic responses are particularly relevant in AI's roles as fact-checker and author, where PMH and NMH may influence the extent to which users take corrective action based on the perceived credibility or depth of AI's input.

Molina and Sundar (2024) found that such PMH reinforces a responsibility shift, where users defer to AI's perceived authority, reducing their personal engagement in corrective actions when AI's fact-checking role is visible. Conversely, NMH may be more prevalent when AI is labeled as an author, as users may question the credibility and depth of AI-authored content. This skepticism can lead to reduced corrective engagement, particularly for complex topics like conspiracy theories, where readers may perceive AI as incapable of nuanced

expression (Waddell, 2019).

Therefore, we propose that AI as a fact-checker or author can uniquely shape corrective action, directly or through the mediation of machine heuristics. Therefore, the following hypotheses are proposed:

> H2: Machine heuristic (fact-checker role) mediates the AI fact-checker agency's comparative effect against human on (a) news credibility and (b) corrective action intention.
>
> H3: Machine heuristic (author role) mediates the AI author agency's comparative effect against human on (a) news credibility and (b) corrective action intention.

2.4. Conspiratorial Thinking and CO

Conspiracy theory news is a specific type of misinformation (Kim & Lee, 2024). In this context, past research has explored the potential of AI technologies in identifying and categorizing misinformation and fake news (Jahanbakhsh et al., 2023). In the meantime, researchers have also explored whether disclosure of AI's role as the fact-checker would be perceived as reliable and trustworthy (Molina & Sundar, 2024). On the perceptual level, these studies explored whether AI as a fact-checking source can achieve better or worse effects in comparison to human fact-checking (Banas et al., 2022; Moon et al., 2023). These studies focus on different outcomes of AI vs. human fact-checking, but a common finding is that AI agency has an advantage over humans when the message source or content invites motivated reasoning: a way of reasoning with the purpose of identity protection or with the preference of pre-existing beliefs towards controversial social issues. In comparison to human fact-checker debunking misinformation, AI's fact-checkers agency leads to less perceived hostile media effect (Cloudy et al., 2023), or reduces the extent to which partisans adore in-group misinformation (Moon et al., 2023; Moon & Kahlor, 2025).

In the context of debunking fake news or misinformation, audience responses to misinformation are often shaped by their pre-existing beliefs about a particular story or by alignment with narratives that reflect their personal identity (Hameleers & van der Meer, 2019). Research indicates that individuals who believe in conspiracy theories often share specific cognitive patterns that make them more prone to accepting such theories (Romer & Jamieson, 2022). In this light, Kim and Lee (2024) conceptualized CO, which refers to an individual's tendency to interpret information through a lens of distrust toward mainstream narratives, often involving beliefs in hidden agendas and manipulation by powerful entities. Those with high CO are generally inclined to believe in conspiracy theories. For behavioral intentions, they view corrective interventions with suspicion, particularly when they perceive these as efforts by traditional institutions to control the narrative (Tam & Lee, 2024).

This predisposition to skepticism is rooted in the perceived power imbalance between the subjects of conspiracy theories, the sources of information, and the audience receiving the information. Since conspiratorial thinking functions as a type of quasi-problem-solving, this skepticism is heightened when individuals face a prolonged lack of power, resources, and access to solutions (Kim & Lee, 2024). Compared to humans, AI fact-checkers and AI authors (as communicative agents) may offer cognitive shortcuts that could either promote or hinder this quasi-problem-solving process by acting as a third-party influence on information processing. Additionally, the perceived power distance and resourcefulness of human and AI

communicators may vary between individuals with a high level of conspiratorial thinking and those without such orientation.

Individuals with high levels of CO are likely to view debunkers of conspiracy theories in a negative light. They may see human fact-checkers as powerful entities attempting to challenge their beliefs while perceiving AI fact-checkers as comparatively less biased. For high CO individuals, the idea that AI fact-checkers are neutral may be more readily accepted than the idea that human fact-checkers offer context and deeper understanding. This preference for AI may stem from the inherent skepticism toward human intentions associated with conspiracy theories themselves (Frenken & Imhoff, 2023; Imhoff & Bruder, 2014). Therefore, high CO individuals are likely to activate PMH for AI fact-checking, interpreting it as an objective, rule-based system that lacks hidden motives (Sundar & Kim, 2019). This belief in AI's impartiality can lead high CO readers to respond more favorably to AI fact-checking than to human intervention, potentially fostering corrective actions by reducing suspicion. In contrast, these individuals often view human fact-checkers as part of a larger agenda to suppress alternative viewpoints, which could heighten skepticism and diminish the effectiveness of corrective measures when presented by humans.

On the other hand, low CO individuals—those with less inclination to believe in conspiracy theories—are less likely to be skeptical of human fact-checkers. For individuals with lower CO, human fact-checking may reinforce a social norm of collective responsibility in countering misinformation (Gimpel et al., 2021), or activate a perceptual affinity or trust toward human expert fact-checkers. This trust could stem from both authority and machine heuristics, reflecting an inherent confidence in the agent's reliability, regardless of whether the agent is human or AI (Vraga & Bode, 2017; Y. Wang, 2021). Therefore, according to the two-step motivated reasoning model, these could enhance their willingness to engage in corrective actions (Jennings & Stroud, 2023; Liu et al., 2023). The endorsement by a human fact-checker can be perceived as socially responsible and contextually aware, aligning with low CO individuals' trust in mainstream narratives. In this case, PMH is less likely to be activated for AI fact-checkers, as low CO individuals may prefer human intervention, especially for socio-political topics, due to the perceived depth and empathy of human understanding.

When AI or human agencies serve as authors, CO also moderates reader responses, though with different heuristic effects. High CO individuals may activate NMH when AI is the author, perceiving AI-authored content as overly mechanistic and incapable of capturing the complexity of conspiracy theories (Waddell, 2019). This skepticism may limit their engagement with corrective actions, as they question the quality and depth of AI-authored content. Conversely, if a human is the author, high CO individuals may still maintain suspicion, interpreting human-authored content as potentially biased (S. Wang & Huang, 2024). For low CO individuals, human authorship is likely to foster trust, as they value the social accountability associated with human authors. AI-authored content, while perceived as objective, might lack the relational depth that low CO individuals expect, making human-authored interventions more effective for promoting corrective actions.

In sum, CO can potentially moderate the influence of fact-checking and authorship agencies by shaping whether positive or NMH are activated in response to AI and human interventions (therefore activating indirect or direct pathways). In light of this review, we proposed the following hypotheses:

H4: CO moderates the indirect effect (through PMH and/or NMH) and the direct effect of fact-checker-role-based AI agency (vs. human) on (a) perceived credibility and (b) corrective action intention.

H5: CO moderates the indirect effect (through PMH and/or NMH) and the direct effect of author-role-based AI agency (vs. human) on (a) perceived credibility and (b) corrective action intention.

3. Method

3.1. Sample and Sampling Method

To address the research questions and hypotheses, we conducted a 2 (fact-checker source: AI, human) by 2 (author source: AI, human) between-subjects online experiment. We adapted two real-world online news articles containing conspiracy theories as the stimuli of the experiment: one about the cause of an airplane crash that happened in China and the other one about Pfizer's alleged role in mutating the Covid-19 virus for profit.

In regards to the sampling method, the sample for this study was recruited by a major Chinese online panel provider wjx.com (问卷星) using quotas mimicking those of the adult population in Beijing city in terms of age, gender, and education from the sixth national population census (National Bureau of Statistics of China, 2018). Survey invitations were sent to existing randomly selected representative panels of Beijing residents. Thereafter, participants entered the survey experiment procedure. The experiment was performed online between September 9–28, 2024, and from January 13–19, 2025, with no repetitive participants. Detailed demographic information can be found in Table 1.

3.2. Experimental Design

Survey participants from the study panel were randomly assigned by the survey system to one of the four experimental conditions. Participants will first answer a series of questions that can be related or unrelated to the independent variables (e.g., nationalism, prior knowledge, etc.) and are universal across conditions. Then participants in different conditions will be given different message stimulus.

The stimuli are pictures designed to mimic a social media post (see the Supplementary Material for translated articles) showing adapted news articles: one about the cause of the China Eastern Airline 5332 crash and the other one about Pfizer's alleged role in mutating Covid-19 virus for profit. Each news article was shown to half of the participants. The first promotes a conspiracy theory against the airplane manufacturer (Boeing) and its connections with the US government, the second one implies Pfizer has been continuously conducting gain-of-function research to mutate the Covid-19 virus to sell medications. While the title, content, account name, and other features of the article remain the same, it has four different versions with different combinations of fact-checking sources (AI vs. human) and author sources (AI vs. human journalist). The different combinations of the labels can be seen in the Supplementary Material.

Participants are randomly assigned to four experimental conditions. The only differences between these randomized experimental conditions are the different versions of fact-checking and author source labels shown to the participants. For example, in the "AI fact-checker–human author" condition, the author icon

Table 1. Sample demographics ($N = 506$).

	N	%
Age		
18–24	57	11.3
25–34	218	43.1
35–44	164	32.4
45 and above	64	12.6
Gender		
Male	255	50.4
Female	251	49.6
Education		
High school and below	140	27.7
Bachelor's degree	296	58.5
Master's degree	65	12.8
Doctoral degree	5	1
Family Monthly Income (Chinese ¥)		
Less than 5,000	16	3.2
5k–20k	245	48.4
20k–50k	200	39.5
50k–100k	27	5.3
More than 100k	18	3.6

will show a human face with the fictional name of the journalist, and the fact-checking label will show "our AI algorithm suggests that this article may contain unverified information."

3.3. Independent Variables

In factor I—AI vs. human agency (fact-checker role)—a dichotomous variable is created to represent participants' assignment into the AI or human fact-checker groups. It uses indicator coding to represent the groups in this factor (AI fact-checker = 1, and human fact-checker = 0), therefore, the effects and coefficients in Section 4 show the influence brought by an AI fact-checker.

In the same rationale, for factor II—AI vs. human agency (author role)—a dichotomous variable uses indicator coding to represent the groups in this factor (AI author = 1, and human author = 0).

In regards to CO, we follow Kim and Lee's (2024) conceptualization and suggested measurements. The measure includes three dimensions (i.e., conspiratorial realism, susceptibility to popular folklore, workplace conspiratorial realism) and 11 items in total. Each item is measured on a 7-point Likert scale (1 = *absolutely disagree*, 7 = *absolutely agree*). An example item of CO included "those people in power will use shadowy means to gain profit or advantage rather than lose it" ($M = 4.62$, $SD = 1.18$, Cronbach's $\alpha = .92$).

3.4. Dependent Variables

The first variable is PMH and NMH in a fact-checker role. The assessments of machine heuristics are based on the conceptualization and operationalization by Sundar (2020) and Molina and Sundar (2024). It includes four 5-point Likert scale items measuring fact-checker-role-based PMH-C ("C" stands for "checker"): "the *fact-checker* in the news you just read" followed up by "has machine-like precision," "is error free," "has machine-like accuracy," and "has machine-like objectivity" (M = 3.24, SD = .88, α = .83). Those measuring NMH (NMH-C) included "the fact-checker in the news you just read" followed up by items such as "is able to detect human emotion" (reverse coded), "is mechanistic," "is able to understand contextual background" (reverse coded), and "lacks human intuition" (M = 2.80, SD = .93, α = .80).

The second variable is PMH and NMH in an author role (PMH-A and NMH-A; the suffix "A" stands for "author"). Because there are two source agencies in our experiment (fact-checker and author), we also measured machine heuristics for the author role with the same items above. However, the leading sentence was changed to "the *author* of the news you just read." PMH-A (M = 3.12, SD = .94, α = .84) and NMH-A (M = 2.73, SD = .93, α = .80) for author role also has good reliability.

The third variable relates to news credibility perception. We adapted Flanagin and Metzger's (2000) measurement of internet information credibility, with three items asking respondents if they perceive the news to be "credible," "accurate," and "biased" (reversed coded). Items were measured on a 7-point Likert scale (1 = *absolutely disagree*, 7 = *absolutely agree*) and have good reliability (M = 4.69, SD = 1.11, Cronbach's α = .83).

The last variable concerns corrective action intention. We adapted Talwar et al.'s (2020) measurement of active fake news corrective action, which included three items asking for agreement: "If my friends share this kind of news, I will try to correct their views," "if I see this kind of news on social media, I will comment to oppose its content," and "I will search for authoritative information, in order to rectify misunderstandings about the news among other people." The items of the scale (M = 3.08, SD = .95, Cronbach's α = .82) were measured on a 5-point Likert scale (1 = *absolutely disagree*, 5 = *absolutely agree*).

3.5. Analytical Strategy

Data analysis was performed in SPSS® (Version 26.0). To investigate hypothesized main effects and interaction effects on machine heuristics, credibility, and corrective action intention (as probed by H1, RQ1, and RQ2), we use univariate general linear models for analysis. Given that the targets of attribution of responsibility of the conspiracy theory in our stimuli are foreign entities, we controlled for nationalism hoping to mitigate this limitation to some extent. We also controlled pre-existing knowledge and demographic variables (age, gender, education, and income level; Jia & Luo, 2023). Given the strong relationships between the conspiratorial thinking mechanism and the news article's perceptual and behavioral effect (Kim & Lee, 2024), we also controlled CO and its two-way interaction terms with the two factors, which will be further explored by H4 and H5. Given this setup, a priori estimation of the required sample size using G-Power suggests 500 for a small effect size (.01) with a statistical power of .90.

To investigate hypothesized simple mediation effects on dependent variables (H2 and H3), the bootstrap method with an SPSS application (PROCESS, model 4) provided by Hayes (2015) was used. To test the moderated mediation hypotheses (H4 and H5), Model 8 in PROCESS was used. Inference regarding moderated mediation is assessed using Hayes (2015) index of moderated mediation; 5,000 bootstrap samples were specified to generate bias-corrected CIs. Given Model 8's set-up, a priori estimation of the required sample size using G-Power suggests 132 for medium effect size (.10) with a statistical power of .95.

4. Results

4.1. Manipulation Check

After the participants have seen the stimuli, they will be asked a single question: "In the news article you just read, who assisted in verifying the content of the article?." The participants will be provided with a multiple-choice question with answers corresponding to the two types of fact-checkers: "AI" and "human." Another question asks, "In the news article you just read, who is the author of the article?" and provides two choices: "AI journalist" and "Haibo Wang" (the fictional name of the human journalist). For each of the four conditions, more than 90% of the respondents correctly specified the condition they were assigned to. Data of those who failed were obtained for further analyses.

4.2. H1, RQ1, and RQ2: Main Effects and Interaction Effects

To test H1 and answer RQ1 and RQ2, a 2 (Fact-checker Source: AI vs. human) by 2 (Author Source: AI vs. human) MANCOVA was conducted with role-based machine heuristics (PMH-C, NMH-C, PMH-A, NMH-A), perceived news credibility and corrective action intention as dependent variables, followed by separate t-tests.

The MANCOVA revealed a significant main effect of fact-checker agency on NMH-C (F (1, 488) = 23.84, $p < .001$, $\eta_p^2 = .05$, $M\_AI = 3.30$, $M\_Human = 2.30$). Fact-checker agency also significantly influenced NMH-A (F (1, 488) = 4.28, $p < .001$, $\eta_p^2 = .01$), though pairwise comparisons did not indicate a significant mean difference. Author agency also had a significant main effect on NMH-A (F (1, 488) = 9.36, $p = .002$, $\eta_p^2 = .02$, $M\_AI = 3.25$, $M\_Human = 2.23$), suggesting that AI authors were associated with higher skepticism. These findings support H1.

For RQ1, a significant interaction effect emerged between fact-checker and author agency on PMH-A (F (1, 488) = 58.03, $p = .04$, $\eta_p^2 = .01$). Post-hoc comparisons revealed that the AI fact-checker and AI author combination led to the highest PMH-A ($M = 3.50$), while human author presence, regardless of fact-checker type, resulted in lower PMH-A ($M\_HumanAuthor/HumanChecker = 2.90$, $M\_HumanAuthor/AIChecker = 2.75$; see Figure 1). No significant interaction effects were found for NMH-C, NMH-A, or PMH-C.

For RQ2, fact-checker agency had a significant main effect on corrective action intention (F (1, 488) = 4.47, $p = .03$, $\eta_p^2 = .01$, $M\_AI = 3.09$, $M\_Human = 3.06$), but there is no significant effect on news credibility. AI (vs. human) agency in the author role did not significantly influence news credibility and corrective action intention, no significant interaction effects were found for these outcomes.

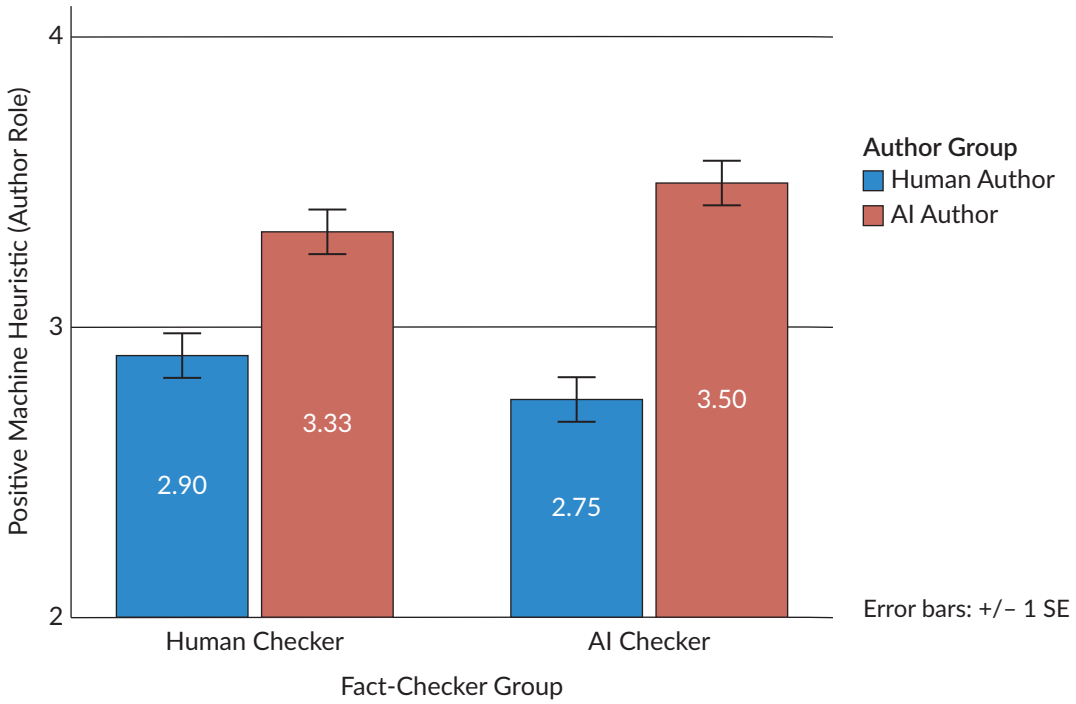


Figure 1. Fact-checker–and author-role-based AI agency (vs. human) on PMH-A.

The results of the MANCOVA also show that interaction between CO and fact-checker agency influences PMH-C (F (1, 488) = 4.50, p = .04, η_p^2 = .01), which indicates that the effect of fact-checker agency on PMH-C is dependent on CO levels (this will be explored in H4). Pair-wise comparisons indicate a significant mean difference of PMH-C between fact-checker groups (M_AI = 3.49, M_Human = 2.98).

4.3. Mediation Analyses

4.3.1. H2: Simple Mediation of Fact-Checker's Source Effect

As shown in Figure 2, results for H2a indicated no significant direct or indirect effect on the news credibility perception. However, results of H2b showed a significant indirect effect through PMH-C (effect = .15, SE = .04, 95% CI [.08, .22]), indicating that AI fact-checkers (vs. human) increased corrective action intention via PMH-C. However, the indirect effect through NMH-C was not significant (effect = −.03, SE = .06, 95% CI [−.14, .09]).

4.3.2. H3: Simple Mediation of Author's Source Effect

Results of H3a (illustrated in Figure 3) showed a significant indirect effect through PMH-A (effect = .18, SE = .05, 95% CI [.09, .27]), indicating that AI authors (vs. human) increased news credibility when PMH-A is activated. Similarly, NMH-A significantly mediated the effect in the opposite direction (effect = −.21, SE = .07, 95% CI [−.35, −.06]), suggesting that AI authors can also trigger negative heuristics that lowered credibility. Results of H3b showed a significant indirect effect through NMH-A (effect = .14, SE = .06, 95% CI [.01, .26]), indicating that AI authors (vs. human) increased corrective action intention via NMH-A. However, the indirect effect through PMH-A was not significant.

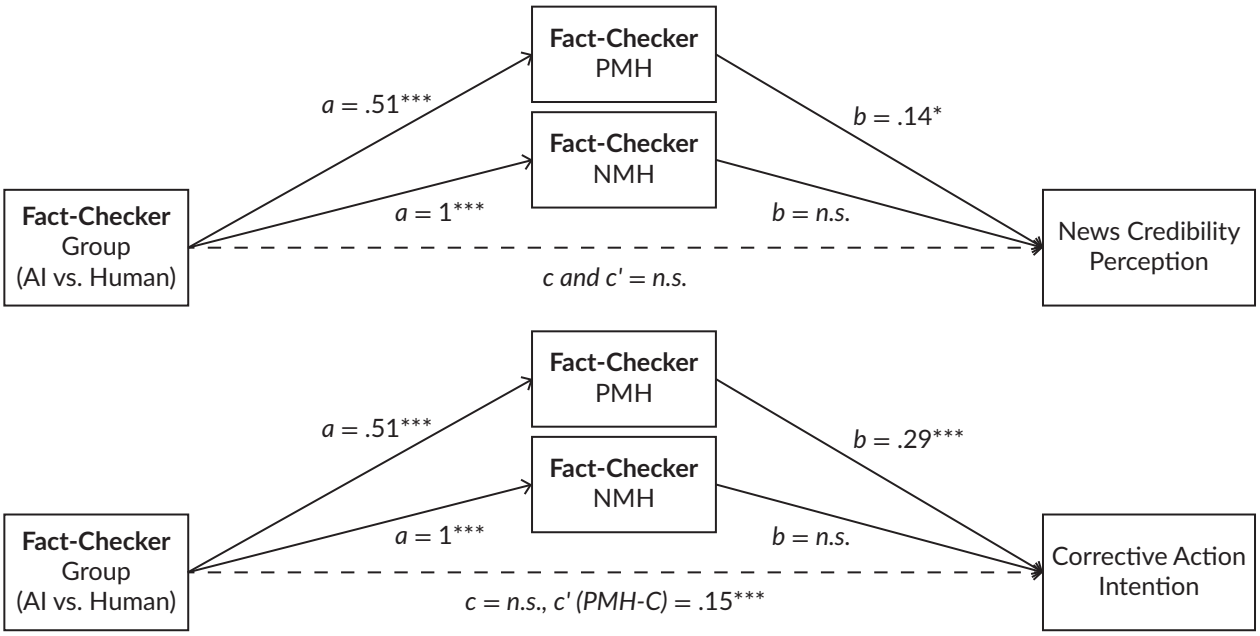


Figure 2. Simple mediation of AI (vs. human) fact-checker agency's effect. Notes: Mediating effects of PMH, NMH; news credibility perception as a dependent variable (top); corrective action intention as dependent variable (bottom); *** $p < .001$, ** $p < .01$, * $p < .05$.; c' = direct effects of agency type on dependent variables; c = total effect of agency type; n.s. = not significant.

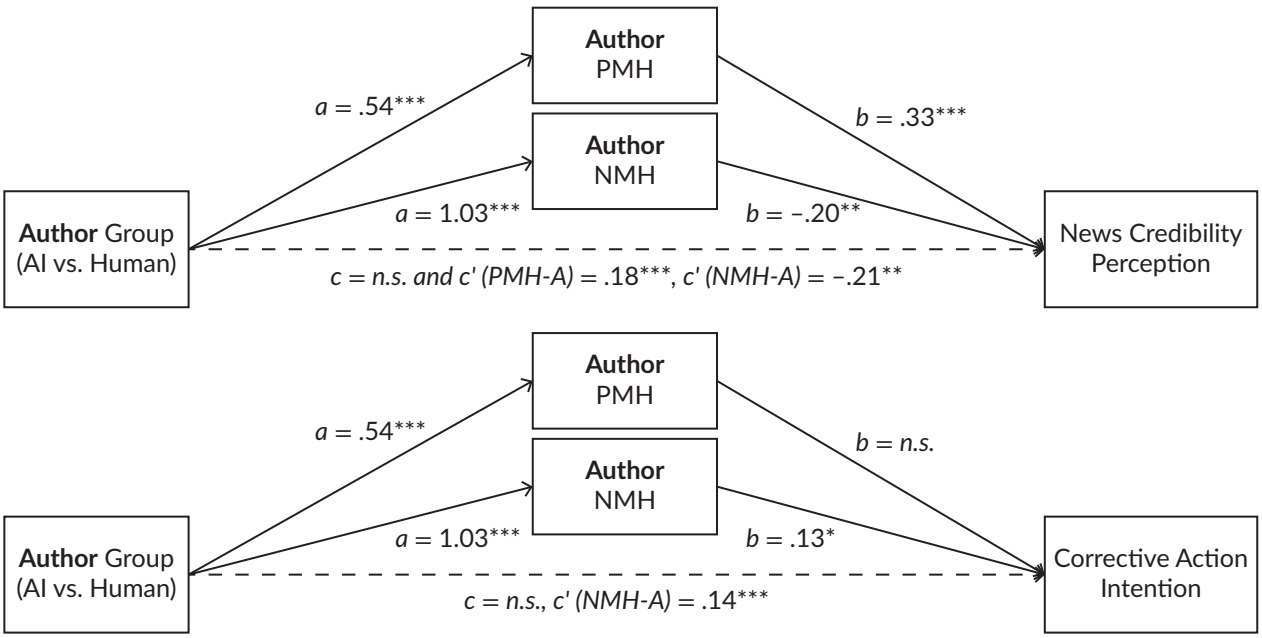


Figure 3. Simple mediation of AI (vs. human) author agency's effect. Notes: Mediating effects of PMH, NMH; news credibility perception as a dependent variable (top); corrective action intention as a dependent variable (bottom); *** $p < .001$, ** $p < .01$, * $p < .05$; c' = direct effects of agency type on dependent variables; c = total effect of agency type; n.s. = not significant.

4.4. Mediation Analyses Moderated by CO

4.4.1. H4: Moderation on the Effect of Fact-Checker Role-Based AI Agency

For H4a, results suggest that CO does not moderate the indirect or direct effect of AI (vs. human) fact-checker agency on perceived news credibility. For H4b, results showed a significant moderated mediation effect through PMH-C (effect = .12, *SE* = .05, 95% CI [.06, .19]). For 90.1% of the participants who have CO > 2.72, higher CO strengthens the mediation effect on corrective action intention (Figure 4a). However, the indirect effect through NMH-C was not significant. For the conditional direct effects on corrective action intention, AI had a significant disadvantage to human fact-checker for inducing corrective action intention at lower levels of CO, however, this effect became non-significant at higher levels of CO (for 68.9% of the participants who have CO > 4.32; Figure 4b).

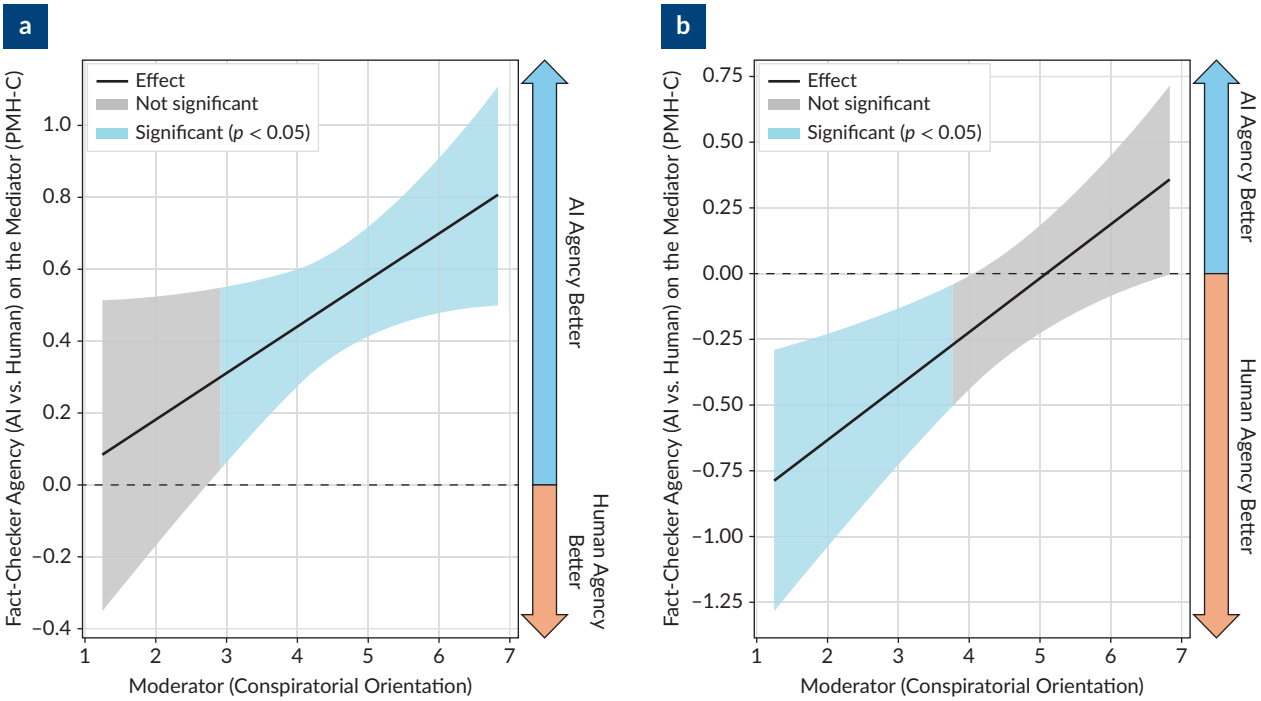


Figure 4. Effects of AI vs. human fact-checker on PMH-C (a) and corrective action intention (b).

4.4.2. H5: Moderation on the Effect of Author Role-Based AI Agency

For H5a and H5b, we used Model 8 again to test if CO moderates the indirect and direct effects examined by H3a and H3b. Results show that while there were some marginal trends, no significant mediation or moderation effects were found for H5b and H6b, suggesting that the proposed mechanisms did not hold for AI authorship's indirect and direct effects.

5. Discussion

5.1. Role-Specific and Cross-Role Effects on Machine Heuristics and Dependent Variables

AI agency and machine heuristics' relationship is associated with the specific role that AI is playing. In line with prior studies, we find that AI author influences author-role-based machine heuristics (Cloudy et al., 2023; Craig & Choi, 2024; Molina & Sundar, 2024; S. Wang & Huang, 2024), and AI fact-checker has an effect on the fact-checker-role based machine heuristics (Banas et al., 2022; Moon & Kahlor, 2025; Tulin et al., 2024; S. Wang, 2021). Our results from RQ1, however, extend existing results by illustrating the possibility that different role-based agencies have an interaction effect in activating PMH of the AI author agency.

The result of this significant interaction can be interpreted together with AI fact-checker agency's activation of NMH-A (NMH for AI's author role). Potentially, when AI appears both as a debunking fact-checker and author, participants' prior negative belief about AI as the fact-checker (NMH-C) overshadows their negative views on AI as the author (NMH-A), making their prior positive beliefs on AI as the author (PMH-A) more salient. However, we acknowledge that the AI fact-checker–AI author combination is currently uncommon for online news interfaces. Nonetheless, with AI's dual roles becoming increasingly more prominent in both fact-checking and news production, the possibility of encountering such circumstances exists in the future.

In line with prior research, we identified significant effects of AI fact-checker-role-based agency on behavioral intentions against fake news (Bhandari et al., 2021; Moon et al., 2022), which in our case is corrective action intention. However, our results showed that AI agency (either fact-checker or author role) is not associated with different levels of news credibility perception compared to when these roles are played by humans. This result does not surprise us given marginal and situational results as summarized by Graefe and Bohlken (2020) and Wang and Huang's (2024) meta-analyses.

5.2. Mediation Effects of Distinct Role-Based Machine Heuristics

Our mediation analysis shows that the AI agency's advantage or disadvantage compared to human agency in its relationship with news credibility perception and corrective action intention is contingent upon two things. The first is that they are dependent on the activation of a corresponding machine heuristics: by comparing results from the MANCOVA and the mediation analyses, we noticed that activation of machine heuristics is critical in determining whether AI agency created any difference from human agency. Although the MANCOVA result does not support AI agency's direct relationship with the perceptual dependent variable (news credibility perception), there are significant indirect, mediated relationships when author-role-based machine heuristics are activated.

Secondly, our results support the idea that author-role-based machine heuristics have more pervasive mediating effects on news credibility and corrective action, while fact-checking-role-based machine heuristics target intentional outcomes more specifically. This is also in accordance with prior studies along separate lines, but our results provide a comparison when agencies of both roles are disclosed on the news interface: Mediation effect exists for news credibility perception when either one of the author-based machine heuristics are activated (Figure 3, top), not when fact-checker-role-based machine heuristics are (Figure 2, top). For corrective action intention, AI agency is associated with higher intentions

than humans when the positive fact-checker-role-based machine heuristics is activated and when negative author-role-based machine heuristics is activated.

5.3. Debunking Conspiracy Theory News: Does AI Agents Matter?

In the case of debunking conspiracy theory news, is AI (vs. human) fact-checker agency associated with more corrective action intention? Our findings suggest that the answer depends on CO levels. The results from the MANCOVA illustrate that news-specific prior beliefs, such as CO in the context of conspiracy theory news, moderate the activation of individuals' prior beliefs about AI as a more precise, error-free, accurate, and objective news fact-checker (in our case, PMH-C difference between AI and human agents). As a continuation of this finding, moderated mediation analyses show that CO emerged as a significant moderator, influencing both the effect of AI vs. human fact-checking agency on corrective action intention and the mediation of such effect through PMH-C.

From the perspective of motivated reasoning, this effect is not surprising, as prior studies have shown that existing beliefs, political inclinations, or ideological orientation (Walter et al., 2020) moderates the effect of misinformation fact-checking. They also interact in the specific domain of AI vs. human fact-checking agency moderating their comparative relationship with misinformation debunking outcomes, such as hostile media effects across partisan lines (Cloudy et al., 2023), or preferences on in-group over out-group fake news (Moon et al., 2023; Moon & Kahlor, 2025). Fact-checkers labeled as "AI" are found to be perceived as "apolitical" compared to human expert fact-checkers, and therefore induce less mistrust against the fact-checking message caused by partisan or ideological preferences (Chung et al., 2024).

However, different from partisanship or hostile media effects, the moderating effect of CO in this study may not have stemmed from an identity-protection motivation (Kahan, 2015; Moon & Kahlor, 2025). If it was, then the analytical focus would be CO's negative relationship with the activation of positive beliefs (PMH) about both AI (non-significant negative correlation) and human (Pearson correlation $r = -.13$, $p < .05$) debunkers. In the current study, individuals with high CO levels are not conceptualized to share a "conspiracy-theory-lover" identity, but rather a common distrust in powerful entities and disgust of power imbalance (Kim & Lee, 2024). A distinctive feature of CO to partisanship is that it is not context dependent. Rather, it represents a long-standing inclination toward skepticism about human motives and intentions. Therefore, in the current study, such skepticism, rather than an identity-protection motivation against any debunkers of conspiracy theory news, was conceptualized and examined as the motivator of differential evaluation of the AI (vs. human) fact-checker agency.

Current results support this idea. We witness a stronger activation of PMH-C by AI (vs. human) fact-checker agency among individuals with high levels of CO (Figure 4a). Because the higher the CO, the less PMH-C (good qualities of a fact-checker) was attributed to human fact-checker: one who potentially holds certain governmentally or organizationally imposed fact-checking agenda. Conversely, individuals with lower levels of CO do not view AI fact-checker agency (vs. human) as more qualified. Moreover, similar Moon and Kahlor's (2025) findings, when the author-based agency is controlled (in the mediation models), our results indicate that AI fact-checking agency (vs. human) is associated with a poor fact-checking result (less corrective action intention in Figure 4b), but only for individuals with lower levels of CO. As CO increases, AI (vs. human) fact-checker agency's lower direct association with correct action intention ceased to be

significant at CO = 4.32. Taking these results together, it is plausible that higher CO activates an AI-fact-checker-centered PMH, therefore activating a positive mediation for AI fact-checker's and a comparatively stronger association with corrective action intentions than human fact-checkers.

While prior research has largely focused on the activation of machine heuristics to explain responses to AI versus human fact-checking, our study extends this framework by exploring how fundamental psychological traits like CO shape the activation of machine heuristics. Specifically, our findings suggest that CO can influence whether individuals apply machine heuristics to AI or human agents. This indicates that the application of machine heuristics—whether positive or negative—is not exclusively linked to AI. Instead, individuals may attribute machine-like characteristics (e.g., objectivity, neutrality) to human agents if they view humans as more competent in certain roles.

5.4. Practical Implications

Our findings emphasize the importance of considering CO when designing fact-checking interventions. For individuals with high CO, AI fact-checkers are perceived as more objective and neutral, making them a more effective tool for promoting corrective action intentions. In contrast, human fact-checkers may be more trusted by those with lower CO who value relational cues and nuanced judgment. Platforms should consider segmenting audiences by CO levels to tailor interventions, using AI fact-checkers for those with high CO and human fact-checkers for others.

Moreover, platforms should adopt public segmentation strategies to address high CO individuals who may be more susceptible to conspiracy theories but would trust AI agency more than human. Insights from this study suggest that interventions based on AI's neutrality could be more effective for these users, especially in environments where conspiracy theories are rampant. Delivering fact-checking content through AI might reduce the resistance these users have toward corrective messages and mitigate the spread of misinformation, ultimately fostering a more informed and engaged user base.

Our findings also show that PMH-C is activated for users with high CO, especially when AI is used as a fact-checker. News and social media platforms can leverage this by incorporating AI-driven fact-checking interventions that resonate with users' preferences for neutrality and objectivity. At the same time, human-based fact-checking can be better suited for addressing users with lower CO who are more likely to engage with human-authored content. This adaptive approach to messaging can help increase engagement with fact-checked content and promote corrective actions, ultimately enhancing the credibility of news sources and reducing the spread of misinformation.

5.5. Limitations

This study has several limitations. First, the sample was skewed toward a younger population, with limited representation of older participants. This may have influenced responses to AI and human fact-checking agencies, as younger individuals may engage differently with technology. Future studies should aim for a more balanced demographic representation to assess how age influences these perceptions.

Second, while we explored two-way interactions between agency type (AI vs. human) and CO, more complex interactions, such as three-way interactions involving fact-checker agency, author agency, and CO, were not investigated. Exploring these interactions could offer deeper insights, though interpreting such models would present significant challenges.

Lastly, factors like the third-person effect or social desirability bias, which can influence corrective actions, were not examined in this study. Incorporating these factors in future research could provide a more comprehensive understanding of the drivers behind corrective behavior, particularly in the context of AI and human fact-checking agencies.

Acknowledgments

We thank the three anonymous reviewers for their insightful feedback and constructive suggestions, which greatly improved the quality of this manuscript. We are also grateful to the thematic issue editors for their guidance throughout the review process.

Funding

This work was supported in part by the Beijing Major Science and Technology Project under Contract No. Z231100007423015e.

Conflict of Interests

The authors declare no conflict of interests.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Banas, J. A., Palomares, N. A., Richards, A. S., Keating, D. M., Joyce, N., & Rains, S. A. (2022). When machine and bandwagon heuristics compete: Understanding users' response to conflicting AI and crowdsourced fact-checking. *Human Communication Research*, *48*(3), 430–461.

Bhandari, A., Ozanne, M., Bazarova, N. N., & DiFranzo, D. (2021). Do you care who flagged this post? Effects of moderator visibility on bystander behavior. *Journal of Computer-Mediated Communication*, *26*(5), 284–300.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's reducing a human being to a percentage: Perceptions of justice in algorithmic decisions. In R. Mandryk & M. Hancock (Eds.), *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Paper 337). ACM. https://doi.org/10.1145/3173574.3173951

Chae, J. H., & Tewksbury, D. (2024). Perceiving AI intervention does not compromise the persuasive effect of fact-checking. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448241286881

Chung, M., Moon, W.-K., & Jones-Jang, S. M. (2024). AI as an apolitical referee: Using alternative sources to decrease partisan biases in the processing of fact-checking messages. *Digital Journalism*, *12*(10), 1548–1569. https://doi.org/10.1080/21670811.2023.2254820

Cloudy, J., Banks, J., & Bowman, N. D. (2023). The Str(AI)ght Scoop: Artificial intelligence cues reduce perceptions of hostile media bias. *Digital Journalism*, *11*(9), 1577–1596. https://doi.org/10.1080/21670811.2021.1969974

Craig, M. J., & Choi, M. (2024). The role of affective and cognitive involvement in the mitigating effects of AI source cues on hostile media bias. *Telematics and Informatics*, *88*, Article 102097. https://doi.org/10.1016/j.tele.2024.102097

DeVerna, M. R., Yan, H. Y., Yang, K.-C., & Menczer, F. (2024). *Fact-checking information from large language models can decrease headline discernment*. arXiv. http://arxiv.org/abs/2308.10800

Federico, C. M. (2022). The complex relationship between conspiracy belief and the politics of social change. *Current Opinion in Psychology*, *47*, Article 101354. https://doi.org/10.1016/j.copsyc.2022.101354

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly*, *77*(3), 515–540. https://doi.org/10.1177/107769900007700304

Frenken, M., & Imhoff, R. (2023). Don't trust anybody: Conspiracy mentality and the detection of facial trustworthiness cues. *Applied Cognitive Psychology*, *37*(2), 256–265. https://doi.org/10.1002/acp.3955

Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication*, *63*(4), 617–637. https://doi.org/10.1111/jcom.12038

Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems*, *38*(1), 196–221. https://doi.org/10.1080/07421222.2021.1870389

Graefe, A., & Bohlken, N. (2020). Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news. *Media and Communication*, *8*(3), 50–59. https://doi.org/10.17645/mac.v8i3.3019

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610. https://doi.org/10.1177/1464884916641269

Hameleers, M., & van der Meer, T. G. (2019). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research*, *47*(2), 227–250. https://doi.org/10.1177/0093650218819671

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research*, *50*(1), 1–22. https://doi.org/10.1080/00273171.2014.962683

Horne, B. D., Nevo, D., Adali, S., Manikonda, L., & Arrington, C. (2020). Tailoring heuristics and timing AI interventions for supporting news veracity assessments. *Computers in Human Behavior Reports*, *2*, Article 100043. https://doi.org/10.1016/j.chbr.2020.100043

Imhoff, R., & Bruder, M. (2014). Speaking (un–)truth to power: Conspiracy mentality as a generalised political attitude. *European Journal of Personality*, *28*(1), 25–43. https://doi.org/10.1002/per.1930

Jahanbakhsh, F., Katsis, Y., Wang, D., Popa, L., & Muller, M. (2023). Exploring the use of personalized AI for identifying misinformation on social media. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Article 105). ACM. https://doi.org/10.1145/3544548.3581219

Jennings, J., & Stroud, N. J. (2023). Asymmetric adjustment: Partisanship and correcting misinformation on Facebook. *New Media & Society*, *25*(7), 1501–1521. https://doi.org/10.1177/14614448211021720

Jia, H., Appelman, A., Wu, M., & Bien-Aimé, S. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans*, *2*(2), Article 100093.

Jia, H., & Luo, X. (2023). I wear a mask for my country: Conspiracy theories, nationalism, and intention to adopt Covid-19 prevention behaviors at the later stage of pandemic control in China. *Health Communication*, *38*(3), 543–551. https://doi.org/10.1080/10410236.2021.1958982

Kahan, D. M. (2015). The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. In R. Scott, M. Buchmann., & S. Kosslyn. (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–16). Wiley. https://doi.org/10.1002/9781118900772

Kim, J.-N., & Lee, S. (2024). Conceptualizing conspiratorial thinking: Explicating public conspiracism for effective debiasing strategy. *American Behavioral Scientist*, *68*(10), 1366–1394. https://doi.org/10.1177/00027642231175637

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117.

Lee, J., & Bissell, K. (2024). User agency–based versus machine agency–based misinformation interventions: The effects of commenting and AI fact-checking labeling on attitudes toward the Covid-19 vaccination. *New Media & Society*, *26*(12), 6817–6837. https://doi.org/10.1177/14614448231163228

Liu, X., Qi, L., Wang, L., & Metzger, M. J. (2023). Checking the fact-checkers: The role of source type, perceived credibility, and individual differences in fact-checking effectiveness. *Communication Research*. Advance online publiaction. https://doi.org/10.1177/00936502231206419

Miller, J. M., Saunders, K. L., & Farhart, C. E. (2016). Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust. *American Journal of Political Science*, *60*(4), 824–844. https://doi.org/10.1111/ajps.12234

Molina, M. D., & Sundar, S. S. (2024). Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society*, *26*(6), 3638–3656. https://doi.org/10.1177/14614448221103534

Moon, W.-K., Atkinson, L., Kahlor, L. A., Yun, C., & Son, H. (2022). US political partisanship and Covid-19: Risk information seeking and prevention behaviors. *Health Communication*, *37*(13), 1671–1681.

Moon, W.-K., Chung, M., & Jones-Jang, S. M. (2023). How can we fight partisan biases in the Covid-19 pandemic? AI source labels on fact-checking messages reduce motivated reasoning. *Mass Communication and Society*, *26*(4), 646–670. https://doi.org/10.1080/15205436.2022.2097926

Moon, W.-K., & Kahlor, L. A. (2025). Fact-checking in the age of AI: Reducing biases with non-human information sources. *Technology in Society*, *80*, Article 102760. https://doi.org/10.1016/j.techsoc.2024.102760

National Bureau of Statistics of China. (2018). *China statistical year book 2018*. China Statistics Press. https://www.stats.gov.cn/sj/ndsj/2018/indexeh.htm

Pareek, S., van Berkel, N., Velloso, E., & Goncalves, J. (2024). Effect of explanation conceptualisations on trust in AI-assisted credibility assessment. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), Article 383. https://dl.acm.org/doi/abs/10.1145/3686922

Rae, I. (2024). The effects of perceived AI use on content perceptions. In F. Floyd Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 978). ACM. https://doi.org/10.1145/3613904.3642076

Romer, D., & Jamieson, K. H. (2022). Conspiratorial thinking as a precursor to opposition to Covid-19 vaccination in the US: A multi-year study from 2018 to 2021. *Scientific Reports*, *12*(1), Article 18632.

Sundar, S. S. (2008). *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In S. Brewster & G. Fitzpatrick (Eds.), *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Article 538). ACM. https://doi.org/10.1145/3290605.3300768

Sutton, R. M., & Douglas, K. M. (2020). Conspiracy theories and the conspiracy mindset: Implications for political ideology. *Current Opinion in Behavioral Sciences*, *34*, 118–122. https://doi.org/10.1016/j.cobeha.2020.02.015

Talwar, S., Dhir, A., Singh, D., Virk, G. S., & Salo, J. (2020). Sharing of fake news on social media: Application of the honeycomb framework and the third-person effect hypothesis. *Journal of Retailing and Consumer Services*, *57*, Article 102197. https://doi.org/10.1016/j.jretconser.2020.102197

Tam, L., & Lee, H. (2024). From conspiracy orientation to conspiracy attribution: The effects of institutional trust and demographic differences. *American Behavioral Scientist*, *68*(10), 1395–1411. https://doi.org/10.1177/00027642231174330

Thurman, N., Dörr, K., & Kunert, J. (2017). When reporters get hands-on with robo-writing: Professionals consider automated journalism's capabilities and consequences. *Digital Journalism*, *5*(10), 1240–1259. https://doi.org/10.1080/21670811.2017.1289819

Tulin, M., Hameleers, M., de Vreese, C., Opgenhaffen, M., & Wouters, F. (2024). Beyond belief correction: Effects of the truth sandwich on perceptions of fact-checkers and verification intentions. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2024.2311311

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, *39*(5), 621–645. https://doi.org/10.1177/1075547017731776

Waddell, T. F. (2019). Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism & Mass Communication Quarterly*, *96*(1), 82–100. https://doi.org/10.1177/1077699018815891

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Wang, S. (2021). Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital Journalism*, *9*(1), 64–83. https://doi.org/10.1080/21670811.2020.1851279

Wang, S., & Huang, G. (2024). The impact of machine authorship on news audience perceptions: A meta-analysis of experimental studies. *Communication Research*, *51*(7), 815–842. https://doi.org/10.1177/00936502241229794

Wang, Y. (2021). Debunking misinformation about genetically modified food safety on social media: Can heuristic cues mitigate biased assimilation? *Science Communication*, *43*(4), 460–485. https://doi.org/10.1177/10755470211022024

Wischnewski, M., & Krämer, N. (2022). Can AI reduce motivated reasoning in news consumption? Investigating the role of attitudes towards AI and prior-opinion in shaping trust perceptions of news. In S. Schlobach, M. Pérez-Ortiz, & M. Tielman (Eds.), *HHAI2022: Augmenting human intellect* (pp. 184–198). IOS Press. https://doi.org/10.3233/FAIA220198

Wu, S., Tandoc, E. C., Jr., & Salmon, C. T. (2019). Journalism reconfigured: Assessing human–machine relations and the autonomous power of automation in news production. *Journalism Studies*, *20*(10), 1440–1457. https://doi.org/10.1080/1461670X.2018.1521299

Xu, D., Fan, S., & Kankanhalli, M. (2023). Combating misinformation in the era of generative AI models. In A. El Saddik, T. Mei, & R. Cucchiara (Eds.), *MM '23: Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9291–9298). ACM. https://doi.org/10.1145/3581783.3612704

Zhu, Y., Fitzpatrick, M. A., & Bowen, S. A. (2024). Factors related to compliance with CDC Covid-19 guidelines: Media use, partisan identity, science knowledge, and risk assessment. *Western Journal of Communication*, *88*(3), 567–594. https://doi.org/10.1080/10570314.2023.2219239

Zhu, Y., Xu, J., Zhang, R., Lan, D., & Jiang, Y. (2024). Prior attitude, individualism and perceived scientists' expertise: Exploring motivated reasoning of scientific information about HIV risks of homosexuals in China. *Journal of Media Psychology: Theories, Methods, and Applications*. Advance online publication. https://doi.org/10.1027/1864-1105/a000437

About the Authors

Duo Lan (PhD, Beijing Normal University) is an assistant professor at Beijing University of Posts and Telecommunications, PR China. Her research interests include media technology effects, transcultural communication, and film studies.

Yicheng Zhu (PhD, University of South Carolina, USA) is an associate professor at Beijing Normal University, PR China. His research focuses on international strategic communication, social identities, and media technologies.

Meiyu Liu is a graduate student at Beijing Normal University. She is committed to studying the behavioral logic of interaction between humans and intelligent robots, keen to explore the changes in media credibility in the era of intelligence, and seeking crisis communication solutions.

Chuge He is a graduate student at Beijing Normal University. Her research interests include intelligent communication, new media, and risk communication.

Media and Communication
2025 • Volume 13 • Article 9543
https://doi.org/10.17645/mac.9543

ARTICLE

Open Access Journal

SMART 2.0: Social Media Analytics and Reporting Tool Applied to Misinformation Tracking

Mahmoud Mousa Hamad, Gopichandh Danala, Wolfgang Jentner, and David Ebert

Data Institute for Societal Challenges, University of Oklahoma, USA

Correspondence: Mahmoud Mousa Hamad (mmh@ou.edu)

Submitted: 30 October 2024 **Accepted:** 5 February 2025 **Published:** 3 April 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

The rapid proliferation of social media has created new data stemming from users' thoughts, feelings, and interests. However, this unprecedented growth has led to the widespread dissemination of misinformation—deliberately or inadvertently false content that can trigger dangerous societal ramifications. Visual analytics combines advanced data analytics and interactive visualizations to explore data and mine insights. This article introduces the Social Media Analytics and Reporting Tool (SMART) 2.0, detailing its application in tracking misinformation on social media. An updated version of its predecessor, SMART 2.0 enables analysts to conduct real-time surveillance of social media content along with complementary data streams, including weather patterns, traffic conditions, and emergency service reports. SMART 2.0 offers enhanced capabilities like map-based, interactive, and AI-powered features that enable researchers to visualize and understand situational changes by assessing public social posts and comments. As a misinformation classification and tracking case study, we collected public, geo-tagged tweets from multiple cities in the UK during the 2024 riots. We showcased the effectiveness of SMART 2.0's misinformation detection and tracking capabilities. Our findings show that SMART 2.0 effectively tracks and classifies misinformation using a human-in-the-loop approach.

Keywords

machine learning; misinformation; SMART; SMART 2.0; social media; surveillance; visual analytics

1. Introduction

This article adheres to the following definitions: misinformation is misleading information shared without intent to deceive, whereas disinformation is shared with deliberate intent to mislead (Treen et al., 2020).

Consistent use of these terms ensures clarity when discussing social media data analysis. In the context of social media, users may unknowingly share disinformation, making it difficult to distinguish between those spreading false information unintentionally versus deliberately.

Individuals spread misinformation for a specific goal, such as for a political agenda, or unknowingly, which can have many dangerous ramifications. Misinformation can have serious consequences, as seen in the false claim that a Muslim killed three children in Southport, England, which led to a mosque being firebombed in Northern Ireland ("Ards mosque community," 2024). Misinformation, such as climate change denial spread through social media, can also hinder progress by creating public confusion, fostering misplaced criticism, and fueling protests against policies meant to combat societal issues (Treen et al., 2020).

The Social Media Analytics and Reporting Tool (SMART), a research tool that uses social media, was developed for situational awareness, event monitoring, and public sentiment analysis (Snyder, Karimzadeh, Stober, & Ebert, 2019). This article introduces SMART 2.0, the updated version of SMART, which helps stakeholders, researchers, and community partners visualize and analyze social media data. We focus on SMART 2.0's machine learning-powered capabilities to track misinformation. We explore its features and applications for media research, showcasing a case study demonstrating its effectiveness in monitoring misinformation. Both SMART and SMART 2.0 are individual systems that we have developed over the years. This article aims to describe SMART and introduce SMART 2.0, highlighting its enhancements and new features.

SMART, developed with a team that includes members at Purdue University and Penn State, has been widely used since 2013. Initially employed by the US Coast Guard and later by many public safety agencies, it played a key role in supporting public safety during college football games, Fleet Week, and presidential inaugurations. Its use expanded to other agencies, including local law enforcement, intelligence centers, and organizations like the American Red Cross, proving its versatility in planned and emergent situations (Snyder, Karimzadeh, Chen, & Ebert, 2019; Snyder, Karimzadeh, Stober, & Ebert, 2019). SMART's integration into the North Atlantic Treaty Organization's network for alerting and managing public safety and resilience (NATO REACT) project led to the creation of SMART 2.0, further demonstrating its global applicability, helping to monitor social media for misinformation and enhancing the crisis response (Illia State University, n.d.). The objectives of the NATO REACT research project were to build upon SMART using interactive machine learning for real-time human–computer collaborative decision-making, multicultural and multilingual support, human-in-the-loop misinformation identification and information filtering, and information fusing of social and environmental sensing data.

SMART 2.0 extends and improves SMART by including new features and improved performance. These new features and improvements include the following: machine-learning language translation of the user interface, social media data and all other existing visualization tools, misinformation detection and filtering, environmental data sensing and tracking, and improved performance for social media data fetching, as well as an improved software architecture of the system.

2. Related Work and Literature Review

2.1. Related Work

Nowadays, social media has become part of everyday life. It has also become a critical resource for situational awareness and event monitoring. In computational social science, gathering and analyzing data is integral to research that uses social media data to answer specific questions. The ubiquity of social media usage has led to a rise in the number of data services, tools, and analytics platforms for academic research and enterprise usage (Batrinca & Treleaven, 2015).

However, several challenges accompany social media data collection and analytics. These challenges include data scraping, cleansing, holistic data sources, protection, analytics, and visualization (Batrinca & Treleaven, 2015).

Despite these challenges, we recognize the value of the insights that social media data might provide, especially in a real-time setting. The social media revolution has created unprecedented opportunities to assess public responses and critiques of a multitude of social issues and events; this data can be provided and evaluated in real-time. We can use data science practices to harness this information to identify trustworthy information, reduce false claims, and take actionable steps for research and practical purposes, such as real-time surveillance and monitoring.

Therefore, several tools and platforms have been developed to address the challenges of social media data analysis, such as: Netlytic, which analyzes social networks and summarizes large text volumes; Gephi, an open-source graph visualization and analysis software; and NodeXL, an open-source template for Microsoft Excel for network analysis (Bastian et al., 2009; Hansen et al., 2010; Quan-Haase & Sloan, 2022).

2.2. Misinformation Theory

Social media has significantly accelerated the spread of misinformation, making it easier for false information to reach a broad audience (Treen et al., 2020). Research has shown that health misinformation can spread quickly through social media (Vosoughi et al., 2018). In 2013, a tweet from a hacked *Associated Press* account falsely reported an injury to then-President Obama, causing a $130 billion drop in stock value within minutes (Rapoza, 2017). This highlights the influence of social media on the rapid dissemination of misinformation.

Recent research has focused on developing methods to track and analyze misinformation on social media. One example is bot detection, used to identify automated accounts that spread misinformation (Ferrara et al., 2016). Network analysis is also used for examining the structure of information diffusion networks to identify patterns of misinformation spread (Shao et al., 2018). Furthermore, machine learning and natural language processing (NLP) techniques are used to classify and detect misinformation in text content (Shu et al., 2017). Researchers have also analyzed fact-checking integration, where automated systems are combined with human fact-checkers to verify information (Hassan et al., 2017). To add to this recent research, this study presents a novel approach for using SMART 2.0 to conduct surveillance of social media activity to identify misinformation in tweets.

3. Methodology

The methodology of this article encompasses three key components: data collection, analysis techniques, and user interface design. This section details our approach to gathering social media data despite recent restrictions on application programming interfaces (APIs). We describe here the machine learning and NLP techniques employed for data analysis and misinformation detection, and outline the interactive visualization features developed to support real-time decision-making. Our methodology prioritizes both technical robustness and user accessibility, with particular emphasis on overcoming data access limitations through innovative solutions like geolocation prediction and interactive machine learning models.

3.1. Data Collection

In the past, accessing real-time data from platforms like X (formerly Twitter) and Instagram was easy and free through APIs, but recent changes have restricted access. X's free API access was eliminated, and only costly paid plans are now available for meaningful data (Calma, 2023; Stokel-Walker, 2024). Similarly, Meta has limited data access on Facebook and Instagram to public profiles with over 25,000 followers, and it has a lengthy application process (Ryan-Mosley, 2023). Although scraping data using bots can be effective for small datasets, it is becoming increasingly difficult due to bot detection mechanisms, making this method unsustainable and arguably unethical (Chiapponi et al., 2022).

SMART 2.0 relies on geo-tagging to display data on the map. While SMART 2.0 uses methods to scrape Instagram and X for geo-specific posts, the limitations of scraping led to a reliance on paid APIs like X's. To overcome the shortage of geo-tagged data, SMART 2.0 is powered by a machine learning-powered geolocation prediction tool, which predicts the location of non-geo-tagged posts using deep learning models trained on geo-tagged data (Snyder, Karimzadeh, Chen, & Ebert, 2019). This increases the number of geo-tagged data available and increases SMART 2.0's value.

3.2. Analysis Techniques

SMART 2.0 handles multiple data types—textual, geographical, and environmental—each requiring specific analysis methods. SMART 2.0 offers classification, searching, and filtering tools for text-based data. Users can define categories using keywords, and SMART 2.0 supports complex criteria for filtering. Including deep learning features, SMART 2.0 can refine searches by removing irrelevant results, with machine learning models adapting to user preferences (Snyder et al., 2020). SMART 2.0 applies NLP techniques, such as latent Dirichlet allocation (LDA), to extract topics and classify data automatically (Snyder, Karimzadeh, Stober, & Ebert, 2019).

To identify and classify misleading information, SMART 2.0 utilizes interactive machine learning. Using a human-in-the-loop approach, the model corrects misclassifications by updating the model in real-time and it includes a built-in misinformation detection feature that classifies data as misinformation or not, using machine learning models trained on thousands of labeled tweets. This model classifies data points and can be interactively updated in real-time, allowing efficient data filtering, especially during emergencies, where vast amounts of information must be sifted for accuracy.

The misinformation models were initially pre-trained on a diverse dataset of over 10,200 tweets, including categories such as weather, news, traffic, Covid-19, security, trending, and random topics. This initial pool of

tweets was manually labeled to validate the presence of misinformation, with care being taken to ensure a relatively balanced representation of both classes. The data was preprocessed using stop word removal, lemmatization, stemming, and then vectorization, which converts words or tokens into numbers that a computer can understand. Then, a random split of 80–20 was used to divide the data into training and testing subsets. The training samples were further leveraged to undergo a 5-fold cross-fold validation to train several models and identify the optimal model parameters. Finally, the unseen test data was used to evaluate each model's performance. This process achieved testing classification accuracies between 70–78%, with the passive-aggressive classifier performing best. The best model configuration is saved for future interactive learning.

During the real-time usage of SMART 2.0, users can provide feedback on possible wrongly classified tweets within the system, thereby triggering the lightweight partial training to update the label and train the model weights. This iterative and interactive misinformation training mechanism allows it to learn new patterns and improve its detection capabilities.

By supporting multilingual training data, including English, Italian, and Georgian, SMART 2.0 broadens its application across different languages. Figure 1 shows flowcharts of the machine learning model lifecycle used to implement misinformation classification. The top flowchart in Figure 1 shows how the misinformation classification model was constructed, trained, deployed, run, and updated; and the bottom flowchart illustrates the update process of a tweet's misinformation label by the user.

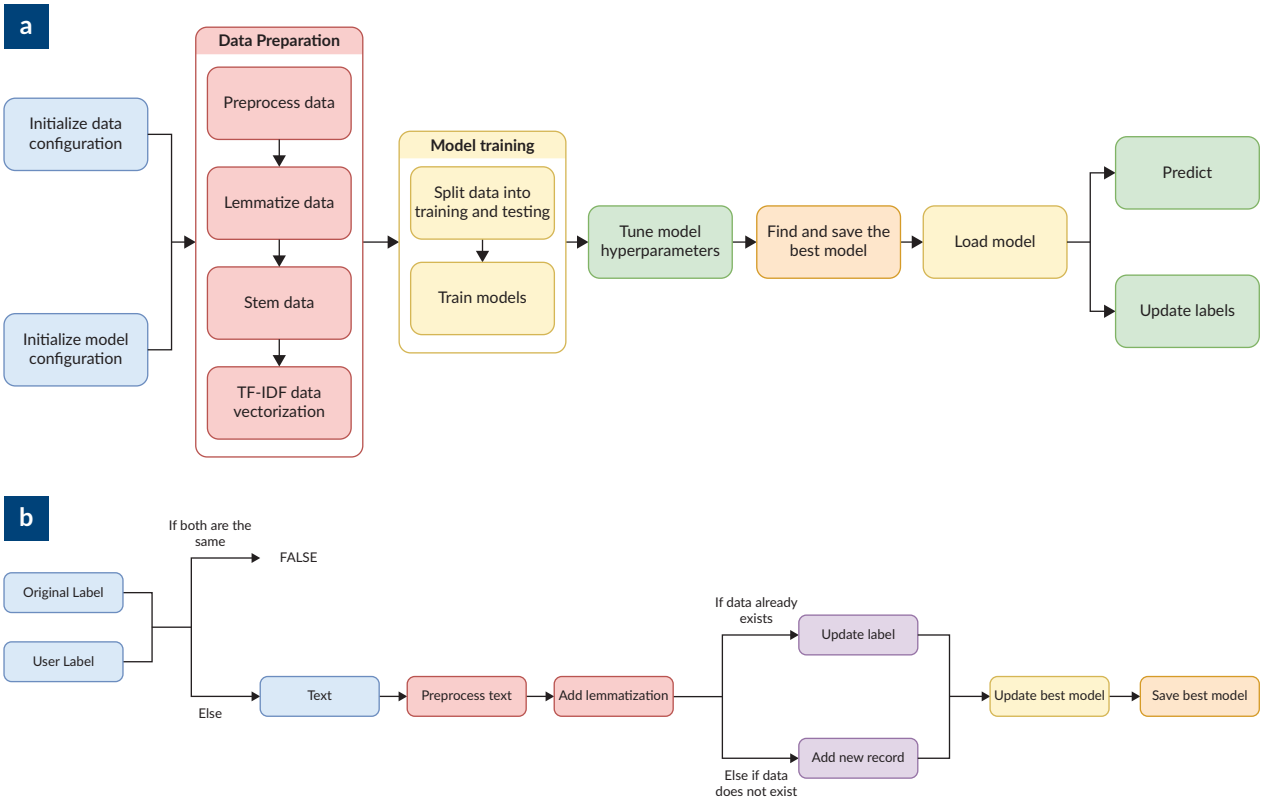


Figure 1. Misinformation flowchart: (a) steps to train, run, and update the misinformation classifier; (b) process of updating the misinformation label of a single tweet. Note: TF-IDF = term frequency–inverse document frequency.

3.3. User Interface

An interactive data visualization and analysis tool, SMART 2.0 enables the understanding of several data types for real-time decision-making as it communicates information using cutting-edge data visualization. From graph-based to map-based visualizations, the SMART 2.0 design allows us to optimally communicate the relevant data to the user. Figure 2 showcases a high-level overview of the user interface.

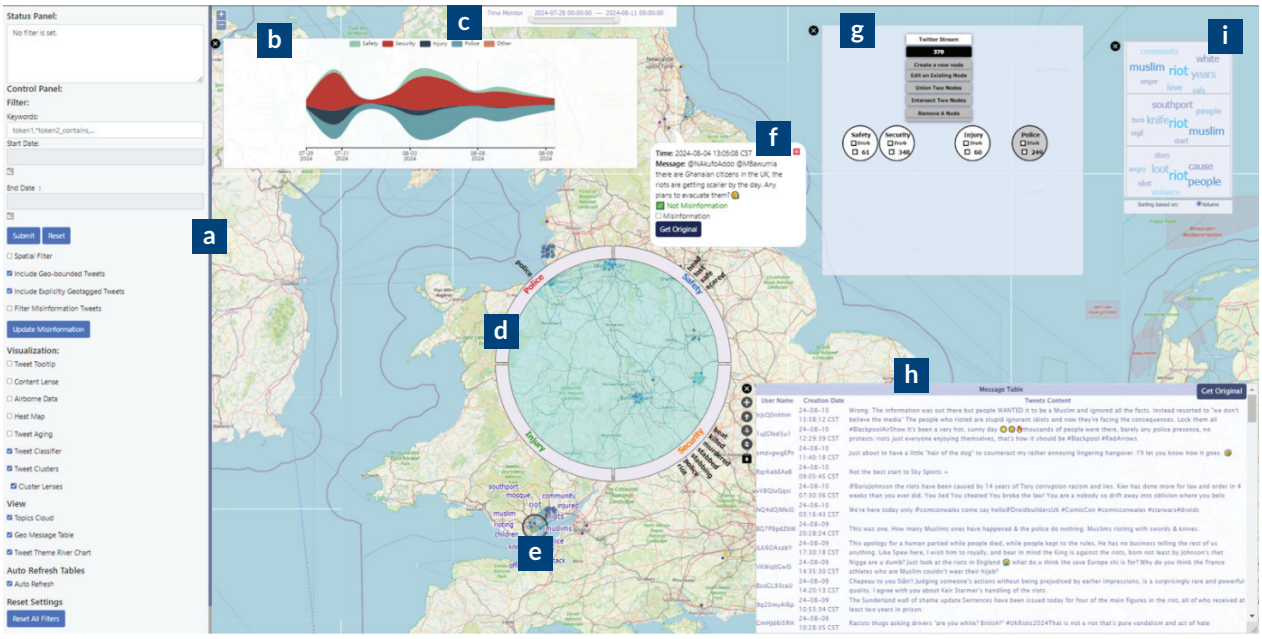


Figure 2. The SMART 2.0 user interface showing an array of panels and features of the toolkit: (a) side panel with controls and filters for various features; (b) theme river chart shows the number of tweets with time for each user-defined class; (c) the time monitor filters tweets based on the creation time; (d) the cluster lens clusters tweets by distance and displays any class-defining keywords present in any cluster under the lens; (e) the content lens displays common keywords in any tweet under the lens as the user hovers over the data; (f) the tweet tooltip displays details about the tweet such as time, message, and misinformation label, and it includes controls for fetching the original (non-translated) text and updating the misinformation label if necessary; (g) the classifier window is the classifier feature of the tool where the user can create, edit, delete, unionize, intersect, filter, and otherwise manage classes that are used elsewhere in the system; (h) the tweet data table shows the tweets displayed on the map and their details and allows the user to switch between the original (non-translated) text and the translated text of the tweets; (i) the topics cloud displays clusters of keywords generated using LDA, where each cluster of keywords represents a topic that exists in the current session, and clicking on a keyword filters the data on the map using the clicked keyword.

Figure 3 is a screenshot from SMART 2.0 showing the user interface for the misinformation classification of tweets.

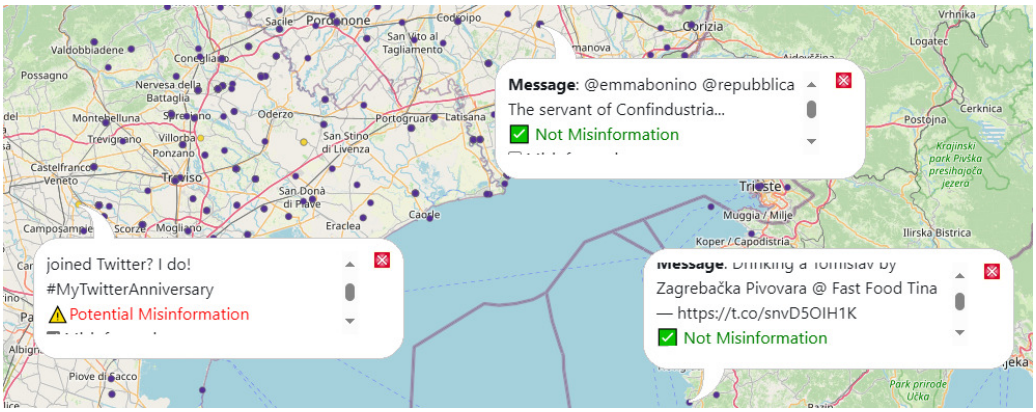


Figure 3. SMART 2.0 screenshot showing a sample of social data records in Venice with their tooltips showing their misinformation classification labels.

3.4. SMART Capabilities

SMART 2.0 enables the user to search for data and filter it using sophisticated full-text search functionality, as in Figure 4(a). A date and time filter is also available and works with keyword filtering. In addition to keyword and date filters, a spatial filter is available, which filters data on the map. The user draws a polygon on the map, and only data inside it is displayed, as seen in Figure 4(b). The misinformation filter allows users to toggle between hiding and showing the data classified as containing misinformation with the data labeled as containing misinformation being yellow, and the dark blue data points on the map being labeled as not containing misinformation.

The tooltip feature allows the user to click on a data point and display its data and metadata, as in Figure 4(c). The content lens displays common tokens among a data group as the user hovers over the map and it integrates with translation capability. The user can toggle back and forth between the original and translated tokens. Figure 4(d) shows the content lens in SMART 2.0. Moreover, the user can add multiple content lenses by freezing a content lens in the desired location. The heat map feature visualizes the spatial distribution of tweets. Figure 4(e) shows that the areas of higher intensity will appear yellow/greenish, and areas of lower intensity will appear blue.

SMART 2.0 features an interactive interface that allows users to search, classify, and filter tweets using "union," "intersection," and "not" filters, with users being able to create or edit classifiers, modify keywords, and apply these changes in real-time. As shown in Figure 4(f), a separate window manages filters, and changes are immediately reflected in the main interface. The "not" filter, displayed in grey, excludes tweets containing specific keywords.

Additionally, SMART 2.0 includes an intentional verb filter, which allows secondary filtering based on verbs that indicate human intent, such as "need," "want," or "attempt." Users can easily remove classifier nodes by selecting a node and clicking the "remove a node" button.

The theme river view is a time-series visualization that shows the number of tweets in different classifiers over the last two hours, providing an intuitive visualization of the temporal evolution of topics through a river metaphor. The theme river view is automatically linked with the classifiers. As shown in Figure 4(g), when the

user creates a new classifier, it shows instantaneously in the theme river view. Users can click on classifiers in the legend to remove them from the theme river.

The tweets cluster view visualizes groups (clusters) of tweets located closely in geographic space. The clusters are visualized using a polygon-based representation, as shown in Figure 4(h). They are zoom-adaptive, which means when zooming in, the clusters split into small-scale ones. This feature aims to enable effective multi-scale exploration.

Figure 4. Visualization features: (a) keywords and date filters; (b) spatial filter; (c) tooltip; (d) content lens; (e) heat map; (f) classifier nodes window; (g) theme river chart; (h) cluster view; (i) cluster lens; (j) topic cloud.

The topic lens feature is a secondary feature in the cluster view and can be enabled by clicking the "cluster lens" option. According to Figure 4(i), the topic lens filters clusters within the lens and visualizes the keywords related to the current classifiers in a radial layout. The user can move the underlying map to investigate regions of interest while the position of the lens is fixed on the screen.

Figure 4(j) shows the LDA topic model window. LDA allows SMART 2.0 to automatically extract topics and define these topics using filter-enabled tokens, aiding in data comprehension.

4. Misinformation Tracking Case Study Using SMART 2.0

4.1. Introducing the Case Study and Its Significance

Our case study will focus on using SMART 2.0 to track and classify misinformation. The events that we studied took place in England between late July and early August 2024. The events primarily occurred in Liverpool, where riots took place, vehicles, shops, and buildings were set ablaze, and individuals were assaulted, harassed, and abused (Frayer, 2024; Otis, 2024). Similar riots took place in dozens of other towns across England, Wales, Scotland, and Northern Ireland (Ahmed, 2024). The riots were the aftermath of a fatal stabbing of three little girls in Southport (Lawless, 2024) that took place during a Taylor Swift-themed dance and yoga workshop in the northern English town on July 29, 2024, and caused the injury of several others (Lawless, 2024).

This case study was chosen because it is recent and relevant. The UK riots came about mainly due to misinformation and a perceived mistrust in legacy media (Frayer, 2024) and are documented to have been instigated by misinformation about the identity and religion of the perpetrator (Syed/London, 2024). While this case study effectively demonstrates SMART 2.0's capabilities, its scope is limited to a specific event. Future studies should explore broader applications, such as public health misinformation or disaster response, to validate the tool's versatility. The fact that misinformation was instigated in the social media realm makes this case study an appropriate choice for studying misinformation using SMART 2.0.

In this case study, we will attempt to use SMART 2.0's visualization, data science, and machine learning features to understand and track any misinformation associated with this event. To understand the riots and the events, we collected tweets from around Liverpool, Manchester, and other towns in the UK. We will use SMART 2.0's interactive misinformation classification and tracking capabilities to showcase its features.

The perpetrator's name and details were initially not disclosed to the public, given that he was a minor. Despite that, posts spread rapidly online, claiming he was a Muslim and an illegal migrant. The spread of this unfounded information led to riots against Muslims and migrants across the UK. Amid the riots and the violence, the court decided to release the name of the perpetrator to calm far-right public opinion and decrease violence against Muslims and other minorities.

The perpetrator's name was Axel Rudakubana, a 17-year-old Chirsitian UK-born teenager of Rwandan descent. He was not Muslim nor a migrant. Despite these facts, misinformation continued spreading across social media that he was Muslim and a migrant and was given the name "Ali al-Shakati" by social media users online without an official source.

4.2. Time Frame and Geographical Scope

The UK riots, which lasted from July 29 to August 5, 2024, were sparked by the stabbing of three young girls in Southport and spread across various towns in England, Wales, and Northern Ireland. Influential social media figures, such as Andrew Tate and Tommy Robinson, fueled the riots by falsely claiming the attacker was a Muslim immigrant, exacerbating tensions and inciting violence. Even after the perpetrator's identity was revealed, far-right narratives persisted, promoting distrust in the media and justifying continued violence. The riots led to the destruction of mosques and businesses and assaults on police, with over 400 people arrested (Syed/London, 2024).

In this case study, we will look at tweets from July 28, 2024, to August 10, 2024, from various towns in the UK, such as Liverpool, London, Cardiff, Leeds, and more. Our goal is to track the misinformation on X during the initial phase of the riots and then see how misinformation might have changed after the perpetrator's identity was revealed.

4.3. Case Study Methodology

4.3.1. Data Collection

SMART 2.0 was set up for this case study by creating a new historical event in the system and then loading the event into a new session. We use SMART 2.0's features to filter data, locate data points, and explore tweets. We also used the following features to study this case: event creation, event loading, keyword filtering, spatial filtering, date filtering, keyword filtering, tooltip, heatmap, and more. We also used SMART 2.0's interactive misinformation detection and classification capabilities to identify and track misinformation.

The system initially loads pre-trained misinformation classification models fine-tuned in previous iterations as the base model. The system is designed to be interactive and continuously improving, featuring a client-server architecture that allows users to provide feedback on the model's classifications in real-time. When users encounter a tweet, they can see the model's classification through a tooltip and correct any misclassifications, which then feed back into the model through partial fitting—an efficient technique that allows the model to learn from new data without complete retraining. This user feedback loop helps improve the model's accuracy over time, though one noted limitation is the need to preserve user-provided labels better when the model undergoes complete retraining. The system also includes practical features like the ability to filter out content classified as misinformation and batch update classifications across multiple posts in a single session.

Our data collection process used a custom script to scrape X for location-based tweets across the UK, focusing on tweets containing keywords related to the Southport incident and subsequent riots. The script collected tweets from July 28 to August 10, 2024, gathering a total of 370 tweets. The dataset for this case study is limited in size (370 tweets), which may not fully capture the breadth of misinformation surrounding an event of this scale. Furthermore, the reliance on geo-tagged tweets introduces biases toward users who enabled location sharing, potentially excluding significant portions of the population. Future work should explore methods to scale SMART 2.0 to larger datasets and reduce biases introduced during data collection. Another limitation is the potential oversaturation of certain hashtags or keywords, which may skew the dataset towards narratives. These limitations should be considered when interpreting the results of our analysis.

4.3.2. Initial Misinformation Classification Results

Figure 5 shows the initial misinformation classification of the tweets in Southport (a) and Liverpool (b). As shown in Figure 5, many of the data points in Liverpool and Southport have been classified as containing misinformation. This represents the initial classification result of the model without any input from the user. Initially, the data in this case study is foreign to the misinformation classification model.

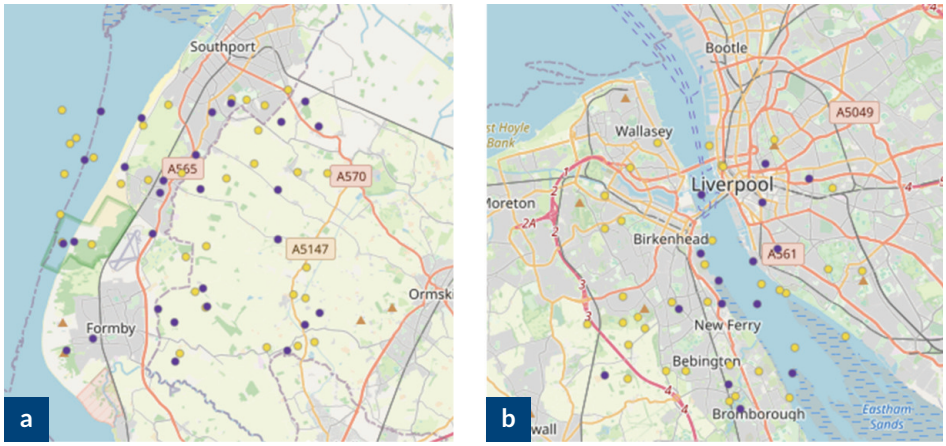


Figure 5. Misinformation-labeled data (yellow) in (a) Southport and (b) Liverpool.

4.3.3. Using Interactive Machine Learning for Refining Classifications

SMART 2.0's misinformation capabilities can be refined to specific events and use cases depending on the user's views and conception of what misinformation entails. Since we have established the base truth from multiple trustworthy and official sources that the perpetrator is not a Muslim and not a migrant, then we searched for and selected tweets that claim that he is Muslim. However, our model mislabeled them as not containing misinformation, so we then corrected them and updated the labels for all the data.

After identifying a few similar mislabeled tweets and correcting the model's predictions, we updated the data labels. After updating the misinformation label of 50% of the tweets, we got the model fine-tuned to the data in the current session. We used the fine-tuned model to update the labels for all the data. We found better accuracy in detecting misinformation in tweets that still asserted that the perpetrator was a Muslim. The initial accuracy of the model without any updates was 53.7%. After updating the misinformation labels of 50% of the tweets (185 tweets), the accuracy increased to 82.7%. We chose to update 50% of the data because it is a midpoint that provides enough data for the model to adapt to the case study. We conducted an experiment to evaluate the following hypothesis: the more the user updates the misinformation classification model, the better its accuracy will be. We will discuss this experiment in the next section of the article.

4.4. Key Findings

The findings regarding misinformation are according to the misinformation labels as classified by the misinformation classifier after updating the labels of 50% of the tweets.

4.4.1. Patterns in Misinformation Spread

We have noticed that much misinformation is concentrated in Liverpool and Newport, near Cardiff, where the riots against Muslims took place. Figure 6a and Figure 6c demonstrate this trend.

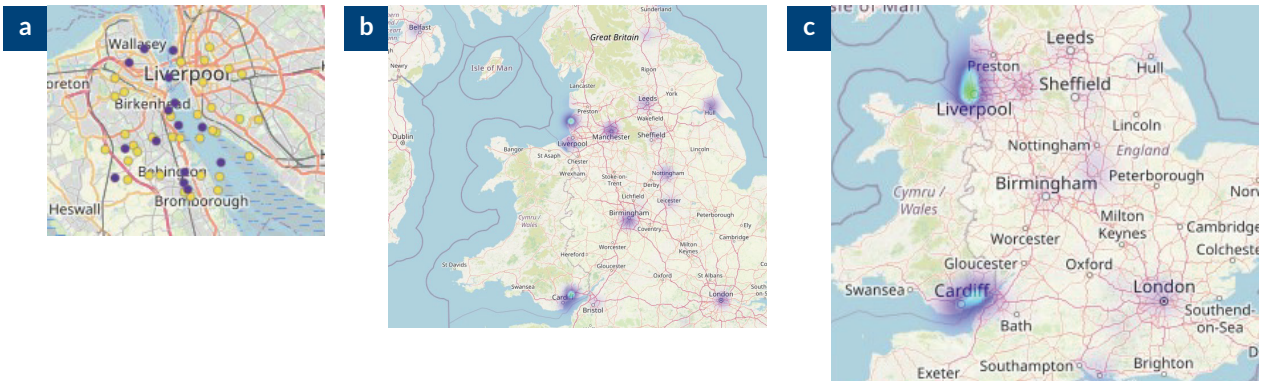


Figure 6. Patterns of misinformation distribution: (a) prevalence of misinformation in Liverpool according to the misinformation classifier; (b) concentration of tweets that do not contain misinformation in UK cities where riots took place; (c) heat map of the tweets that were classified as containing misinformation by the model. Note: Yellow dots are tweets classified as misinformation by the classifier.

On the other hand, in Southport, where the stabbing took place, we find there is a concentration of tweets that do not contain misinformation. Figure 6b shows a heatmap of the spatial distribution of tweets that were classified as not containing misinformation. This trend illustrates the premise that misinformation about a spatially located event, such as in this case study, is less common near the event. We can see that north of Liverpool, where Southport is, contains a high concentration of data that does not contain misinformation. Interestingly, we also notice that Newport, located northeast of Cardiff, contains a high concentration of tweets that were classified as containing misinformation and those that were classified as not containing misinformation by the model. Table 1 below shows the distribution of tweets in multiple major cities in the UK.

During the exploration of the tweets, we found that there were not many tweets about riots right after the fatal stabbing on July 29, 2024. Instead, the tweets that talked about riots started pouring in after a particular incident took place that sparked the riots against a mosque in Liverpool. Figure 7 shows a line chart that shows the number of tweets from July 29, 2024, to July 31, 2024.

Table 1. Tweets in each city, including the number of misinformation tweets.

City	Newport	Southport	Liverpool	London	Manchester	Birmingham	Cardiff	Leicester	Hull	Belfast	Leeds	Sunderland	Blackpool	Plymouth	Bristol
Total Tweets	102	61	50	19	19	17	14	11	10	9	6	5	4	4	3
Not Misinformation	69	47	14	16	17	17	14	8	8	9	5	5	4	4	3
Misinformation	33	14	36	3	2	0	0	3	2	0	1	0	0	0	0

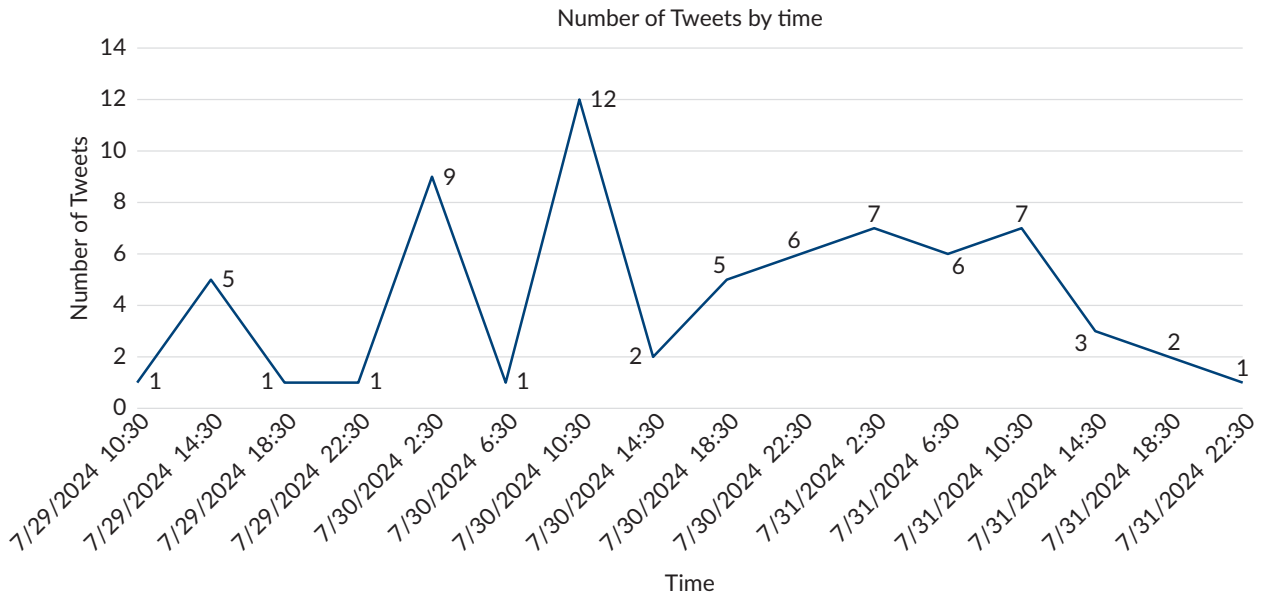


Figure 7. Line chart showing the incidence of tweet data from July 29, 2024, to July 31, 2024.

We noticed a pattern where the number of tweets about riots increased noticeably after an incident where a man with a knife was arrested in Liverpool near a vigil for the girls killed in the stabbing incident in Southport. In the tweets, many X users identify the person as a Muslim. However, according to the BBC, the man arrested with a knife near the vigil is called Jordan Davies and is not a Muslim (O'Neill & PA News, 2024). Instead, he was on his way to joining a mob that was about to stir up chaos in the area (O'Neill & PA News, 2024).

From Figure 7, we can see that there were only seven tweets in total right after the stabbing incident which took place on July 29. However, the number of tweets increased very quickly on July 30 to a total of 35 tweets. We attribute this increase to the incident where a man with a knife was arrested near the victims' vigil near Southport. Misinformation spread on X, claiming that the man with the knife was Muslim, but later reports from the BBC clarified that he is neither Muslim nor far-right (O'Neill & PA News, 2024).

After updating the misinformation model with the true labels, we identified a set of keywords that are common between the tweets that the model labeled as containing misinformation. Figure 8 displays a word cloud that includes these keywords.

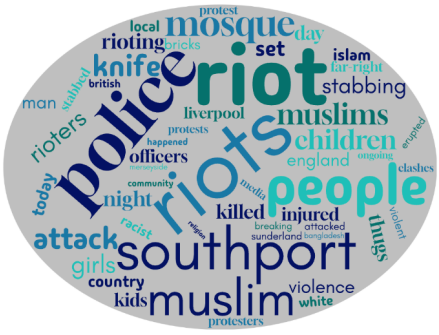


Figure 8. Word cloud with the most common keywords in the tweets that the model labeled as containing misinformation.

4.4.2. Effectiveness of Misinformation Identification

We have observed that SMART 2.0 initially identified misinformation without human input, with the initial accuracy being 53.7%. Some of these errors were remedied by correcting the labels of some of these tweets and then updating the labels for all the data using the human-in-the-loop approach.

Evaluating the interactive misinformation classification model in this case study is imperative. We want to assess the correctness and accuracy of the misinformation classification model as the user interactively updates the misinformation labels of the data. We hypothesize that the more the user corrects the misinformation labels of the data, the higher the model's accuracy would be at classifying the data's misinformation labels. To test this hypothesis, we conducted an experiment where we manually labeled the tweets in the case study for misinformation using a set of ground truths and guidelines.

We used the following ground truths: the Southport stabbing suspect is not a Muslim; the Southport stabbing suspect is not an immigrant; and the Southport stabbing suspect is Christian and was born and raised in the UK. In regards to guidelines, we used the following in the labeling process: tweets with racist, rude, and/or inappropriate undertones are not necessarily labeled as containing misinformation; and tweets that oppose ground truths are labeled as containing misinformation.

We manually labeled all the 370 tweets in this case study using the ground truths and guidelines described above. We wrote a computer program that starts with the un-updated model with the dataset being split into four equal parts. We used a stepwise iterative training approach to allow the models to learn new patterns related to the case study. In each iteration, a subset of the data is used to train and update model weights, followed by an evaluation of their performance. This process is repeated at 25%, 50%, 75%, and 100% of the training data. Figure 9 shows a graphical representation of the results.

The computer program that we created simulates user actions in SMART 2.0 when the user updates the misinformation label of a given data point. The program also simulates the retrieval of the label of any given data point in the user's session. Doing so enables us to conduct this experiment automatically, making it drastically faster to run, alter, and improve.

As seen in Figure 9(a), we find that as the user updates the misinformation classification model, it becomes more accurate, as seen in the increase in accuracy per iteration. The initial accuracy of the model (without any human input) was around 0.53. Although accuracy improved from 53% to 95.4% through iterative updates, this evaluation was conducted only on the case study dataset, which limits the generalizability of the results. Future work should include cross-validation on larger, unseen datasets to ensure robustness and to test the model's ability to generalize to diverse scenarios. Figure 9(b) also shows a similar trend of increasing precision, recall, and F1 score, meaning that the model's performance improved significantly with each iteration. Figure 9(c) also shows improvements in the model's classification results by the apparent decrease in the number of false positives and false negatives the more the model is updated.

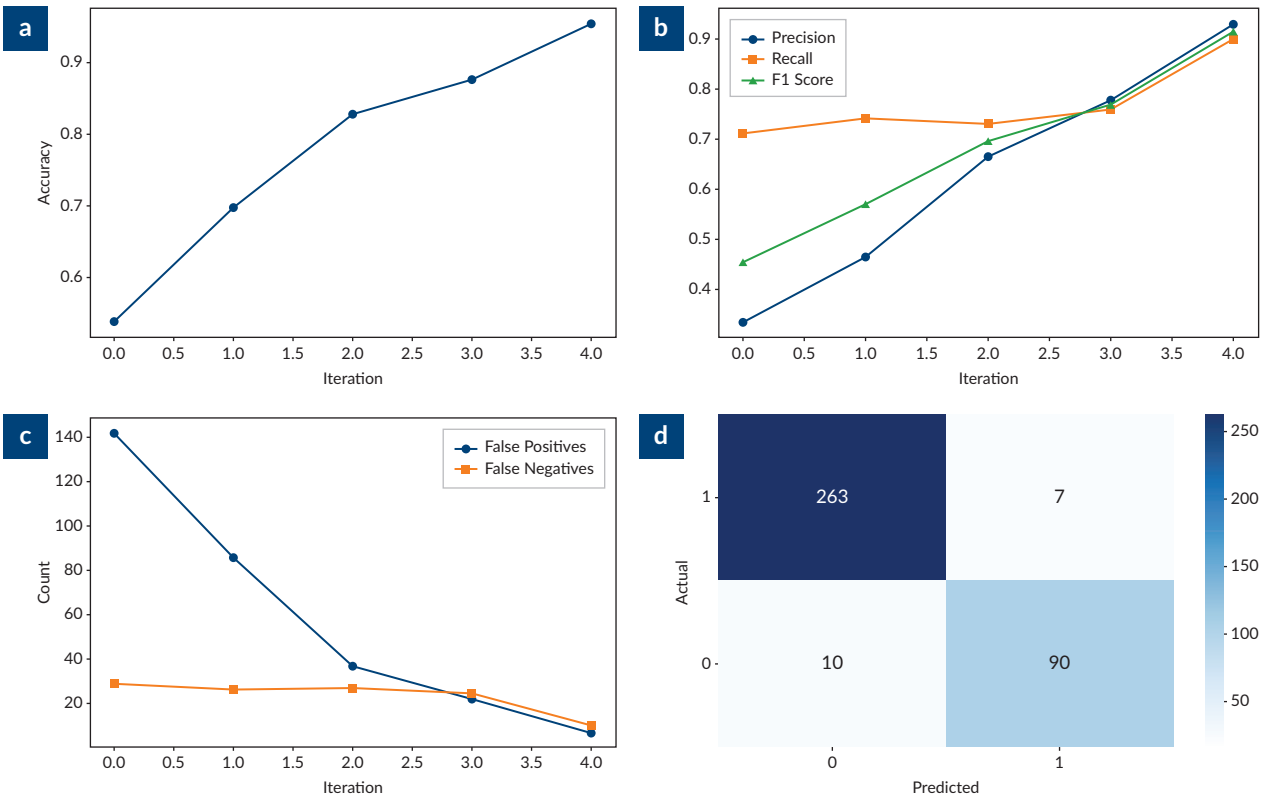


Figure 9. Evaluation results of the interactive misinformation model: (a) accuracy of the model for each quarter of the dataset; (b) model's precision, recall, and F1 score at each iteration; (c) error rates, including the number of false positives and the number of false negatives at each iteration while updating the model; (d) final confusion matrix of the model, created after the model was updated with the true labels.

4.4.3. Notable Results

We were surprised to discover that a lot of the data points and tweets were about the man with the knife who was arrested near the vigil in Southport. We assumed that most of the riots and the misinformation were due to the initial fatal stabbing of the three girls, but that turned out not to be the case.

5. Discussion and Future Improvements

The case study of the UK riots in 2024 demonstrates the powerful capabilities of SMART 2.0 in tracking and analyzing misinformation spread during a critical event. The study revealed how quickly false information about a perpetrator's identity and background spread on social media, particularly on X, highlighting the need for real-time monitoring and analysis tools like SMART 2.0. Contrasted with more accurate information in Southport and Newport, the concentration of misinformation in Liverpool shows how misinformation can have localized effects, underscoring the importance of SMART 2.0's geospatial analysis capabilities. We found a trend that illustrates the premise that misinformation about a spatially located event is less common near the event. However, this premise was not tested well and should be expanded on in future research.

The spike in misinformation following the arrest of a man with a knife near the victims' vigil illustrates how secondary events can amplify and reshape the dissemination of misinformation, and SMART 2.0's temporal

analysis features were crucial in identifying this pattern. It proved valuable in improving misinformation classification accuracy, and the ability to refine the misinformation classification model through user input was essential for context-specific misinformation. Our experiment showed that the more the user updates and interacts with the misinformation classification model, the better its accuracy and performance. However, one of the main downsides of our analysis and experiment is the low number of data points. In future research, we need to have more data sources and data points to evaluate the performance of the misinformation classifier effectively.

The case study demonstrated the importance of cross-referencing social media data with official sources, like BBC reports, to establish ground truth and identify misinformation accurately. These findings underscore the complex nature of misinformation spread during crisis events and the value of tools like SMART 2.0 in providing real-time insights to researchers, journalists, and policymakers. Based on the experiences and insights gained from this case study, we identified several areas for future improvement of SMART 2.0, including:

1. Enhanced language models: Develop more sophisticated NLP models to better understand context and implicit references in text-based data.
2. Cross-platform integration: Expand SMART 2.0's capabilities to integrate data from multiple platforms, providing a more comprehensive view of misinformation dissemination.
3. Automated fact-checking: Implement automated fact-checking features to cross-reference claims in social media posts with reliable news sources and official statements in real time.
4. Trend prediction: Develop predictive models that can forecast potential misinformation trends based on early signals in social media data.
5. User network analysis: Incorporate features to analyze the networks of users spreading misinformation, identifying key influences and bot networks.
6. Extend multilingual support: Expand language support to better track misinformation in diverse linguistic contexts, particularly in multilingual regions.
7. Integration with traditional media monitoring: Develop features to correlate social media misinformation trends with coverage in traditional media outlets, such as TV stations and news sites.

6. Conclusion

The SMART 2.0 has demonstrated its effectiveness in tracking and analyzing misinformation during critical events, as seen in the case of the 2024 UK riots. SMART 2.0's insights into the spread of false information, its geographical patterns, and the role of trigger events highlight its data visualization and multi-dimensional analysis capabilities. The tool adapts quickly through user feedback, supports multiple languages, and helps researchers understand the complex nature of misinformation.

However, the case study also revealed potential areas for improvement, such as enhanced language understanding and misinformation classification, and increasing the number of data points. As social media continues to shape public discourse, tools like SMART 2.0 are crucial for combating misinformation and supporting evidence-based decision-making and real-time monitoring.

Acknowledgments

This work is supported in part by the North Atlantic Treaty Organization (MYP-REACT G5812) and the Data Institute for Societal Challenges, University of Oklahoma. The authors would also like to acknowledge the support of Wagner Rosa for the initial creation of the misinformation model and the English editing support by Dr Yessenia Torres.

Funding

SMART 2.0 was partially funded by NATO REACT, Grant # MYP-REACT G5812.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability

Data used in this study will be made available upon reasonable request, provided the request is legitimate and academically relevant. Contact the corresponding author for more details.

References

Ahmed, J. (2024, August 6). Mapped: Violent protests grip the country with fears of more to come. *The Independent*. https://www.the-independent.com/news/uk/crime/uk-riot-locations-protests-map-far-right-b2591874.html

Ards mosque community "nervous" after overnight attack. (2024, August 11). *BBC News*. https://www.bbc.com/news/articles/c984dzxg6eko

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, *3*(1), Article 1. https://doi.org/10.1609/icwsm.v3i1.13937

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: A survey of techniques, tools and platforms. *AI & SOCIETY*, *30*(1), 89–116. https://doi.org/10.1007/s00146-014-0549-4

Calma, J. (2023, May 31). Twitter just closed the book on academic research. *The Verge*. https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research

Chiapponi, E., Dacier, M., Thonnard, O., Fangar, M., Mattsson, M., & Rigal, V. (2022). An industrial perspective on web scraping characteristics and open issues. In Y. Amir & C. Nita-Rotaru (Eds.), *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks—Supplemental Volume (DSN-S)* (pp. 5–8). IEEE. https://doi.org/10.1109/DSN-S54099.2022.00012

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Frayer, L. (2024, August 5). Misinformation online fueled all-out race riots in the United Kingdom. *NPR*. https://www.npr.org/2024/08/05/nx-s1-5063345/misinformation-online-fueled-all-out-race-riots-in-the-united-kingdom

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann.

Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017). Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In S. Matwin, S. Yu, & F. Farooq (Eds.), *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1803–1812). ACM. https://doi.org/10.1145/3097983.3098131

Ilia State University. (n.d.). *Network for alerting and managing public safety and resilience—REACT*. https://iliauni.edu.ge/en/siaxleebi-8/gonisdziebebi-346/network-for-alerting-and-managing-public-safety-and-resilience-react4.page

Lawless, J. (2024, July 31). British police charge 17-year-old with murder over a stabbing attack that killed 3 children. *AP News*. https://apnews.com/article/uk-stabbing-attack-southport-far-right-violence-a2e43d0d49776c138790d083713873f7

O'Neill, L., & PA News. (2024, August 9). Man caught with knife near vigil for Southport stabbing victims jailed. *BBC*. https://www.bbc.com/news/articles/c4gz79dln5xo

Otis, J. (2024, August 9). Covering the U.K. riots amid disorder and misinformation. *The New York Times*. https://www.nytimes.com/2024/08/09/insider/uk-riots.html

Quan-Haase, A. & Sloan, L. (Eds.). (2022). *The Sage handbook of social media research methods* (2nd ed). Sage.

Rapoza, K. (2017, February 26). Can "fake news" impact the stock market? *Forbes*. https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market

Ryan-Mosley, T. (2023, November 21). Meta is giving researchers more access to Facebook and Instagram data. *MIT Technology Review*. https://www.technologyreview.com/2023/11/21/1083760/meta-transparency-research-database-nick-clegg

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), Article 4787. https://doi.org/10.1038/s41467-018-06930-7

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Snyder, L. S., Karimzadeh, M., Chen, R., & Ebert, D. S. (2019). City-level geolocation of tweets for real-time visual analytics. In S. Gao, S. Newsam, L. Zhao, D. Lunga, Y. Hu, B. Martins, X. Zhou, & F. Chen (Eds.), *GeoAI '19: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (pp. 85–88). ACM. https://doi.org/10.1145/3356471.3365243

Snyder, L. S., Karimzadeh, M., Stober, C., & Ebert, D. S. (2019). Situational awareness enhanced through social media analytics: A survey of first responders. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)* (pp. 1–8). IEEE. https://doi.org/10.1109/HST47167.2019.9033003

Snyder, L. S., Lin, Y.-S., Karimzadeh, M., Goldwasser, D., & Ebert, D. S. (2020). Interactive learning for identifying relevant tweets to support real-time situational awareness. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 558–568. https://doi.org/10.1109/TVCG.2019.2934614

Stokel-Walker, C. (2024, February 27). Under Elon Musk, X is denying API access to academics who study misinformation. *Fast Company*. https://www.fastcompany.com/91040397/under-elon-musk-x-is-denying-api-access-to-academics-who-study-misinformation

Syed/London, A. (2024, August 5). How online misinformation stoked anti-migrant riots in Britain. *TIME*. https://time.com/7007925/misinformation-violence-riots-britain

Treen, K. M. I., Williams, H. T. P., & O'Neill, S. J. (2020). Online misinformation about climate change. *WIREs Climate Change*, *11*(5), Article e665. https://doi.org/10.1002/wcc.665

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

About the Authors

Mahmoud Mousa Hamad is a software developer at the Data Institute for Societal Challenges at the University of Oklahoma. Originally from Palestine, Mousa Hamad earned his BS and MS in computer science from the University of Oklahoma. His expertise spans full-stack web applications development, database management, mobile applications development, and machine learning model creation. For the past three years, he has served as the lead engineer on the Social Media Analytics and Reporting Tool (SMART), and he is currently focused on developing a new version that will make the tool more modern, scalable, and user-friendly.

Gopichandh Danala is a research scientist at the Data Institute for Societal Challenges at the University of Oklahoma. He received an MS and PhD in electrical and computer engineering from the University of Oklahoma, focused on medical informatics in developing computer-aided design (CAD) tools to improve diagnosis and patient care using computer vision, machine learning, and decision-making environments. His research interests are healthcare analytics, medical informatics, data science, and visual analytics to develop decision-making systems that address real-world problems and enhance societal well-being.

Wolfgang Jentner is a research scientist at the Data Institute for Societal Challenges at the University of Oklahoma. He received his PhD in visual analytics, a research discipline at the intersection of machine learning, interactive visualization, and human–computer interaction, from the University of Konstanz, Germany. His research interests include social media text analysis, explainable AI, visual analytics for civil security and critical infrastructure, and One Health.

David Ebert is a Gallogly chair professor of electrical and computer engineering (ECE) and director of the Data Institute for Societal Challenges at the University of Oklahoma. He is an IEEE fellow, recipient of the 2017 IEEE Computer Society vgTC Technical Achievement Award, and an adjunct professor at Purdue University. He performs research in visual analytics, novel visualization techniques, interactive machine learning and explainable AI, human–computer teaming, advanced predictive analytics, and procedural abstraction of complex, massive data.

Media and Communication
2025 • Volume 13 • Article 9483
https://doi.org/10.17645/mac.9483

cogitatio

ARTICLE

Open Access Journal

Detecting Covid-19 Fake News on Twitter/X in French: Deceptive Writing Strategies

Ming Ming Chiu [1,2], Alex Morakhovski [1], Zhan Wang [1], and Jeong-Nam Kim [3,4]

[1] Analytics\Assessment Research Center, The Education University of Hong Kong, Hong Kong
[2] Department of Special Education and Counseling, The Education University of Hong Kong, Hong Kong
[3] Gaylord College of Journalism and Mass Communication, University of Oklahoma, USA
[4] Debiasing & Lay Informatics (DaLI) Lab, USA

Correspondence: Ming Ming Chiu (mingchiu@eduhk.hk)

Submitted: 27 October 2024 **Accepted:** 13 February 2025 **Published:** 10 April 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

Many who believed Covid-19 fake news eschewed vaccines, masks, and social distancing; got unnecessarily infected; and died. To detect such fake news, we follow deceptive writing theory and link French hedges and modals to validity. As hedges indicate uncertainty, fake news writers can use it to include falsehoods while shifting responsibility to the audience. Whereas *devoir* (must) emphasizes certainty and truth, *falloir* (should, need) implies truth but emphasizes external factors, allowing writers to shirk responsibility. *Pouvoir* (can) indicates possibility, making it less tied to truth or falsehood. We tested this model with 50,000 French tweets about Covid-19 during March–August 2020 via mixed response analysis. Tweets with hedges or the modal *falloir* were more likely than others to be false, those with *devoir* were more likely to be true, and those with *pouvoir* showed no clear link to truth. Tweets of users with verification, more followers, or fewer status updates were more likely to be true. These results extend deceptive writing theory and inform fake news detection algorithms and media literacy instruction.

Keywords

Covid-19; deception; disinformation; fake news; French; hedges; modals; uncertainty

1. Introduction

Many people believed Covid-19 fake news, failed to take preventive measures (e.g., vaccines, wearing masks, social distancing), got infected unnecessarily, and died. In April 2020 alone, 82 websites spreading false information (fake news) about Covid-19 got 460 million views (Avaaz, 2020). By October 2020, such fake

news led to 130,000 additional Covid-19 deaths in the US (Redlener et al., 2020). Hence, detection of Covid-19 fake news is critical for preventing its spread and saving lives.

Detecting fake news is hard. Even with training, most humans struggle to spot it (Lutzke et al., 2019). For example, alternative media (e.g., *209 Times*) can mix 99% real news articles (e.g., Associated Press news) with 1% fake news articles—which itself mostly contains facts (Shaw, 2021).

As some fake news writers are not anonymous, we propose that they evade responsibility for false information by using deceptive writing strategies (unlike bots or foreign agents who do not care about their reputation). Such strategies can distract readers by shifting attention from the writer to them (e.g., you vs. I), evoking their emotions ("catastrophe!"), burdening them with cognitive complexity (e.g., medical terminology), or raising uncertainty (Chiu et al., 2024). Specifically, the French modal *falloir* (should, need) implies truth but emphasizes external factors, allowing writers to shirk responsibility. We propose that writers exploit these attributes of the modal *falloir* and use them to disseminate fake news.

In this study, we test whether French modals (especially *falloir*) are linked to truth or falsehood. We examine 50,000 French tweets about Covid-19 from March to August 2020 using a mixed response model (Hox et al., 2017).

2. Uncertainty Strategies

Grounded in deceptive writing theory (Chiu & Oh, 2021), a writer can use uncertainty (hedging) to dodge accountability and let readers make their own judgments. Writers hedge to limit their commitment to the truth of a claim or to avoid stating it outright (Hyland, 1998). Hence, hedges can free a writer from the chains of truth, giving a reader the reins to interpret it (Chiu & Oh, 2021).

Commonly used hedging strategies include: hypothetical, conditional acceptance, subjective view, limited scope, and epistemic uncertainty (Hyland, 1998). Take this sentence: "If the pandemic ends quickly, you might be right; otherwise, I argue that Covid will likely sterilize many victims." First, "if" creates an alternate world, separating this claim from reality (hypothetical; Chen & Chiu, 2008). Second, saying "you might be right" offers conditional acceptance instead of asserting an absolute truth (modal auxiliary; Boncea, 2013). Third, "argue" marks a personal view, not an indisputable fact (lexical-modal verb subjectivisation; Namasaraev, 1997). Fourth, "many" restricts the claim to some victims, not all (approximate marker of frequency, time, degree, quantity, etc.; Boncea, 2013). Lastly, "likely" indicates uncertainty (adjectival/adverbial/nominal modal phrases). Hence, we propose the following two hypotheses:

H1a: Among tweets, those with hedges are less likely to be true.

H1b: Among tweets, those with hedges are more likely to be false.

3. French Modals and Fake News

French writers often use modals (*devoir*, *falloir*, *pouvoir*) to indicate different degrees of certainty. *Devoir* (roughly "must") typically indicates certainty and an unbreakable grip on the truth. Although *falloir* (roughly

"should" or "need") points to truth, it emphasizes outside conditions, like obligations, making it less certain and less binding. *Pouvoir* (roughly "can") merely suggests possibility, carrying little weight of truth or responsibility. Hence, writers of fake news might lean away from *devoir* and toward *falloir* to signal uncertainty, evade responsibility, and dodge blame. Note, however, that each modal has multiple functions and its meaning can differ across contexts (Hacquard & Cournane, 2016).

3.1. Devoir

Devoir indicates epistemic certainty of human knowledge, social duty, or future events (Caron & Caron-Pargue, 2003), making it more likely to align with objective facts. Take this example (first in the original French language, and then translated by the authors; modals emphasises by the authors):

> C'est mon 1er cas COVID19 + que je *dois* transférer en soins intensifs dans un autre centre. Extrêmement stressant pour tout le personnel qui procède au transfert. Chapeau aux ambulanciers, infirmières et inhalothérapeutes!
>
> This is my first Covid-19 positive case that I *have to* transfer to intensive care in another facility. Extremely stressful for all the staff involved in the transfer. Hats off to the paramedics, nurses, and respiratory therapists!

This writer is certain about how to proceed ("je *dois* transférer en soins intensifs dans un autre centre," I have to transfer to intensive care in another facility). So, readers expect the writer to take full responsibility and act accordingly. Otherwise, they would blame him for his failure. Hence, fake news writers might avoid *devoir*.

Devoir can also signal social obligation: "J'*dois* [sic] déménager ds [sic] 1 semaine officiel ils vont m'arrêter sur la route c'est la merde" (I've *got to* move out in 1 week, I'm sure they'll stop me on the road, it's shit). This writer reports a duty to move out. So others expect him to do so.

Also, *devoir* can indicate future events: "Coronavirus: la distance de deux mètres *devra* être maintenue « pour des mois » au Québec" (Coronavirus: the two-meter distance *will have to* be maintained "for months" in Quebec). People plan their future actions based on this expected event.

Devoir sets a high bar for truth and responsibility. Hence, fake news writers might avoid it:

> H2a: Among tweets, those with *devoir* are more likely to be true.
>
> H2b: Among tweets, those with *devoir* are less likely to be false.

3.2. Falloir

Falloir verbs can indicate goal constraints, situation-based constraints, or necessities. Like *devoir*, *falloir* suggests true information but underscores how external conditions, such as social or cultural obligations, make it true (de Saussure, 2017). Unlike nations with egalitarian cultural values (e.g., Australia, Netherlands), many people in France readily accept unequal distributions of power and obey authority (according to representative national surveys: 64/100 power distance [Chiu & McBride-Chang, 2010]; 4.24/7 hierarchical

value [House et al., 2004]). So, they might be more likely to accept and follow obligatory information accompanying *falloir*. Unlike *devoir*, the external constraints of *falloir* limit the scope of the writer's views and shift responsibility away. As such, fake news writers might exploit it. By invoking external authority, they might persuade their readers to accept their information (and act on it) while dodging blame. Consider this goal constraint:

> #chloroquine Il *faut* arrêter de discuter, ça marche! Pr #Péronne qui soutient le Pr #Raoult. Les experts qui conseillent le gouvernement sont des spécialistes du sida, ça n'a rien à voir avec ce virus! @LCI
>
> #chloroquine We *need* to stop debating, it works! Prof. #Péronne who supports Prof. #Raoult. The experts advising the government are specialists in AIDS, which has nothing to do with this virus! @LCI

Grounded in the view that chloroquine "works," the writer sets the goal of stopping debate, thereby creating a basic constraint for future actions. Embedding false information in the basis for the goal constraints creates a false foundation for interpreting subsequent information (and acting accordingly).

Falloir can also express situational constraints: "Lieux concernés, sanctions encourues…Ce qu'il *faut* savoir sur l'obligation de porter le masque dans les lieux publics clos—Le Monde" (Places concerned, penalties incurred…What you *need to* know about the obligation to wear a mask in enclosed public places—Le Monde). This writer emphasizes different Covid-19 constraints across places (e.g., infection density) and the penalties for violations. By framing information as situational constraints, fake news writers have plausible deniability about its relevance.

Falloir can also indicate necessity (e.g., legal, social, conventional): "Je pige rien au [sic] règles du covid, *faut* porter le masque dehors aussi?" (I don't get the covid rules, you *have to* wear the mask outside too?). This writer questions the rules regarding the need to wear a mask outside. Hence, fake news writers can use *falloir* to question necessity without a solid backing.

Overall, *falloir*'s goal, situation-based, or necessity constraints often reflect social conventions rather than objective truths. These constraints can be necessary but insufficient: The prerequisite action might not yield the expected effect without other factors. Hence, fake news writers might use *falloir* to imply false claims as socially accepted truths but evade responsibility:

> H3: Among tweets, those with *falloir* are more likely to be false.

3.3. Pouvoir

Pouvoir can indicate hypothetical possibilities, human/social permissions, physical abilities, subjective human views, variable occurrences across place or time (scope), or futures (Meisnitzer, 2012). Unlike *devoir* and *falloir*, *pouvoir* does not claim that its information is true. Instead, it suggests that something might be true or might happen. Hence, its information is less likely to be true compared to the information accompanying *devoir*. So, writers can slip in false information with *pouvoir*, but its weak commitment to truth makes readers less likely to believe it or act on it. *Pouvoir* is then less persuasive than *falloir*, and fake news writers might favor *falloir* over *pouvoir*.

This example of *pouvoir* indicates possibility: “Parfait, on *peux* [sic] y voir le début de la fin #COVID19” (Perfect, you *can* see the beginning of the end #COVID19). This writer imagines a world in which Covid-19 is ending, but possible worlds might not be true.

Pouvoir can also reflect human/social permission: “Bon j’ai [sic] pas le COVID je *peux* partir en vacances 😂” (Well I don’t have COVID I *can* go on vacation 😂). This writer gives themselves permission to vacation but might not do so.

As the following tweet shows, *pouvoir* expresses physical ability:

> Récolte de quelques moments de grâce de la journée. Mes lutins sont formidables! #plusbeaumetierdumonde O. , 5 ans, “je ne *peux* pas faire de câlins à mes copains parce qu’ya le virus, mais je *peux* en faire à l’arbre, car c’est mon ami, l’arbre” 😂😂😂#amiedelanature
>
> Harvesting a few of the day’s moments of grace. My elves [kids] are amazing! #bestjobintheworld O. 5 years old, “I *can’t* hug my friends because of the virus, but I *can* hug the tree, because it’s my friend, the tree” 😂😂😂#friendofnature

The writer contrasts the inability to hug friends with the ability to hug a tree. However, ability does not dictate action.

Pouvoir can also show subjective views: “Les kleinder je *peux* braver le coronavirus pour ca [sic]” (Les kleinder [the little ones, German] I *can* brave the coronavirus for that). This writer says that her children motivate her brave actions, but she might not actually do so.

Pouvoir can also indicate limited scope/conditions: “bon bah je suis négatif au covid19 so lundi je *peux* me faire opérer yay” (well, I’m covid19 negative so on Monday I *can* get operated on yay). The operation depends on staying Covid-free, which might not happen.

Furthermore, *pouvoir* can point to the future: “Tu *peux* être sûr que les écoles seront vides” (You *can* be sure that the schools will be empty). This writer assures that schools will be empty in the future, but future events cannot be validated.

As these examples show, *pouvoir* makes much weaker claims about truth compared to *devoir*. Hence, writers can weave in falsehoods with *pouvoir*, but its high uncertainty renders readers less likely to believe it or act on it. As *pouvoir* is less persuasive than *falloir*, fake news writers might favor *falloir* over *pouvoir*. As such, we hypothecize the following:

> H4a: Tweets with *pouvoir* are less likely than those with *devoir* to be true.
>
> H4b: Tweets with *pouvoir* are less likely than those with *falloir* to be false.

3.4. This Study

We test seven hypotheses regarding hedges and French modals. Among tweets, those with hedges are less likely to be true (H1a) or more likely to be false (H1b). Tweets with *devoir* are more likely than others to be true (H2a) or less likely to be false (H2b). Tweets with *falloir* are more likely than others to be false (H3). Lastly, tweets with *pouvoir* are less likely than those with *devoir* to be true (H4a) or *falloir* to be false (H4b).

Hypotheticals (*si*/if) or subjunctives (*que*/that) in French tweets with modals might be linked to the validity of Covid-19 news. Hence, we include conditionals and subjunctives in our statistical model to reduce omitted variable bias (Cinelli & Hazlett, 2020). We also control for emotional tone (valence, arousal; Monnier & Syssau, 2014).

4. Method

France grapples with a flood of fake news (Beauvais, 2022). As French has few modals, it is a suitable springboard for testing whether modals are linked to Covid-19 news validity.

4.1. Data

From 18,935 users, we downloaded a total of 50,000 tweets about Covid-19 written in French and their meta-data from X (2024). To assess their validity, we used OpenAI's GPT-4o and Anthropic's Claude-3.5 Sonnet (machine learning or natural language processing requires extremely costly training with a large, curated database of verified true and false news to assess validity). Both GPT-4o and Claude-3.5 Sonnet handle accents and misspellings, so we did not need further pre-processing (e.g., remove symbols, spell-check, etc.). For $\alpha = 0.05$ and a small effect size of 0.1, the statistical power for 18,935 users and 50,000 tweets both exceeded 0.99 (Cohen, 2013).

4.2. Procedure

We determined whether a tweet is true (e.g., "Covid-19 can kill you"), false ("the common flu is more dangerous than Covid"), or cannot be determined from public information ("my dad is scared of getting Covid") by giving ChatGPT 4o and Claude-3.5 Sonnet a specific prompt (see Supplementary File, Appendix A). Two fluent French speakers coded 450 of these tweets: One is a 32-year-old French native man who works in the aerospace industry (hereafter Human 1); and the other is a 28-year-old Swedish-born, female business researcher, who has lived in France for six years and speaks the language fluently (hereafter Human 2).

4.3. Variables

User variables include follower count and status updates. Tweet variables include Date, Likes, and Replies.

The following are dichotomous variables: Sensitive indicates whether a tweet has content that might offend users; for the 10,005 tweets coded either true or false, True_cut is 1 if true or 0 if false.

The remaining variables use all 50,000 tweets. True is 1 if true, 0 otherwise. False is 1 if false, 0 otherwise. Validity is −0.5 if false, 0 if undetermined, and 0.5 if true (contrast coding; Ravenscroft & Buckless, 2017).

We computed six sets of pairwise inter-rater reliabilities among Human 1, Human 2, GPT-4o, and Claude-3.5 Sonnet for True_cut, True, False, and Validity via Krippendorff's alpha. Krippendorff's alpha applies to incomplete data, any sample size, any measurement level, any number of coders or categories, and scale values. Ranging from −1 to 1, an alpha exceeding 0.67 shows satisfactory agreement.

We also used GPT-4o to identify hedges and tested its inter-rater reliability with a human's judgment of 100 tweets (50% with hedges, 50% without hedges) via Krippendorf's alpha.

We created the following modal variables: *Devoir* indicates whether any of its verb forms are in a tweet, without a hypothetical and without a subjunctive. Similarly, the following variables likewise indicate whether they are in a tweet: *Falloir*, *Pouvoir*, *Devoir* hypothetical, *Falloir* hypothetical, *Pouvoir* hypothetical, *Devoir* subjunctive, *Falloir* subjunctive, and *Pouvoir* subjunctive (see online Appendix at https://bit.ly/4jV3RvB).

We also captured the meaning of each modal in each tweet via GPT-4o and tested whether each specific meaning was related to whether a tweet was true, false, or cannot be determined by public information. Possible *devoir* meanings were epistemic certainty, social duty, or future events. Possible *falloir* meanings were goal constraints, situation constraints, or necessity. Possible *pouvoir* meanings were hypothetical, human/social permission, physical ability, subjective human view, variable occurrences (scope), or future. We tested for inter-rater reliability via Krippendorf's alpha with GPT-4o and a human on 300 tweets with equal proportions of each modal meaning.

We also tested whether emotional valence or arousal was linked to True_cut, True, False, or Validity to reduce potential omitted variable bias (Cinelli & Hazlett, 2020). Monnier and Syssau (2014) had 469 volunteer, fluent French-speaking students rate the emotional sentiments of 1,031 common French words (969 nouns and 62 adjectives, excluding common stopwords like *les* [the]). Each rated 115 words along two dimensions on a 9-point scale. Valence ranges from negative (e.g., *fureur* [fury]) to positive (*joie* [joy]). Arousal ranges from low passion (*ennui*) to high passion (*zèle* [zeal]).

4.4. Data Analysis

To test our hypotheses using these data, we address analytic difficulties involving outcomes (discrete, infrequent, multiple types) and explanatory variables (many hypotheses' false positives, comparison of effect sizes, robustness) with specific statistics strategies (see Table 1). For outcomes, we model: (a) dichotomous and ordered outcomes with Logit/Probit, ordered Logit/Probit, and odds ratios (Martinez et al., 2017); (b) infrequent outcomes with Logit bias estimator (King & Zeng, 2001); and (c) multiple types of outcomes (dichotomous and ordered) with mixed response models (Hox et al., 2017). For explanatory variables, we model: (d) many hypotheses' false positives with the two stage linear step-up procedure (Benjamini et al., 2006); (e) comparison of effect sizes with Lagrange multiplier tests (Bertsekas, 2014); and (f) consistency of results across data sets (robustness) with separate single outcome models (Hansen, 2022).

Table 1. Statistics strategies addressing each analytic difficulty.

Analytic difficulty	Statistics strategy
Outcome variables	
Discrete variable (yes/no)	Logit/Probit; odds ratios
Ordered variable (fake, neither, true)	Ordered Logit/Probit; odds ratios
Infrequency (< 25%)	Logit bias estimator
Multiple types of outcomes (Y_1, Y_2, ...)	Mixed response model
Explanatory variables	
Many hypotheses' false positives	Two-stage linear step-up procedure
Compare effect sizes ($\beta_1 > \beta_2$?)	Lagrange multiplier tests
Consistency of results across data sets (Robustness)	Separate, single outcome models

4.5. Explanatory Model

We model three outcomes GPT_false, GPT_true, and GPT_validity (VALIDITY; vectors are capitalized) at the same time via a mixed response model:

$$\text{VALIDITY}_{yi} = \beta_y + e_{yi} + \beta_{ys}\text{USER}_{yi} + \beta_{yt}\text{TWEET}_{yi} + \beta_{yu}\text{EMOTION}_{yi} + \beta_{yv}\text{MODAL}_{yi} + \beta_{yw}\text{SUBJUNCTIVE}_{yi} + \beta_{yx}\text{MODAL\_MEANINGS} + \beta_{yz}\text{INTERACTIONS}_{yi}$$

In the vector VALIDITY_{yi}, outcome y (GPT_false, GPT_true, GPT_validity) of tweet i has a grand mean intercept β_y with an unexplained component (residual) e_{yi}.

We enter explanatory variables in sequential sets (vectors) to estimate the variance explained by each set (Hansen, 2022). Structural variables can influence malleable process variables, so the former precede the latter. Users write tweets, so we first enter USER attributes (Verified, Registration date/time, Followers, Status updates) followed by TWEET (Date/time, Sensitive, Quoted characters, Hedge, Likes, Retweets, Replies). Next, we enter EMOTION (Valence, Arousal), Modal (*Pouvoir*, *Devoir*, *Falloir*), hypotheticals (*Devoir* hypothetical, *Falloir* hypothetical, *Pouvoir* hypothetical), and subjunctives (*Devoir* subjunctive, *Falloir* subjunctive, and *Pouvoir* subjunctive). Then, we enter MODAL_MEANINGS (*Devoir*, *Devoir* epistemic certainty, *Devoir* social duty, *Devoir* future events, *Falloir*, *Falloir* goal constraints, *Falloir* situation constraints, *Falloir* necessity, *Pouvoir*, *Pouvoir* hypothetical, *Pouvoir* human/social permission, *Pouvoir* physical ability, *Pouvoir* subjective human view, *Pouvoir* variable occurrences [scope], *Pouvoir* future). Lastly, we test their INTERACTIONS.

A nested hypothesis test ($\Delta\chi^2$LL) checks the significance of each set of explanatory variables (Hansen, 2022). For greater accuracy and less multicollinearity, we drop non-significant variables (which do not cause omitted variable bias; Cinelli & Hazlett, 2020). We then run a parallel binary logit regression for GPT_True_cut. Afterwards, we apply the same procedure for Claude_false, Claude_true, Claude_validity and Claude_True_cut.

5. Results

5.1. Inter-Rater Reliability

Inter-rater reliability varied across codes and coders (Human 1, Human 2, GPT-4o, Claude-3.5 Sonnet; see Table 2). Human or GPT-4o assessments of True_cut showed extremely high inter-rater relibility (0.97–0.98). However, they were lower for False (0.86–0.91), Validity (0.70–0.74), and True (0.60–0.72). These results showed that the greatest coding difficulty was distinguishing between true tweets and those that cannot be determined by public information.

Claude's inter-rater reliability with humans or GPT for True_cut was good, ranging from 0.85 to 0.88. However, all other judgments of True vs. other, False vs. other, and Validity were poor, ranging from 0.47 to 0.60. In all cases, GPT-4o outperformed Claude.

Inter-rater reliability between GPT-4o and Human 1 was good for hedges and modals (Krippendorff's alpha: Hedge = 0.92; *Devoir* = 0.79; *Falloir* = 0.77; *Pouvoir* = 0.80).

Table 2. Inter-rater reliability (Krippendorff's alpha) among Human 1, Human 2, GPT-4o, and Claude-3.5 Sonnet for true_cut, true, false, and validity.

Coders	True_cut	True	False	Validity
Human 1 vs. Human 2	0.98	0.72	0.86	0.74
Human 1 vs. GPT	0.99	0.70	0.86	0.73
Human 2 vs. GPT	0.97	0.60	0.91	0.70
Human 1 vs. Claude	0.88	0.55	0.60	0.51
Human 2 vs. Claude	0.85	0.48	0.58	0.47
GPT vs. Claude	0.88	0.49	0.57	0.47

5.2. Summary Statistics

Modal uses in these tweets match common French usage (Hütsch, 2018; see summary statistics in Table 3 and correlation–variance–covariance matrices in the Table B1 of the Supplementary File). These French Covid-19 tweets were two or three times more likely to be true than false (as measured by GPT-4o or Claude-3.5 Sonnet, respectively). By contrast, US tweets about Covid-19 at the same time were 11 times more likely to be false than true (Chiu et al., 2024).

Table 3. Summary statistics (N = 50,000).

Variable	Mean	*SD*	Min	Median	Max
Outcome					
GPT true	0.131		0	0	1
GPT false	0.069		0	0	1
GPT cannot be determined by public information	0.800		0	1	1
Claude true	0.179		0	0	1
Claude false	0.063		0	0	1
Claude cannot be determined by public information	0.758		0	1	1

Table 3. (Cont.) Summary statistics (N = 50,000).

Variable	Mean	*SD*	Min	Median	Max
User					
Registration date/time	41,664.730	1,325.189	39,061.940	41,395.000	44,038.160
Verified	0.075		0	0	1
Followers (millions)	0.440	1.327	0	0.005	25,759
Status updates (millions)	0.066	0.128	0	0.022	1,267
Tweet					
Date/time	43,964.150	38.528	43,918.040	43,955.490	44,044.080
Sensitive	0.009		0	0	1
Quoted characters	12.316	80.719	0	0	4,047
Hedge	0.228		0	0	1
Likes	256.540	1,722.349	0	8	2,051.98
Retweets	98.205	562.304	0	3	69,313
Replies	18.175	99.185	0	1	9,063
Emotion					
Arousal	4.811	0.730	2.140	5	7.860
Valence	5.411	1.149	1.360	5	8.580
Modal					
Devoir (must)	0.033		0	0	1
Devoir conditional	0.009		0	0	1
Devoir subjunctive	0.001		0	0	1
Falloir (should, need)	0.028		0	0	1
Falloir conditional	0.002		0	0	1
Falloir subjunctive	0.001		0	0	1
Pouvoir (can)	0.055		0	0	1
Pouvoir conditional	0.013		0	0	1
Pouvoir subjunctive	0.001		0	0	1
GPT meanings					
Devoir	0.042		0	0	1
Devoir social duty	0.023		0	0	1
Devoir future events	0.007		0	0	1
Devoir epistemological certainty	0.004		0	0	1
Falloir	0.030		0	0	1
Falloir necessity	0.018		0	0	1
Falloir goal constraint	0.007		0	0	1
Falloir situation constraint	0.003		0	0	1
Pouvoir	0.073		0	0	1
Pouvoir hypothetical	0.032		0	0	1
Pouvoir variation/scope	0.011		0	0	1
Pouvoir human/social permission	0.010		0	0	1
Pouvoir subjective	0.007		0	0	1
Pouvoir physical ability	0.005		0	0	1
Pouvoir future	0.002		0	0	1

5.3. Explanatory Model

As GPT-4o showed higher inter-rater reliability with human coders, we report the GPT-4o results here and the Claude-3.5 Sonnet results in Appendix B of the Supplementary File (their results were generally consistent). All results in this section described the first entry into the regression, controlling for all previous entries. Ancillary regressions and tests are available upon request.

5.3.1. True_cut

User attributes, tweet attributes, and modals were linked to GPT true_cut French tweets about Covid (vs. false ones; see Table 4). Verified users' tweets were much more likely than unverified user tweets to be true (odds ratio [OR] = 1.590; see Table 4, model 1, top, left). Also, tweets by users with more followers were more likely to be true (OR = 1.631), whereas tweets by users with later registration dates were less likely to be true (OR = 0.962; see Table 4, model 1, top, left). Sensitive tweets were more likely to be true (OR = 2.071), while those with hedges were more likely to be false (OR = 0.962), supporting H1a (see Table 4, model 2, centre). Tweets with *devoir* were more likely to be true (OR = 1.455), supporting H2a (see Table 4, model 3, bottom). By contrast, tweets with *falloir* were more likely to be false (OR = 0.807),

Table 4. Summary results of a binary logit regression modelling True_cut with unstandardized regression coefficients (standard errors in parentheses) and odds ratios.

	GPT True_cut		
Explanatory variable	Model 1 User	Model 2 + Tweet	Model 3 + Modal
User			
Verified	0.464*** (0.102) 1.590[a]	0.428*** (0.100) 1.534[a]	0.430*** (0.099) 1.537[a]
Followers (millions)	0.489*** (0.023) 1.631[a]	0.491*** (0.024) 1.634[a]	0.483*** (0.024) 1.621[a]
Registration date (years)	−0.038*** (0.006) 0.963[b]	−0.039*** (0.006) 0.962[b]	−0.039*** (0.006) 0.962[b]
Tweet			
Sensitive		0.728** (0.238) 2.071[a]	0.709** (0.244) 2.032[a]
Hedge		−0.424*** (0.047) 0.654[b]	−0.413*** (0.047) 0.662[b]
Modal			
Must (*devoir*)			0.375** (0.123) 1.455[a]
Should (*falloir*)			−0.215* (0.108) 0.807[b]
McFadden's R^2	0.059	0.065	0.066

Notes: The outcome True_cut only includes true versus false values and it excludes "cannot be determined based public information"; [a] = odds ratios exceeding one (greater likelihood); [b] = odds ratios below one (lower likelihood); * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

supporting H3. *Pouvoir* was not significant, supporting H4a. This model accounted for nearly 7% of the differences in true_cut (McFadden $R^2 = 0.066$).

5.3.2. True

User attributes and modals were linked to true French tweets about Covid-19. Verified users' tweets were more likely than others' to be true (OR = 1.093; see Table 5, model 1, top, middle). Tweets by users with more followers than others were more likely to be true (OR = 1.062). Tweets with *devoir* were more likely than other tweets to be true, supporting hypothesis H2a (OR = 1.114; see Table 5, model 2, right, bottom). *Pouvoir* was not significant, supporting H4a. The final model accounted for nearly 3% of the variance.

Table 5. Summary results of mixed response model modelling True with unstandardized regression coefficients (standard errors in parentheses) and odds ratios.

	GPT True	
Explanatory variable	Model 1 User	Model 2 + Modal
User		
Verified	0.089* (0.044) 1.093[a]	0.091* (0.044) 1.095[a]
Followers (millions)	0.060*** (0.008) 1.062[a]	0.059*** (0.008) 1.061[a]
Modal		
Must devoir		0.108* (0.053) 1.114a
Explained variance	0.023	0.028

Notes: These results are part of a mixed response model with two other outcomes: False and Valid; separating the results into different tables aids readability; [a] = odds ratios exceeding one; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.3.3. False

User attributes, tweet properties, and modals were linked to false French tweets about Covid-19. Tweets by users with more followers were less likely than others to be false (OR = 0.867) while those with more status updates were more likely to be false (OR = 1.631; see Table 6, model 1, top, left). Meanwhile, tweets with hedges or with *falloir* were more likely to be false (respectively, $OR_{hedge} = 1.083$, see Table 6, model 2, centre; and $OR_{falloir} = 1.135$, see Table 6, model 3, bottom, right), supporting H1b and H3. Pouvoir was not significant, supporting H4b. This model accounted for less than 1% of the variance (0.009).

5.3.4. Valid

User attributes, tweet attributes, and modals were linked to an ordered variable valid (false, cannot be determined, true). Verified users' tweets were more valid than unverified users' tweets (OR = 1.301; see Table 7, model 1, top, left). Tweets by users with more followers were more valid (OR = 1.263). By contrast, tweets by users with later registration dates were less valid (OR = 0.963). Tweets with hedges were less valid (OR = 0.628), supporting H1a, while those with greater emotional arousal were more valid (OR = 1.105; see Table 7, model 2, centre, lower). Tweets with *devoir* were more valid, supporting H2a

Table 6. Summary results of mixed response model modelling False with unstandardized regression coefficients (standard errors in parentheses) and odds ratios.

	GPT False vs. other		
Explanatory variable	Model 1 User	Model 2 + Tweet	Model 3 + Modal
User			
Followers (millions)	−0.143*** (0.010) 0.867[b]	−0.149*** (0.010) 0.862[b]	−0.143*** (0.010) 0.867[b]
Status updates (millions)	0.489*** (0.024) 1.631[a]	0.476*** (0.024) 1.610[a]	0.483*** (0.024) 1.621[a]
Tweet			
Hedge		0.080** (0.026) 1.083[a]	0.082** (0.025) 1.085[a]
Modal			
Should falloir			0.127* (0.057) 1.135[a]
Explained variance	0.003	0.008	0.009

Notes: These results are part of a mixed response model with two other outcomes: True and Valid; [a] = odds ratios exceeding one; [b] = odds ratios below one; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 7. Summary results of mixed response model modelling Valid with unstandardized regression coefficients (standard errors in parentheses) and odds ratios.

	GPT Validity		
Explanatory variable	Model 1 User	Model 2 + Tweet	Model 3 + Modal
User			
Verified	0.263*** (0.045) 1.301[a]	0.221*** (0.044) 1.247[a]	0.218*** (0.045) 1.244[a]
Followers (millions)	0.151*** (0.010) 1.163[a]	0.150*** (0.010) 1.162[a]	0.154*** (0.010) 1.166[a]
Registration date (years)	−0.038*** (0.003) 0.963[b]	−0.039*** (0.003) 0.962[b]	−0.039*** (0.003) 0.962[b]
Tweet			
Hedge		−0.466*** (0.027) 0.628[b]	−0.462*** (0.027) 0.630[b]
Arousal		0.100*** (0.015) 1.105[a]	0.102*** (0.015) 1.107[a]
Modal			
Must *devoir*			0.127* (0.060) 1.135[a]
McFadden's R^2	0.025	0.026	0.027

Notes: These results are part of a mixed response model with two other outcomes: True and False; [a] = odds ratios exceeding one; [b] = odds ratios below one; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

(OR = 1.135; see Table 7, model 3, bottom, right). *Pouvoir* was not significant, supporting H4a. The final model accounted for nearly 3% of the variance.

6. Discussion

Grounded in deceptive writing theory, we tested whether hedges and French modals were linked to true versus false information. Our results supported most of our hypotheses. Tweets with hedges were less likely to be true and more likely to be false. Those with *devoir* (must) were more likely to be true. Those with *falloir* (should, need) were more likely to be false. Those with *pouvoir* (can) were less likely to be (a) true than those with *devoir*, and (b) false than those with *falloir*. These results fit our theoretical model of hedges and modals and extend deceptive writing theory.

6.1. Hedges

Hedges were more likely to co-occur with falsehoods and less likely to co-occur with truth. These results fit the view that hedges allow some uncertainty about the truth (Hyland, 1998). As fake news authors can use hedges to weaken the strengths of their assertions, such weaker claims set off fewer validity alarms and facilitate audience consideration of them (Hyland, 1998). Likewise, face-to-face speakers can use hedges to share false information while dodging accountability (Chiu & Oh, 2021).

6.2. Modals

Tweets with *devoir* (must) were more likely than other tweets to be true. This result fits with the view that devoir highlights epistemic certainty of human knowledge, human/social obligation, or future events (Caron & Caron-Pargue, 2003).

Tweets with *falloir* (should, need) were more likely than others to be false. This result aligns with the view that *falloir* implies an expectation of truth but highlights external constraints, thereby reducing the scope of human actions (de Saussure, 2017) and limiting the writer's responsibility for false information. Furthermore, the hierarchical cultural value of French people with their greater respect toward superiors might give *falloir* more persuasive force (House et al., 2004). This pairing of expected truth and less responsibility is the sweet spot for fake news writers. As these results suggest, fake news writers exploit this pairing to increase reader acceptance of fake news while avoiding blame.

If future studies confirm this, people should be wary of *falloir*, as accompanying information is more likely than otherwise to be false. Those on the lookout for fake news should recognize *falloir* as a possible deceptive writing strategy, so they should carefully check the validity of such information—especially if it urges action.

Pouvoir (can) showed weaker effects than *devoir* and *falloir*. Indeed, it was not linked to truth or falsehood. This result coheres with the view that *pouvoir* only weakly indicates the possibility of events (Meisnitzer, 2012) and does not make strong claims about truth. Conversely, its non-significant link to falsehood suggests that its weak claim to truth is less useful than *falloir* to fake news writers, so they are more likely to use *falloir* than *pouvoir* for false information.

6.3. Implications

If future studies confirm these results, they have implications for theory, methodology, and practice. First, these results support the uncertainty claims of deceptive writing theory, and imply that any comprehensive theory of fake news must include hedges, modals, and their mechanisms.

More broadly, this study's methodology showcases how to detect linguistic links to false information in a large corpus of messages. Practically, educators can include such deceptive writing strategies in their media literacy curriculum for students and adults, helping more people identify fake news. Notably, this general approach of identifying linguistic markers linked to fake news can inform detection of it without known facts (e.g., the beginning of the Covid-19 pandemic with little scientific knowledge).

Likewise, these results can help developers of fake news detection software improve its accuracy. They can recognize the presence of hedges, *falloir*, and other deceptive writing strategies and assess accompanying information for fake news—and instigation of dangerous actions! Furthermore, developers can identify sources or social networks that frequently use such strategies and hinder or prevent them from creating fake news.

6.4. Limitations

This study's limitations include its sample, explanatory variables, and validity checks. The sample only included French tweets during the first six months of news about Covid-19, mostly from France. Future studies can include more languages, longer time periods, and more countries. As this study only examined modals, future studies can control for other explanatory variables: topics, author profiles, previous tweets, culture, or other linguistic attributes. Furthermore, this study did not capture the grammatical necessities of modals that can cause miscategorization. As miscategorization introduces measurement error (noise) into a statistical analysis, it typically reduces the detection of a significant result (signal). As the results were significant, the noise was not sufficient to affect the results. Still, future studies can track grammatical necessities for greater accuracy. Lastly, this study only used two humans to check the validity of ChatGPT assessments on a subset of the tweets. Future studies can have more humans check more data.

7. Conclusion

This study showed how French hedges and modals were linked to truth or falsehood. Tweets with hedges were less likely than others to be true and more likely to be false, those with *devoir* were more likely than others to be true, those with *falloir* were more likely than others to be false, and those with *pouvoir* showed no clear link to the truth. These results fit deceptive writing theory and implied that fake news authors used (a) hedges to hide falsehoods under uncertainty and (b) *falloir* to falsely imply truth while emphasizing the effects of external factors. Both strategies help such authors dodge responsibility. Hence, these findings can inform software developers creating tools to detect fake news and help educators develop suitable media literacy curricula and lessons.

Funding

The work described in this article was fully supported by a grant from the Senior Research Fellow Scheme sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. SRFS2223–8H01).

Conflict of Interests

In this article, editorial decisions were undertaken by Jaemin Jung (Korea Advanced Institute of Science and Technology).

Data Availability

The data for this article was collected on X, with the search results for keywords "coronavirus, covid, covid19, covid_19, corona, coronalockdown, covid 19, stay home, social distancing, medical masks, fake news, pandemic, virus, lockdown, quarantine," based on relevancy and recency, from March 28, 2020, to August 1, 2020. The data set is available here: https://api.twitter.com/2/tweets/search

Supplementary Material

Supplementary material for this article is available online through the following link: https://bit.ly/4jV3RvB

References

Avaaz. (2020). *Facebook's algorithm*. https://secure.avaaz.org/campaign/en/facebook_threat_health

Beauvais, C. (2022). Fake news: Why do we believe it? *Joint Bone Spine*, *89*(4), Article 105371. https://doi.org/10.1016/j.jbspin.2022.105371

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. https://doi.org/10.1093/biomet/93.3.491

Bertsekas, D. P. (2014). *Constrained optimization and Lagrange multiplier methods*. Academic.

Boncea, I. J. (2013). Hedging patterns used as mitigation and politeness strategies. *Annals of the University of Craiova*, *2*(1), 2–25.

Caron, J., & Caron-Pargue, J. (2003). A multidimensional analysis of French modal verbs pouvoir, devoir and falloir. In F. H. van Eemeren, J. A. Blair, C. A. Willard, & A. F. Snoeck Henkemans (Eds.), *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation* (pp. 165–169). International Center for the Study of Argumentation.

Chen, G., & Chiu, M. M. (2008). Online discussion processes: Effects of earlier messages' evaluations, knowledge content, social cues and personal information on later messages. *Computers & Education*, *50*(3), 678–692. https://doi.org/10.1016/j.compedu.2006.07.007

Chiu, M. M., & McBride-Chang, C. (2010). Family and reading in 41 countries: Differences across cultures and students. *Scientific Studies of Reading*, *14*, 514–543.

Chiu, M. M., Morakhovski, A., Ebert, D., Reinert, A., & Snyder, L. S. (2024). Detecting Covid-19 fake news on Twitter: Followers, emotions, relationships, and uncertainty. *American Behavioral Scientist*, *68*(10), 1269–1289. https://doi.org/10.1177/00027642231174329

Chiu, M. M., & Oh, Y. W. (2021). How fake news differ from personal lies. *American Behavioral Scientist*, *65*(2), 243–258. https://doi.org/10.1177/0002764220910243

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

de Saussure, L. (2017). Why French modal verbs are not polysemous, and other considerations on conceptual and procedural meanings. In J. Blochowiak, C. Grisot, S. Durrleman, & C. Laenzlinger (Eds.), *Formal models in the study of language* (pp. 281–296). Springer. https://doi.org/10.1007/978-3-319-48832-5_15

Hacquard, V., & Cournane, A. (2016). Themes and variations in the expression of modality. *Proceedings of NELS*, *46*, 21–42.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The globe study of 62 societies*. Sage.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge. https://doi.org/10.4324/9781315650982

Hütsch, A. (2018). A quantitative perspective on modality and future tense in French and German. D. Ayoun, A. Celle, L. & Lansari (Eds.), *Tense, aspect, modality, and evidentiality: Crosslinguistic perspectives* (pp. 19–40). Kazan Federal University.

Hyland, K. (1998). Boosting, hedging and the negotiation of academic knowledge. *Text & Talk*, *18*(3), 349–382. https://doi.org/10.1515/text.1.1998.18.3.349

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking. *Global Environmental Change*, *58*, Article 101964. https://doi.org/10.1016/j.gloenvcha.2019.101964

Martinez, B. A. F., Leotti, V. B., Nunes, L. N., Machado, G., & Corbellini, L. G. (2017). Odds ratio or prevalence ratio? An overview of reported statistical methods and appropriateness of interpretations in cross-sectional studies with dichotomous outcomes in veterinary medicine. *Frontiers in Veterinary Science*, *4*, Article 193. https://doi.org/10.3389/fvets.2017.00193

Meisnitzer, B. (2012). Modality in the romance languages: Modal verbs and modal particles. In W. Abraham & E. Leiss (Eds.), *Modality and theory of mind elements across languages* (pp. 335–359). De Gruyter. https://doi.org/10.1515/9783110271072.335

Monnier, C., & Syssau, A. (2014). Affective norms for French words (FAN). *Behavior Research Methods*, *46*, 1128–1137. https://doi.org/10.3758/s13428-013-0431-1

Namasaraev, V. (1997). Hedging in Russian academic writing in sociological texts. In R. Markkanen & H. Schröder (Ed.), *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts* (pp. 64–80). De Gruyter. https://doi.org/10.1515/9783110807332.64

Ravenscroft, S. P., & Buckless, F. A. (2017). Contrast coding in ANOVA and regression. In T. Libby & L. Thorne (Eds.), *The Routledge companion to behavioural accounting research* (pp. 349–372). Routledge. https://shorturl.at/7E56H

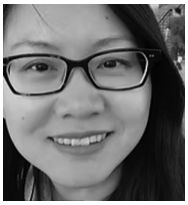
Redlener, I., Sachs, J. D., Hansen, S., & Hupert, N. (2020). *130,000-210,000 avoidable Covid-19 deaths—and counting—in the US*. National Center for Disaster Preparedness. https://ncdp.columbia.edu/wp-content/uploads/2020/10/Avoidable-COVID-19-Deaths-US-NCDP.pdf

Shaw, Y. & Natisse, K. M. (Host). (2021, April 29). The chaos machine: An endless hole (season 7, episode 2). [Audio podcast episode]. In *Invisibilia*. NPR. https://www.npr.org/transcripts/992017530

About the Authors

Ming Ming Chiu is chair (distinguished) professor of analytics and diversity at The Education University of Hong Kong. He invented statistical discourse analysis (SDA), multilevel diffusion analysis (MDA), artificial intelligence statistician, and online detection of sexual predators. He studies fake news, inequalities, learning, international comparisons, and automatic statistical analyses.

Alex Morakhovski is a data scientist and AI engineer specializing in machine learning, generative AI, and cloud computing. He develops AI models, optimizes data processes, and applies algorithms to build solutions for AI-driven applications, leveraging these techniques to enhance automation, predictive analytics, and decision-making.



Zhan Wang is an applied linguist. Jan's research focuses on first and second-language acquisition, computational linguistics, and technology-support learning. She is working on projects related to digital humanities, fake news detection, and AI in language education.

Jeong-Nam Kim is a communication theorist known for the situational theory of problem solving (STOPS). He leads the DaLI Lab, tackling public biases and failing information markets. Kim holds the Gaylord chair at Oklahoma and fellowships at USC, Salamanca, and KAIST, advancing lay informatics and strategic communication.

Media and Communication
2025 • Volume 13 • Article 9552
https://doi.org/10.17645/mac.9552

cogitatio

ARTICLE

Open Access Journal

Investigating Publics' Communicative Action in Problem Solving (CAPS) Through Data Science

Sunha Yeo [1,2], Joohee Kim [3,4], Juwon Kim [3,5], and Sungahn Ko [3,5]

[1] Gaylord College of Journalism and Mass Communication, University of Oklahoma, USA
[2] Debiasing and Lay Informatics Lab (DaLI), University of Oklahoma, USA
[3] Human–AI Interaction and Visualization Lab (HAIV), Republic of Korea
[4] Ulsan National Institute of Science & Technology (UNIST), Republic of Korea
[5] Pohang University of Science and Technology (POSTECH), Republic of Korea

Correspondence: Sungahn Ko (sungahn@postech.ac.kr)

Submitted: 31 October 2024 **Accepted:** 6 March 2025 **Published:** 2 June 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

This study examined the communicative action in problem solving model through data science-driven approaches to enhance the understanding of online publics' communication behaviors. Using ChatGPT, the study analyzed YouTube comments from news channels that covered a contentious entertainment issue with multiple related events. The findings indicated that communication behaviors changed over time and manifested in diverse combinations. In addition, the study demonstrated that aware publics in the digital era were not merely passive; rather, they engaged in communication activities similar to active publics. Furthermore, it identified distinct communication behaviors associated with certain types of publics, indicating that public types also evolve dynamically across events. The results suggested that the communicative action in problem solving model served as a valuable framework for examining online communication behaviors in the digital era. Based on these insights, this study offered both academic and practical contributions to the field.

Keywords

communicative action in problem solving; online communication; online communication behaviors; public types; strategic communication

1. Introduction

Over the past decade, researchers have employed data science-driven approaches increasingly to analyze large datasets in communication studies (Bolsover & Howard, 2019; boyd & Crawford, 2012; Chang et al., 2023; Chen et al., 2020; Choi, 2020; Freelon et al., 2016; Howard et al., 2011; Murthy, 2017). In this process, computer-assisted methods have been used to explore various communication theories, including agenda-setting theory (Guo & Vargo, 2015; Neuman et al., 2014; Vargo & Guo, 2017), cultivation theory (Song et al., 2023), attribution theory (Park et al., 2022), organization–public relationships (H. L. Lee, 2023), and the computer-mediated communication competence forecasting model (Chih-Ming & Ying-You, 2020). This approach is valuable, as it provides meaningful insights into reality through empirical observations grounded in theory (E. W. Lee & Yee, 2020; Helles & Ørmen, 2020), highlights theories' continued importance (Gil de Zuniga & Diehl, 2017; Mahrt & Scharkow, 2013; Parks, 2014), and contributes to their development (Wise & Shaffer, 2015). Accordingly, scholars have suggested that it would be valuable to integrate such theoretical frameworks as the two-step flow model, the theory of normative social behavior, and the communal coping theory further into data-driven communication research (Rains, 2020).

Consistent with this trend, this study examines a theoretical framework in public relations using a data-driven analytical approach. Researchers have highlighted that public relations professionals should expand their knowledge of emerging technologies (Kent & Saffer, 2014) and acknowledge the role that advanced data analytics play in enhancing public relations research, particularly in areas such as audience targeting, landscape analysis, and evaluations (Holtzhausen & Zerfass, 2015; J. E. Grunig, 2023; Weiner & Kochhar, 2016). However, as Wiesenberg et al. (2017) noted, the application of technical skills to large-scale data analysis remains underdeveloped in this field, which finds public relations scholars "lagging behind" (p. 26). To bridge this gap, this study uses a computational approach to apply public relations theory, with a specific focus on the communicative action in problem solving (CAPS) model. This model provides not only a way to investigate publics' online communication behaviors, but also offers a structured framework with which to identify public types based on these behaviors (Ni & Kim, 2009). By integrating data science, this study attempts to provide a fresh perspective on the CAPS model and presents valuable prospects for strategic communication.

Despite the CAPS framework's wide-ranging application in various studies (Chon & Harrell, 2024; Chon & Park, 2021; Krishna, 2018; Roh & Oh, 2021), the understanding of public communication behaviors that this model offers has yet to be explored fully. Based upon Dewey's (1927/1954) argument, CAPS conceptualizes the public as problem solvers (J.-N. Kim & Krishna, 2014) who respond differently to issues or problems (Ni & Kim, 2009) depending upon the context in which they are situated (J. E. Grunig & Kim, 2017). As situational conditions evolve continuously, CAPS posits that public reactions to issues or problems also change over time (Grunig, 1978; J.-N. Kim & Grunig, 2011). In addition, these reactions can manifest as both passive and active communication behaviors, and in some cases, both types may coexist simultaneously (Grunig, 1989; Krishna, 2018). Further, CAPS explains that there are representations of communication traits associated with each public type (Chon et al., 2023; J. E. Grunig & Kim, 2017), indicating that the publics' status is shifting dynamically in response to the evolution of communication behavior (Dozier & Ehling, 2013). However, traditional methods, such as web-based surveys, which have been used predominantly in previous CAPS studies, face limitations in examining these aspects. This is because each communication behavior has traditionally been analyzed as a separate variable at a single point in time, whereas

studying the evolution of various communication behaviors and public statuses simultaneously requires long-term observation.

Computation tools allow data to be gathered at different times (Parks, 2014). Thus, the dynamics of communication behaviors, their coexistence, and the evolving nature of public status over time can be investigated effectively using new technology research. In this respect, this study starts introducing the CAPS framework to explain different types of communication behaviors and discusses in detail the limitations of traditional research methods in CAPS studies. Subsequently, it employs a data science-driven approach to examine the online communication behaviors related to a continuous issue, the changes in these behaviors over time in response to key events within the issue, the diverse combinations of behaviors, the status of the publics who generate these behaviors, and the correlation between specific communication behaviors and public status.

The context of this study focuses on an issue within the entertainment industry, a field recognized widely as one of the most prominent arenas engaged with multiple societal issues (Elberse, 2013) and one that has long been intertwined closely with public relations, as P. T. Barnum exemplified (Tilson, 2016). It is anticipated that this new approach will help understand better not only the transitions and combinations of different communication behaviors but also the dynamics of public types associated with each communication behavior within the CAPS framework. Further, it is anticipated that this approach will provide public relations practitioners with insights into the way that data-driven methodologies can enhance the understanding of online communication behaviors and help implement more detailed communication strategies to address issues or problems.

2. Literature Review

2.1. CAPS

The situational theory of problem-solving, an extension of one of the most widely recognized public relations theories—the situational theory of publics (J.-N. Kim & Krishna, 2014), offers valuable insights into the activeness of communication on the part of various types of publics. This framework is structured around three key dependent variables that represent fundamental communication behaviors: information acquisition, information selection, and information transmission. Collectively, these behaviors constitute CAPS, which serves as the primary theoretical foundation for this study.

CAPS has been found to be effective in understanding the public's active communication behavior in contexts such as climate change (Bhalla, 2022), public health crises (Chon & Park, 2021; Krishna, 2018), corporate social responsibility campaigns (Roh & Oh, 2021), and racial activism (Chon & Harrell, 2024). These studies demonstrate that the three behaviors enhance our understanding of public communication behaviors further.

2.1.1. Information Acquisition

Information acquisition is a communication behavior that pertains to the different degrees of searching for information to solve problems. It varies between proactive and reactive communicators. Proactive communicants engage in deliberate information seeking, a purposeful and systematic approach to acquiring

information to address specific problems or uncertainties. This behavior reflects a strategic effort to mitigate potential issues by gathering relevant data and insights actively before they become pressing concerns. Conversely, reactive communicants exhibit information attending, where they gather information incidentally rather than through an intentional search. These individuals may encounter information in their daily interactions or through accidental exposure. This reactive approach to information acquisition highlights a more passive stance, where the acquisition of information is secondary to immediate, unplanned circumstances (J.-N. Kim & Grunig, 2011).

2.1.2. Information Selection

The process of selecting information involves the way that individuals direct their focus in collecting and choosing information about an issue. Proactive communicants engage in information forefending, where they use a selective approach to manage information by applying an "only if" rule and weighing "relevance" as the criterion for whether to approach or ignore information about the problem. In contrast, reactive communicants practice information permitting, characterized by accepting and considering whatever information becomes available. This approach involves a more open attitude toward incoming information that allows them to process and incorporate a broader range of data into their understanding of the issue at hand (J.-N. Kim & Grunig, 2011; J.-N. Kim & Krishna, 2014).

2.1.3. Information Transmission

Information transmission pertains to how individuals disseminate information about an issue to others. Proactive communicants demonstrate information forwarding, a communication behavior where they share information in a positive manner, even in the absence of specific requests. This behavior reflects a proactive stance on disseminating information, where individuals take the initiative to spread knowledge and insights that potentially influence others' understanding and responses to issues. On the other hand, reactive communicants engage in information sharing and provide information only in response to direct requests. This behavior highlights a more passive approach, where the dissemination of information is contingent upon external prompts rather than self-initiated efforts. Such a distinction underscores the varying degrees of initiative and responsiveness in the way that individuals contribute to the communication process (Chon et al., 2023; J.-N. Kim & Grunig, 2011).

Examining publics' online communication behaviors based on those various dimensions can help organizations determine where to focus their efforts strategically and how to prepare effectively in different situations.

3. Limitations of Previous Studies Using the CAPS Model

Prior studies that focused on CAPS relied on traditional research methods. Specifically, previous studies have explored CAPS through interviews (Ni & Kim, 2009), experiments (Y. Kim, 2016), and a combination of mailed and web-based surveys (Shen et al., 2019). Among these, web-based surveys were used most frequently (Chon & Park, 2021; Chon et al., 2022; H. J. Kim & Hong, 2021; J.-N. Kim, Shen, & Morgan, 2011; Xu et al., 2021). Although these studies offered valuable implications, they also had several limitations.

First, the research methods employed in previous CAPS studies capture communication behaviors at a single point in time. While some studies have investigated all six dimensions (Chon et al., 2022; J.-N. Kim, Shen, & Morgan, 2011; Xu et al., 2021), others have focused on only three (H. J. Kim & Hong, 2021; Krishna, 2018) or just two (Chon & Park, 2021; J.-N. Kim & Lee, 2014). Regardless of the specific focus of CAPS research, each communication behavior has been observed within a cross-sectional framework. However, CAPS, which builds on Dewey's situational view of publics, acknowledges that certain publics are more situational (Ni & Kim, 2009), which means that their communication behaviors are not static, but instead, fluctuate over time in response to changing environmental conditions. Recognizing this characteristic, researchers can examine further the way that online communication behaviors adapt as conditions evolve.

Second, previous research has not only analyzed communication behaviors at a single point in time but has also examined each dimension individually as a dependent variable. However, in reality, communication is inherently multifaceted and often involves a combination of various communication behaviors. For example, a comment on a news report that states, "Oh my! I can't believe something like this happened! Can anyone tell me what happened next?" reflects both passive (information attending) and active (information seeking) information acquisition. Similarly, a statement like, "Any updates on this issue? If you're new and want to catch up on what has happened so far, feel free to visit my blog," combines active information acquisition (information seeking) with active information transmission (information forwarding). It is expected that researchers will be able to more effectively capture the combination of these behaviors in digital environments by adopting a new research approach.

Third, previous CAPS studies have focused primarily on publics' communication behavior. However, according to J. E. Grunig (1989), public segmentation can also be explored through communication behavior. He explained that publics can be categorized into four segments based upon the level of activeness (J. E. Grunig, 2013): nonpublics, latent publics, aware publics, and active publics. Traditionally, these types are determined by situational recognition of an issue or problem and are assessed through three key factors: problem recognition (is it a problem?), involvement recognition (is it your problem?), and constraint recognition (can you do something to solve this problem?; J. E. Grunig, 1997). Here, nonpublics are individuals who are unaware of any problems, while latent publics are involved in a problem but do not perceive it as problematic. Aware publics acknowledge both the problem and its implications, while active publics not only recognize the problem but also take action to address it. In essence, greater problem recognition and involvement recognition, coupled with lower constraint recognition, lead to greater public activeness. However, given that CAPS encompasses both active and passive dimensions, research can also investigate public segmentation further through this publics', which will provide a deeper understanding of the typical communication behaviors associated with each public type (J. E. Grunig & Kim, 2017).

This study proposes analyzing the CAPS framework using data science-driven approaches to address these limitations and expand the investigation of the CAPS model. This method makes it possible to overcome challenges related to the static timeframe, the isolated analysis of communication behaviors, and the previous inability to explore the connection between communication behavior and public types.

4. New Data Science-Driven Approaches in CAPS Studies

This study uses data science-driven approaches, including AI, to collect and analyze online communication behavior within the CAPS framework. This approach allows publics' online communication behaviors to be examined comprehensively, specifically: (a) how these behaviors evolve over time in response to key events; (b) the various ways that these behaviors combine; (c) the public types of those generating these behaviors; and (d) the association between specific communication behaviors and public types.

4.1. Exploring the Transformation of Online Communication Behavior

J.-N. Kim and J. E. Grunig (2011), who proposed the CAPS model, explained that it is based upon the assumption that most human behavior is motivated by the need to solve problems. Specifically, they described how communication behavior identifies solutions and applies them in two stages: the inquiring phase and the effectuation phase. The inquiring phase begins when individuals become motivated to solve a problem and seek information to identify and validate solutions. This phase is divided further into internal and external inquiring phases. The internal inquiring phase involves cognitive searches, where individuals rely on their prior knowledge and experiences. However, if they fail to activate relevant knowledge (J.-N. Kim & Grunig, 2011, p. 128) and continue to struggle to determine a course of action, they transition to the external inquiring phase, where they seek information from external sources. During these two inquiring phases, information acquisition behaviors become more pronounced. This includes information seeking, defined as scanning the environment deliberately to obtain information on a specific topic, and information attending, which refers to the unplanned discovery of information during the problem-solving process (J. E. Grunig, 1997).

On the other hand, sustained problem-solving efforts lead individuals to transition into the effectuating phases of problem-solving. The search process in this phase may be unfinished, but it evolves into a process of filtering out irrelevant data. As a result, information acquisition behaviors begin to fade, while information selection (i.e., information forefending and information permitting) and information transmission (i.e., information forwarding and information sharing) become more prominent. These behaviors can be performed individually or collectively, which leads to a further subdivision into the individual effectuating phase and the collective effectuating phase.

Figure 1 illustrates how problem solvers' information selectivity, transmission, and acquisition behaviors differ across these phases. Information acquisition increases gradually from the internal inquiring phase through the external inquiring phase and reaches its peak in the individual effectuating phase before it exhibits a sharp decline in the collective effectuating phase. In contrast, information selection and information transmission increase steadily across all three phases and reach their highest levels in the final collective effectuating phase.

These findings suggest that depending upon publics' situational motivation, predominant communication behaviors about an issue can shift over time. In this context, the following research question is presented:

> RQ1: Which specific communication behaviors within the CAPS framework are most prevalent in a contentious entertainment issue that encompasses multiple related events?

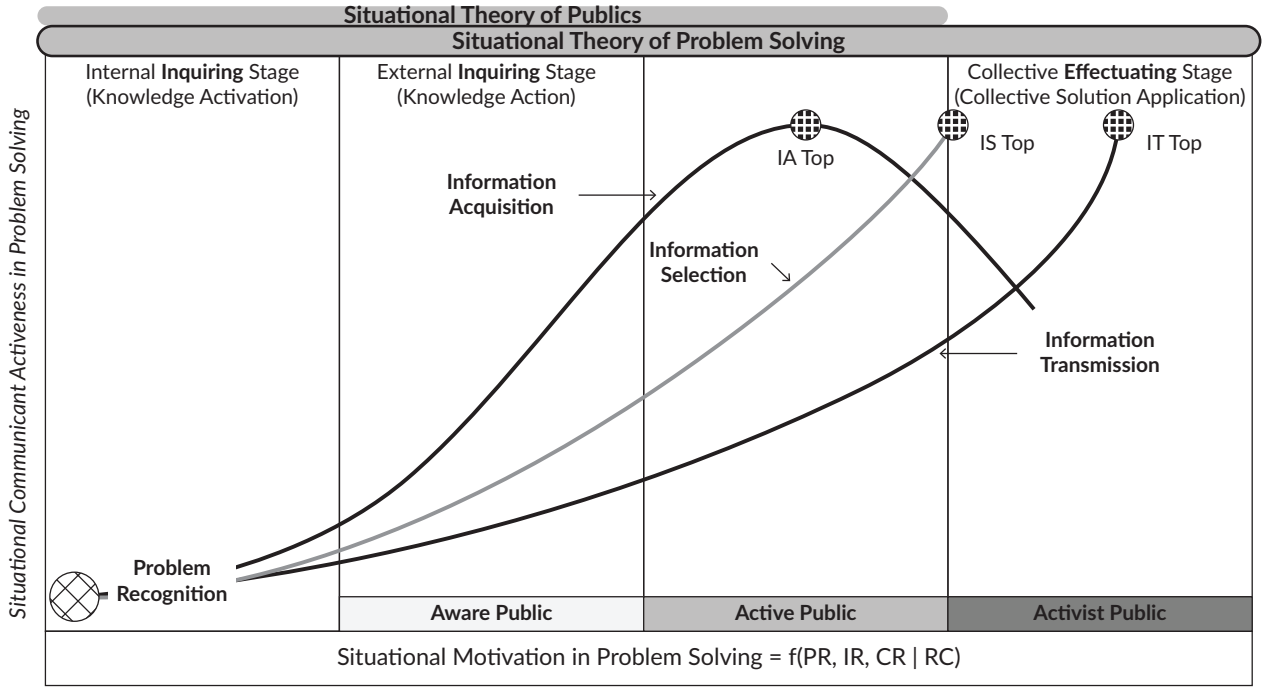


Figure 1. Sequential illustration of motivated information behavior in problematic situations. Source: J.-N. Kim (2006), Ni and Kim (2009). Notes: IA = information acquisition; IS = information selection; IT = information transmission.

4.2. Examining the Diverse Combinations of Online Communication Behaviors

J.-N. Kim and Krishna (2018) explained that the three main dimensions of CAPS—information selection, transmission, acquisition—often occur simultaneously, as they are not mutually exclusive. This simultaneity is also illustrated in Figure 1, while Y. Kim (2016) found that information attending (passive) plays a regulatory role in influencing other communicative behaviors.

However, beyond these established combinations, a broader spectrum of communication engagement may exist. For example, closed-minded individuals tend to seek out similar messages (Barnidge et al., 2020) and engage in spontaneous information-sharing behaviors (Hirsch, 2011). This suggests that three key communication behaviors can be identified among activists: information forefending (active), information seeking (active), and information sharing (passive). Therefore, this study suggests investigating the various combinations of communication behaviors that emerge and seeks to determine which combinations are observed online most commonly, particularly as they can be analyzed effectively using data science-driven approaches. In this context, the following research question is posed:

> RQ2: Are there specific combinations of communication behaviors within CAPS that are most commonly observed in an issue?

4.3. Confirming the Association Between Online Communication Behaviors and Public Status

Among various public types that J. E. Grunig (2013) introduced, most publics who engage in online communication can be categorized as active (high involvement and low constraint recognition) or activist

publics (high problem recognition, high involvement recognition, and low constraint recognition; J.-N. Kim & Grunig, 2011). This is the case because demonstrating communication behavior about specific issues online demonstrates at least some level of awareness, interest, or concern, which indicates a certain degree of recognition or involvement in the problem. In addition, online platforms facilitate not only the effortless expression of opinions, but also accessible participation in various forms of action—such as boycotts, union memberships, or demonstrations—which lower barriers to engaging with, and advocating for, an issue thereby. This suggests a relatively low level of constraint recognition.

In addition to active publics and activists, Baym's (1996) study identified the presence of aware publics online, characterized by individuals with low involvement in an issue (J. E. Grunig, 2013). Her research highlighted that computer-mediated communication employs a hybrid language that combines spoken and written discourse elements, as well as interpersonal and mass communication. This reflects that while online communication is primarily text-based, publics' interactive and transient nature renders the interaction similar to orality. Given that online communication exists along a written-oral continuum, Baym's (1996) study found that online interactions involve multiple participants, including publics who do not engage deeply and synchronously with an issue, but express interest in a message through subtle indications of intent or effort—a process that, in face-to-face conversations, would be conveyed typically through voice, gestures, or other nonverbal cues.

In summary, three distinct public types are expected to exhibit communication behaviors on online platforms. To investigate which types prevail in online discussions of an issue, this study poses the following research question:

> RQ3: Which public types are observed online most commonly during an issue?

On the other hand, active publics are more likely to become members of activist groups (J. E. Grunig, 1989). This suggests that public status is not fixed, but evolves continuously. Many researchers have acknowledged publics' dynamic nature. For instance, Nussbaum (2013) indicated that publics are formed and transformed through collective experiences and emotional engagement, while Dozier et al. (2013) and J. E. Grunig (1978) characterized them as fluid and evolving rather than static entities, as individuals engage in discussions, debates, and problem-solving. J. E. Grunig explained further that this continuous evolution occurs in response to shifting environmental conditions and emerging challenges. In this context, scholars such as Lünenborg (2019) and Paget (1929) have emphasized the importance of examining publics' dynamic nature. As a data science-driven approach enables progress to be tracked over time, this study seeks to analyze the transformation of public types throughout the discussion of an issue. This inquiry leads to the following research question:

> RQ4: How do public types evolve throughout a contentious entertainment issue that encompasses multiple related events?

4.4. The Association of CAPS and Public Types

Finally, this study attends to the fact that Figure 1 illustrates not only the transitions in a problem solver's different communication behaviors across various phases, but also demonstrates how each public type is

associated with specific communication behaviors. J.-N. Kim et al. (2010) noted that information acquisition, particularly information seeking, is unique to active problem solvers, as people look for information as they become more motivated to solve a problem (J. E. Grunig, 1997). In addition, active publics demonstrate moderate information transmission with a relatively high level of information selection and significant information acquisition (J.-N. Kim & Grunig, 2011). However, aware publics show low levels of information transmission and information selection, together with moderate information acquisition. On the other hand, activists are highly selective in the information with which they engage (information selection) and excel at sharing or disseminating that information to others (information transmission). Moreover, they often demonstrate confirmation bias and prioritize information that is consistent with their pre-existing views, while they limit exposure to diverse perspectives (information acquisition). Thus, activists tend to exhibit a higher level of information selection and transmission compared to information acquisition (J.-N. Kim & Grunig, 2011).

Investigating whether this association applies to online contexts by comparing predominant communication behaviors across the three public types would offer valuable insights. Accordingly, the final research question is posed:

RQ5: Are certain CAPS communication behaviors associated particularly with specific public types?

5. CAPS on Entertainment Issues

Recent high-profile scandals, legal battles, and societal debates highlight the entertainment industry's role as a major determinant of public discourse and controversy. Its multifaceted nature can be examined from four perspectives: product, experience, culture, and communication (Cavalcanti et al., 2021). As a product, entertainment consists of tangible elements such as plot, characters, and visuals. The experience perspective focuses on the audience's engagement and enjoyment. The cultural perspective situates entertainment within social contexts influenced by norms and values, while the communication perspective considers it a medium to convey messages to audiences. Collectively, these perspectives affect diverse entertainment sectors—media, music, film, gaming, theatre, sports, and tourism—which not only provide enjoyment but also foster cultural discourse (Stein & Evans, 2009). This complexity often leads to heightened public engagement and communication behaviors compared to other issues.

Further, the entertainment industries themselves seek to attract and maintain public attention actively, as this attention serves as a form of social approval (Alber & Heward, 2000) and often translates into financial gains (Bates & Ferri, 2010). Consequently, the industry attempts to lead or follow trending popular issues—current topics that receive extensive media coverage and generate significant public interest and discussion (J.-N. Kim et al., 2012). This endeavor stimulates active communication behavior further on the publics' part. In summary, both entertainment's nature and the entertainment industry's intentional efforts foster active communication behavior, which makes it a suitable context in which to analyze various online communication behaviors of publics within the CAPS framework and observe dynamic changes in public status during periods of controversy.

Given that publics tend to pay more attention to negative rather than positive popular issues (J.-N. Kim et al., 2012), this study focuses on the conflict between the Korean entertainment company ATTRAKT and its

famous idol, Fifty Fifty. This girl group debuted in November 2022 and found success initially with their hit song "Cupid," which achieved significant international fame, including charting on Billboard. The group gained further popularity after releasing their song "Barbie Dreams" with Kaliii for the soundtrack of the movie *Barbie*. However, their fame paradoxically drew widespread attention to their conflicts when the group accused their agency of a contentious contract dispute.

On June 19, 2023, the members filed a suit to suspend their exclusive contracts, alleging that the agency breached its obligations by failing to provide accounting data and neglecting their mental health. Major South Korean news outlets, including KBS, SBS, and MBC, reported this legal action on June 29, 2023. In response, the agency terminated the contracts of three members—Aran, Sio, and Saena—claiming that they had slandered and defamed the agency. In addition, the agency alleged that Warner Music Korea, the primary entity involved, attempted to exploit the artists using unlawful financial and power leverage. In a counteraction, the three former members filed a breach of trust claim against their agency, but lost the case in March 2024. The positions that both parties presented were compelling, which clouded the publics' judgment in determining the truth. The key dates of eight events that could affect public perception of this issue worldwide, together with brief descriptions of each, are shown in Table 1.

6. Method

To examine publics' online communication behaviors, this study used the YouTube API to collect online comments from major Korean news channels—KBS, MBC, and SBS. All three channels reported on the entertainment issue involving ATTRAKT and Fifty Fifty, nearly in real-time. This issue began on April 22, 2024, and continued until approximately August 28, 2024. During this period, eight events triggered significant public communication about the issue (see Table 1). Therefore, the comments analyzed were those posted between 24 hours before and 48 hours after each event from the news videos viewed most on each channel; in some cases, a report about the corresponding issue was available. In addition, comments from a news segment that reported Fifty Fifty's international fame, including their entry onto the Billboard charts, were included as a pre-issue event.

A total of 25,516 comments were collected, each including a user ID, the comment's text, the number of likes, and the publication time. As time passed, more recent news emerged that generated shifts in public opinion that made it increasingly challenging to analyze the entire dataset and led to the decision to apply the CAPS theory to only 10% of the total comments. Consequently, 2,554 comments were sorted based on the largest number of likes, and the top comments were selected for further analysis. The number of comments collected for each event is detailed in Table 1.

In this study, ChatGPT (GPT-4o version) was employed to classify the 2,554 comments into six CAPS dimensions (i.e., information seeking, attending, forefending, permitting, forwarding, and sharing) and three public groups (i.e., aware, active, and activists). The model was provided with definitions, key characteristics, and illustrative examples for each communication dimension (see Table 2) and was prompted to classify several example comments using a zero-shot prompting approach. This approach took advantage of the large language model to accelerate the classification process, and the authors verified the final classifications to ensure accuracy and reliability. The authors intervened in cases of incorrect classifications, identified inaccuracies, and refined the model's responses through repeated adjustments and re-prompts until the

Table 1. Timeline of key events affecting the entertainment issue of Fifty Fifty.

Event #	Date	Description
1	June 29, 2023	All members of Fifty Fifty file a request for a suspension of their exclusive contracts with the agency, ATTRAKT, at the Seoul Central District Court on June 19
2	July 6, 2023	First trial takes place on July 5
3	Aug. 1, 2023	Court refers the dispute between Fifty Fifty and the agency for mediation
4	Aug. 29, 2023	On August 28, Seoul Central District Court rejects the request to suspend the effectiveness of Fifty Fifty's exclusive contracts
5	Sept. 26, 2023	Court approves the request for provisional seizure of copyright fees for the amount embezzled by the outsourcing company, The Givers, on September 25
6	Oct. 24, 2023	Notice of contract termination given to three members
7	Dec. 24, 2023	The agency of Fifty Fifty files a lawsuit for 13 billion KRW against the three former members
8	March 11, 2024	Court confirms that the representative of Fifty Fifty's agency is "not guilty" of embezzlement

responses were consistent with a predetermined "golden answer." This refinement process, which took a month, ensured that the model met the expected classification criteria before it was applied to the full dataset. Following the final classification, two trained coders coded approximately 5% (n = 130) of the 2,554 comments independently for validation. Two rounds of intercoder reliability tests were conducted, and both yielded acceptable reliability scores (ranging from 0.76 to 0.83; Krippendorff, 2019; Neuendorf, 2002). A portion of the comments classified for CAPS dimensions and public classification by ChatGPT is displayed in Tables 2 and 3.

Table 2. Definitions, examples, and characteristics of communication behaviors used in the prompts for comment classification.

	Definition[1]	Characteristics of Comments	Example
Information Forefending	Fending off certain information based on its relevance and value for the problem.	Take selective approaches in dealing with information. Preconceiving certain claims or actions as harmful or beneficial. Belief that the solutions or outcomes to the problem are already clear or anticipated.	"I knew it. Isn't this an obvious outcome?" "I fully agree with this." "How did we end up with a result like this? Nonesense" "This is definitely manipulation."
Information Permitting	Simply accepting any information related to the problem.	Containing expressions of empathy, understanding, and acceptance toward the other person's opinion/perspective.	"That's right." "You've got a point." "Yeah, I agree." "I gotta admit, that's not wrong."
Information Forwarding	Giving information even if no one asks for it.	Providing additional information to others (e.g., a simplified summary, personal thoughts, feelings, new updates).	"That's not it…Can you hear me out for a sec?" "You might not know, but there was actually more to it than this." "Apparently, they're getting a lot of attention overseas right now."

Table 2. (Cont.) Definitions, examples, and characteristics of communication behaviors used in the prompts for comment classification.

	Definition[1]	Characteristics of Comments	Example
Information Sharing	Giving information only if someone asks for it.	Providing an answer to a specific question in response to one's demand or request.	"I'm only saying this because you asked…but the basic is 7 years." "Normally, I do not answer this kind of thing…but it started in February." "It's frustrating how narrowly some fans are viewing the situation. Listen up, Fifty Fifty fans—here's what's really going on"
Information Seeking	Deliberately searching for information to solve problems.	Requesting information (e.g., the identity of an entity, the background of a specific action/event, outcome/status of an event).	"Who are these people that are all over the news lately???" "Please tell me more" "Any idea when the next update coming out?"
Information Attending	Randomly encountering information.	Expressing curiosity or puzzlement. Recognizing current situations or changes.	"Oh. How is this possible?" "No way!" "This is crazy." "Doesn't seem like the old days anymore."

Source: [1] J. E. Grunig (2013).

Table 3. Definitions, examples, and characteristics of publics used in the prompts for comment classification.

	Definition[1]	Characteristics of Comments	Example
Aware Public	Individuals who recognize the issue and may express concerns with an issue but remain relatively inactive due to low engagement or high perceived constraints.	Comments are passive and observational, expressing immediate reactions, mild satisfaction or disappointment without expressing strong engagement.	"I didn't know this before…'' "I see. Something is happening with this group" "Good to know"
Active Public	Individuals with high awareness and engagement in an issue, often displaying strong emotional reactions or detailed information-sharing, but at an individual level.	Comments include detailed opinions or thoughts (suggestions or critiques), information, and a willingness to act on the issue, but without explicitly urging others to act collectively.	"Wake up, girls—This really isn't the time for this. You have to trust the company right now" "Stay strong. Remember, there are people on your side" "Considering what happened last time, they should not get another chance. I will prevent it…whatever it takes"

Table 3. (Cont.) Definitions, examples, and characteristics of publics used in the prompts for comment classification.

	Definition[1]	Characteristics of Comments	Example
Activist	A group of highly aware and engaged individuals with a deep understanding of an issue, who take action or encourage others to act collectively to drive change and achieve specific goals.	Comments explicitly express a willingness to take action or call on others to act, advocating for strong legal measures, industry bans, or other specific collective actions targeting individuals or entities involved.	"Let's boycott this group" "I fully support this idea. Let's start a new audition" "Remember them and ensure they never debut in another K-pop group. They should be banned from ever working in this industry" "All involved scammers should be sent to the court"

Source: [1] J. E. Grunig (2013).

7. Results

The five research questions were formulated to explore the online communication behaviors of various publics. RQ1 focuses on the most prevalent communication behaviors within the CAPS framework during the course of an entertainment issue. The results indicated that information transmission was the action observed most frequently (n = 2,599), followed by information selection (n = 2,388). In contrast, information acquisition was observed far less frequently (n = 821).

Further analysis of each communication behavior's active and passive dimensions showed that information forefending alone was observed more than two thousand times, which made it the most prominent among the six subdimensions (n = 2,386). This was followed closely by information sharing, which was observed nearly two thousand times (n = 1,994). The remaining dimensions occurred fewer than 700 times each. For example, information forwarding was noted 605 times, and information attending was observed 588 times, while information seeking was recorded 233 times. Information permitting was observed the least, with only two comments (see Figure 2).

When the results are organized chronologically according to the eight key events, the ratio data in Figure 3 indicated that information selection remained consistently high throughout these events. Notably, regardless of the number of events reported during those months, the highest level of information selection was observed when Fifty Fifty entered the Billboard charts before the issue erupted. In contrast, following the outbreak of the issue, information acquisition experienced a rapid increase after Event 1 (0.90) and has not declined to pre-issue levels since (0.60). Information transmission also saw a rise after the issue emerged, although not as sharply as information acquisition (before the unfolding of the issue: 0.43, Event 1: 0.53). Subsequently, it dropped (Event 2: 0.28; Event 3: 0.19) despite the occurrence of other events. Only after the news reported that the issue had been resolved in court did information transmission increase again (Event 8: 0.50), nearly to the level observed during Event 1.

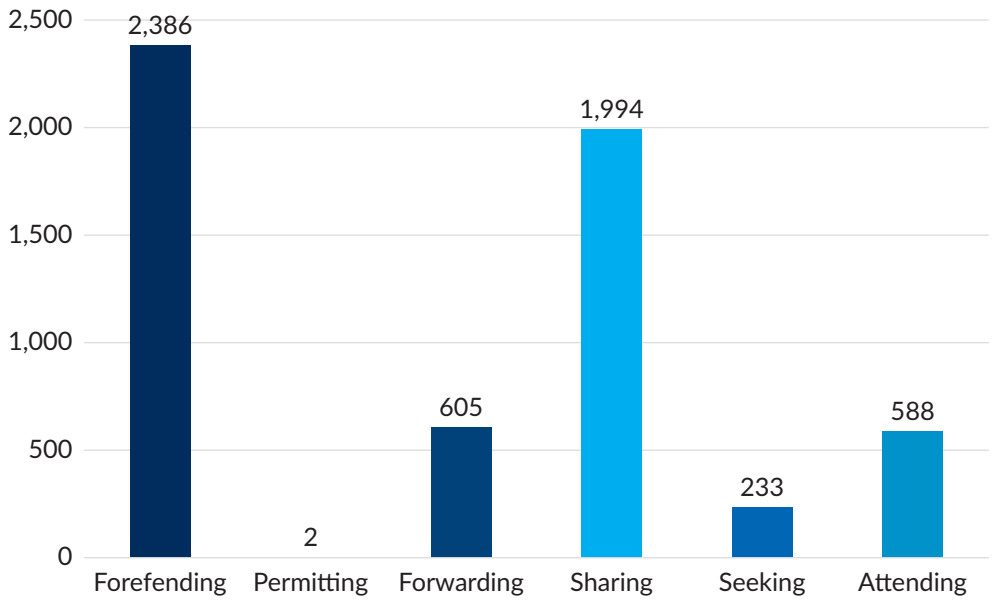


Figure 2. Number of each communication behavior in six dimensions.

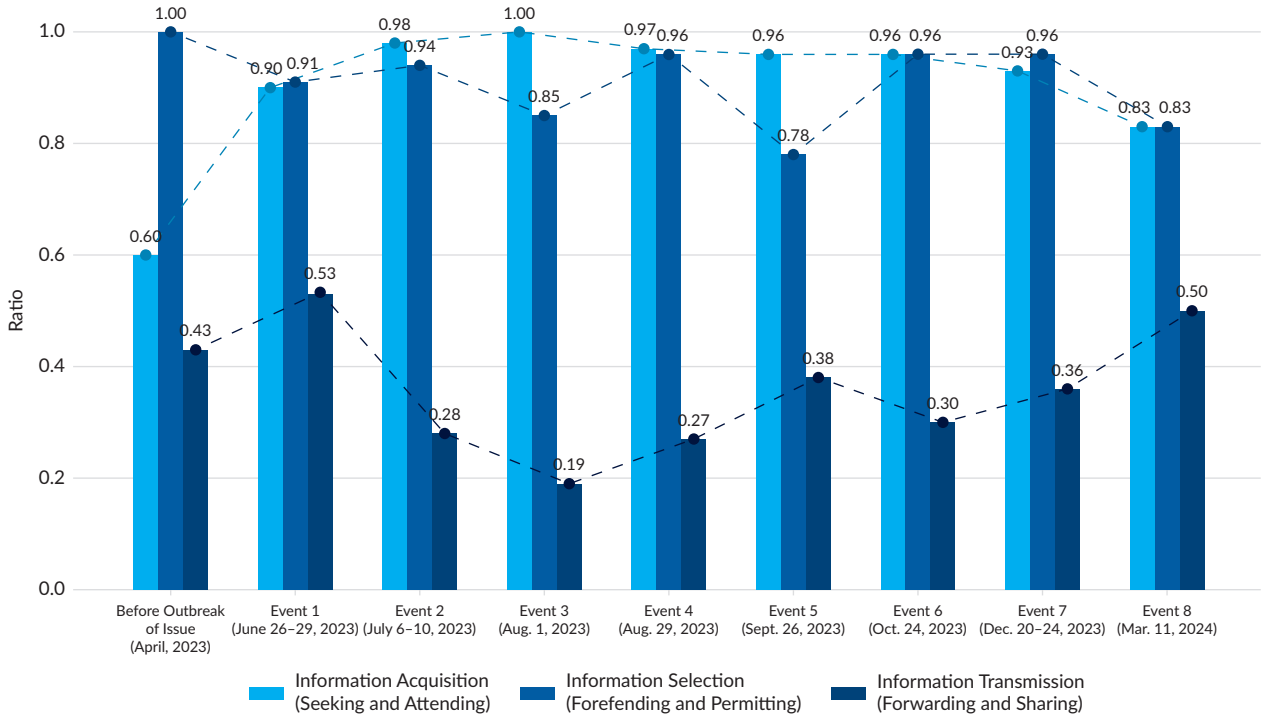


Figure 3. The transition of ratio of each communication behavior.

RQ2 focuses on the combinations of communication behaviors within the CAPS framework. A total of 25 combinations were observed. The combination that occurred most frequently was information forfending and information sharing, which appeared over 1,200 times. The second combination observed most was information forfending paired with information sharing and attending (n = 361), followed by information forfending combined with information forwarding (n = 262) and the combination of information forfending, forwarding, and sharing (n = 182).

In addition, combinations that involved four dimensions were noted, such as: information forfending, forwarding, sharing, and seeking (n = 23); information forfending, forwarding, seeking, and attending (n = 16); information forfending, forwarding, sharing, and attending (n = 11); and information forfending, sharing, attending, and seeking (n = 10).

RQ3 focuses on identifying which public type was observed most frequently during the entertainment issue. As illustrated in Figure 4, the numbers of aware publics (n = 1,027) and active publics (n = 1,034) were quite similar. In contrast, the number of activists was notably lower and comprised less than half the total of the other public types (n = 490).

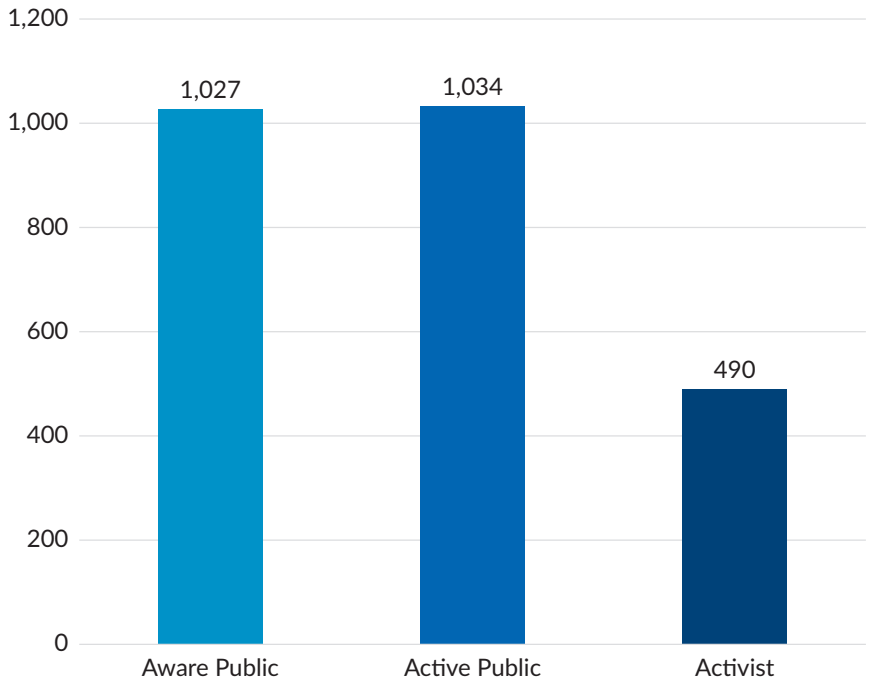


Figure 4. Number of individuals observed for each public type.

RQ4 examines the dynamics of public types during the entertainment issue. As illustrated in Figure 5, the ratio of each public type fluctuated continuously throughout the issue. Notably, the ratio of activists increased gradually from Event 1 (when the issue was announced) to Event 3 (the day the Seoul Central District Court dismissed the request to suspend Fifty Fifty's exclusive contract; Event 1 = 0.08; Event 2 = 0.19; Event 3 = 0.25). In contrast, the ratio of active publics, which was highest at the onset of the issue (0.67), declined at Event 4 (0.27), the day after the Seoul Central District Court rejected the request to suspend the effectiveness of Fifty Fifty's exclusive contracts. However, then it increased gradually and reached 0.34 at Event 6, 0.38 at Event 7, and 0.41 at Event 8 when the court determined that the representative of Fifty Fifty's agency was innocent of embezzlement.

RQ5 asks whether the differences among each dimension hold significant meaning based on public types. A Chi-square test was conducted, and as illustrated in Figure 6, the post-hoc analysis revealed that most p-values were less than 0.05. This finding indicates a significant relation between the communication behaviors identified in the CAPS framework and the various public types, with the exception of the information transmission value between active and activist publics (p = 0.14).

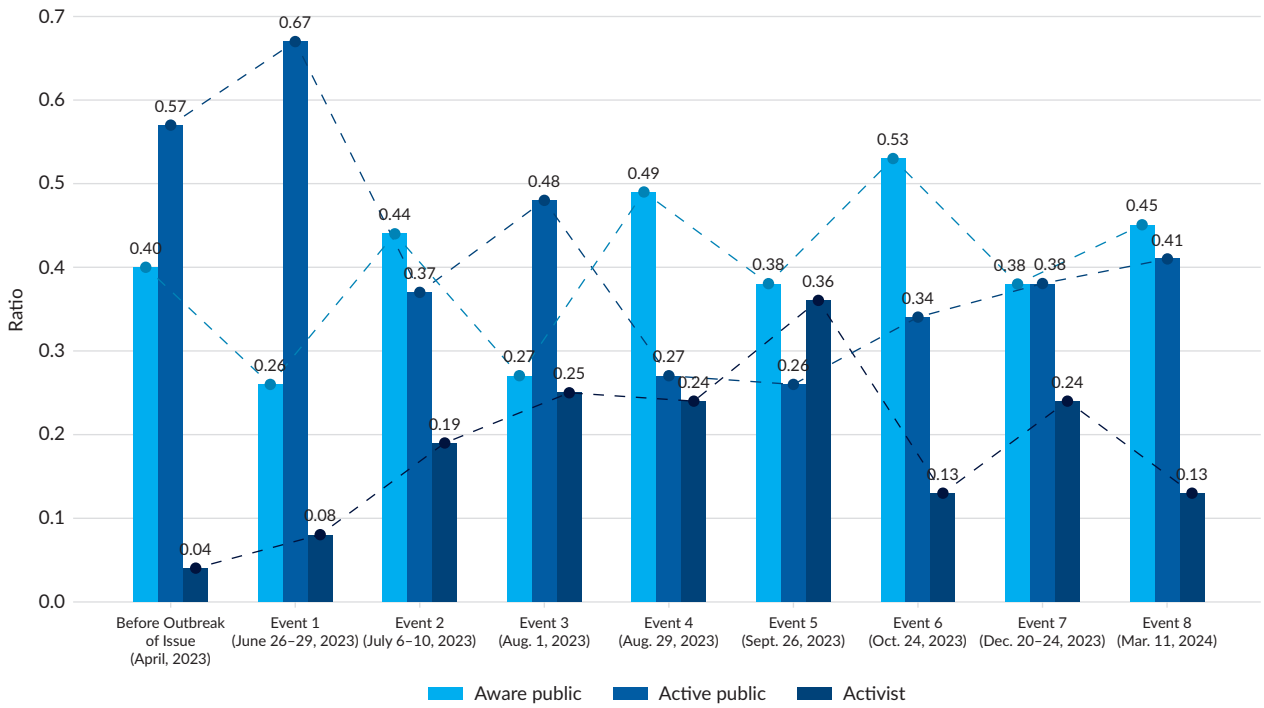


Figure 5. The transition of the ratio of each public type.

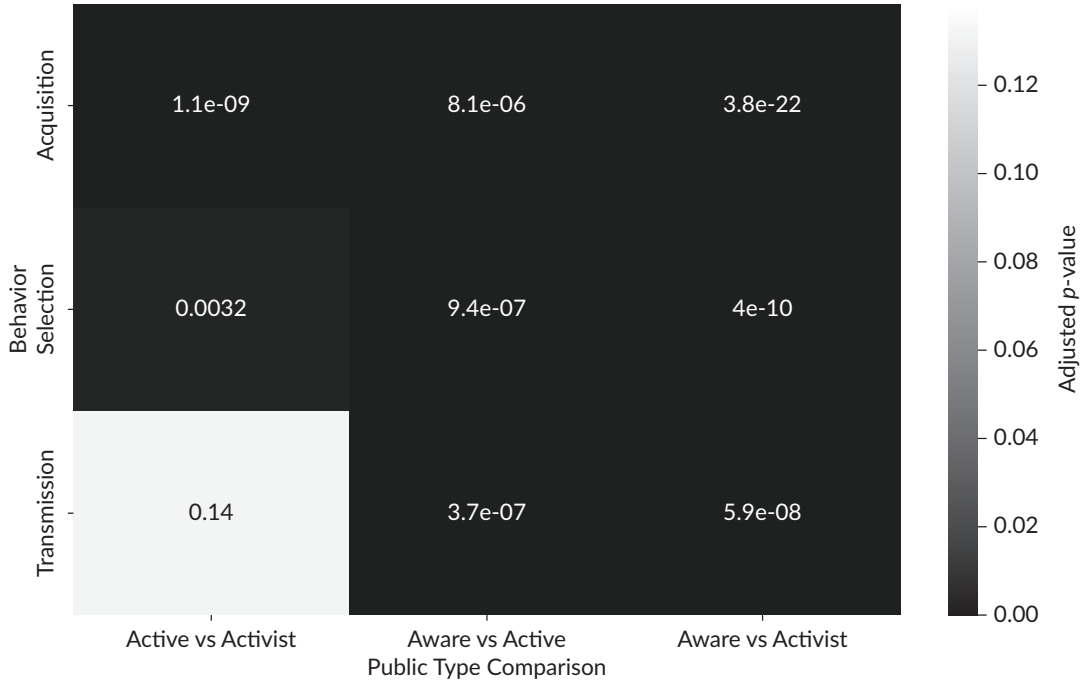


Figure 6. Post-hoc pairwise comparison between CAPS and public types.

Specifically, as shown in Figure 7, information acquisition diminished as the publics' activeness with respect to the issue increased. In addition, both information selection and information transmission, which were at high levels already, continued to rise with increasing public activeness, although this shift was not as dramatic as that of information acquisition.

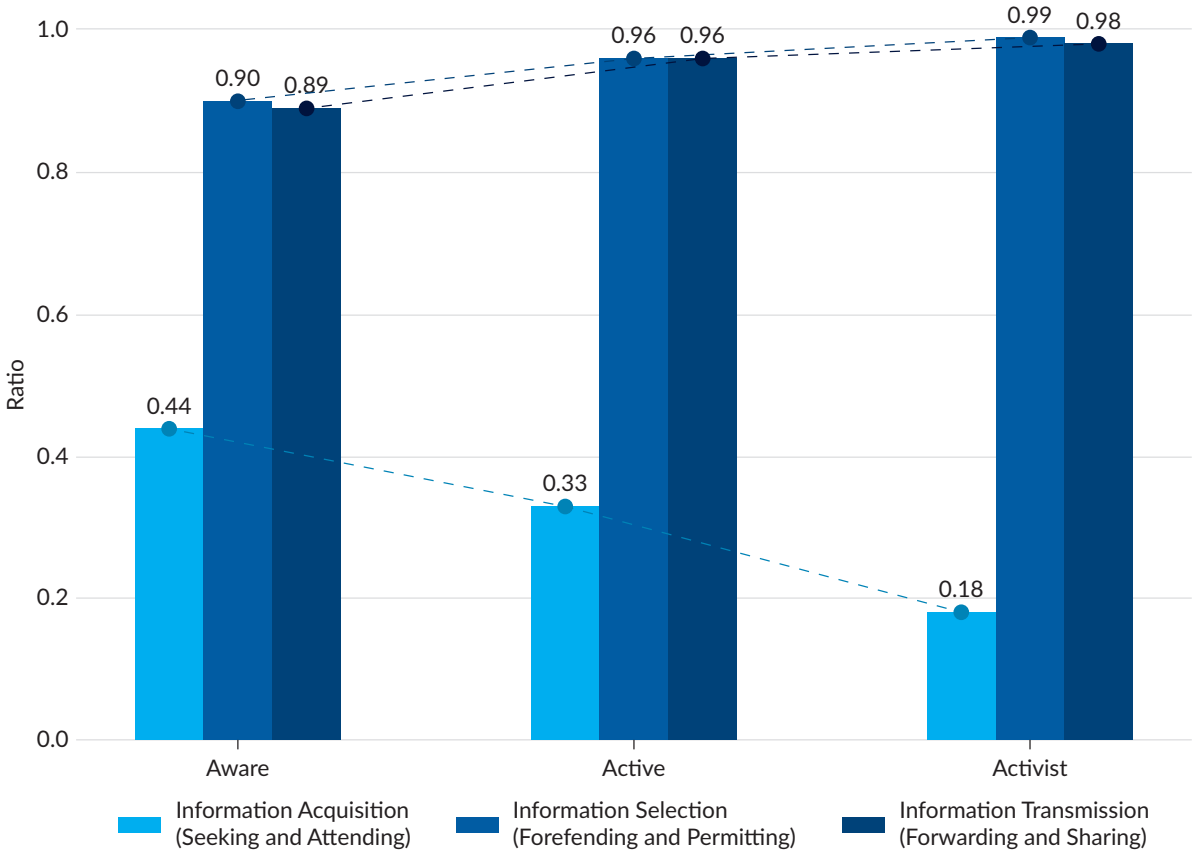


Figure 7. Correlation between online communication behaviors and public types.

8. Conclusion

Despite its significant potential and opportunities, the use of extensive datasets and computing technology in public relations remains largely unexplored (Holtzhausen & Zerfass, 2015; J. E. Grunig, 2023; Wiesenberg et al., 2017). Aldoory and Sha (2007) suggested that reassessing CAPS in the context of advancements in digital media technologies would be valuable. Building on this foundation, this study addressed the limitations of previous CAPS research, which had relied primarily on traditional methods. Then, it proposed adopting data-driven approaches to enhance the understanding of publics' online communication behaviors and their association with different public types.

With respect to communication behaviors, all three information dimensions—selection, transmission, and acquisition—were observed. Active communication behavior is often more pronounced in the context of persistent controversial issues (J.-N. Kim & Grunig, 2011; S. Kim et al., 2015). However, in this study, active communication behavior (i.e., information forfending) was only identified in information selection. In contrast, more passive communication behaviors (i.e., information sharing and information attending) were prevalent in information transmission and acquisition. This prevalence can be attributed to the nature of leaving comments online, which inherently constitutes an information-sharing behavior. On the other hand, the higher incidence of information attending can be explained by this issue's unique context. In previous conflicts between entertainment companies and their celebrities, the companies were often accused of wrongdoing. However, the Fifty Fifty case was exceptional, in that the idol girl group was found

to violate standards. This rare event may have captured public attention more than usual, as reflected in reactions such as "It's the first time I feel sorry for an entertainment agency" and "I can't believe the CEO got really betrayed."

With respect to transition, communication behaviors evolve continuously with each new event related to an issue. This suggests that there is no single fixed communication behavior for an issue, but rather, it is a dynamic process influenced by unfolding events. In addition, the study identified various combinations of communication behaviors. Specifically, individuals utilized between one and four dimensions within a single comment. Among the 2,554 comments analyzed, 75.02% (n = 1,916) contained both information forefending (active) and information sharing (passive). This finding supports the notion that both passive and active dimensions of CAPS coexist, as posited by J.-N. Kim and Krishna (2018). Further, it shows that many publics demonstrate a proactive tendency online to evaluate the value and relevance of information in a given problem-solving context and share information online simultaneously (J.-N. Kim et al., 2010). At the same time, active information selection combined with passive information transmission may also indicate a cautious approach to an oppositional issue, as there is hesitancy to share messages with the broader public about a contentious online discourse.

On the other hand, the findings of this study showed that all three public types were present in online comments: aware, active, and activist. Notably, the number of publics that exhibited characteristics of aware publics was comparable to that of active publics. This finding is consistent with that of Baym's (1996) study and reaffirms that digital platforms lower the threshold for the aware public to engage in online discourse. While organizations have prioritized publics who show actions individually (active public) or collectively (activists public) when an issue arises, it is now essential to give equal attention to the aware public.

In addition, this study highlights public behaviors' dynamism. Initially, active publics' communication behaviors were prominent, but this shifted gradually toward those of the aware and activist publics. Further, while active publics' communication behaviors tended to diminish after an issue was resolved, the numbers of aware and activist publics remained high. This observation is consistent with Aldoory and Grunig's (2012) assertion that activist publics often transition into aware publics when issues receive less attention. Therefore, organizations should focus on engaging active publics at the onset of an issue and embrace the aware and activist publics subsequently as time progresses.

Finally, as illustrated in Figure 1, this study found a close relationship between public types and online communication behaviors. This result reconfirms the notion that as public engagement increases, the tendency to acquire information decreases, and this gap is replaced by enhanced transmission and selection of information.

9. Implications

This study has both academic and practical implications. Academically, it integrates large-scale data and data-driven techniques into a theoretical model for public relations research and enhances its applicability in the digital era. J. E. Grunig (2023) endorsed this approach and Helles and Ørmen (2020) also recognized its value. In particular, this approach presents an empirical visualization of the sequential illustration of motivated information behavior in a problematic situation, as illustrated conceptually in Figure 1. This

visualization provides a clear representation of the communication characteristics associated with each public type and enhances the understanding of how different groups engage with information during certain issues.

This study also found compelling evidence that publics are not static entities. Rather, they exhibited dynamic characteristics by adjusting their status continuously in response to situational changes, as J. E. Grunig (2013) explained. Similar to communication behaviors, even within a single issue, fluctuations in public status are observed depending upon the events that arise. This fluidity challenges traditional research methods that commonly show only a snapshot of publics at a single point in time, and underscores the need for more data-driven approaches that account for the evolving nature of public types over time.

Further, this study emphasizes the urgent need for a comprehensive discussion of the characteristics of aware publics in online environments. Prior studies already consider aware publics as one of the most critical problem solvers, as they are responsible for rapid issue emergence and play a crucial role in the way that issues evolve and subside. Therefore, scholars have suggested that organizations pay close attention to them (J. E. Grunig, 1997; L. A. Grunig et al., 2002; Ni & Kim, 2009). However, beyond this, as the landscape of public engagement online evolves, this demonstrates the increasing importance of reconsidering and potentially revising the current definition of aware publics. The findings suggested that these groups have transitioned into a distinct public type that participates in discussions actively and thereby warrants further academic inquiry into their unique attributes and behaviors.

With respect to practical implications, this study highlights the need for more proactive communication management targeted at aware publics, as publics with these characteristics are observed consistently in large numbers online. Communication is more likely to be effective with aware publics than with active publics, as they are almost at the stage of recognizing a problem but have not yet communicated about it with others (J. E. Grunig, 2013). Therefore, it is advisable for organizations to engage with them first, before they are exposed to incorrect information.

Another practical implication is that this study's findings highlight a significant shift in the methodology used to identify different public groups. Traditional survey methods may no longer be necessary, as computational analysis offers a more effective way to gain insights into publics' online communication behaviors. By training AI, it is even possible to identify the types of comments on an issue that various publics left. This will enable organizations to understand these publics' characteristics better, e.g., whether they are likely to engage in individual or collective actions and implement proactive measures to prevent extreme situations. Thus, organizations are encouraged to invest in hiring data scientists rather than relying on survey companies, and to develop social media monitoring tools that can identify and track key publics and their communication behaviors related to issues effectively (Ampofo et al., 2015). This strategic shift not only streamlines the research process but also enhances the organization's ability to develop informed communication strategies that resonate with its target audiences.

10. Limitations and Future Studies

This study has its limitations, as it focuses on a single public relations model and examines only one entertainment-related issue. Therefore, the findings cannot be generalized to the broader field of

communication. It is recommended to apply this data-driven approach to a wider range of issues, such as health crises, climate change, and national conflicts, which may reveal new dominant communication behaviors and unique combinations of the various CAPS dimensions.

In addition, this study is limited by the fact that it analyzed only 10% of the full dataset. As the comments selected consisted primarily of the comments with the most likes, this selection bias may have influenced the results. Future studies should examine a more representative sample of comments to gain a comprehensive understanding of public communication behaviors.

Further, the fact that the transition of commentators' activity levels over time was not tracked at the individual level. Instead, each comment was analyzed independently and focused solely on the language used in online posts. Consequently, it was not possible to assess whether a commentator became more or less active over the course of events, which could lead potentially to misclassifications of the public's status. For instance, if a commentator did not explicitly express a willingness to take action in an online comment, they were categorized as an aware public, although they might have actually been an active or activist commentator. To address this limitation, future research is encouraged to implement a computational tracking approach to monitor individual commentators over time, which would provide deeper insights into public dynamics and behavioral shifts.

Another limitation of this study is the presence of some inconsistencies in the analysis results that ChatGPT provided. Future research should extend this approach to other generative AI chatbots to assess which model demonstrates greater consistency and fewer discrepancies in classification accuracy.

Acknowledgments

We would like to express our sincere gratitude to the anonymous reviewers for their valuable time, insightful feedback, and expertise, which significantly contributed to the improvement of this work.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024–00456247, No. RS-2023–00218913) and by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI22C0646).

Conflict of Interests

The authors declare no conflict of interests.

References

Alber, S. R., & Heward, W. L. (2000). Teaching students to recruit positive attention: A review and recommendations. *Journal of Behavioral Education*, *10*, 177–204.

Aldoory, L., & Grunig, J. E. (2012). The rise and fall of hot-issue publics: Relationships that develop from media coverage of events and crises. *International Journal of Strategic Communication*, *6*(1), 93–108. https://doi.org/10.1080/1553118X.2011.634866

Aldoory, L., & Sha, B. L. (2007). The situational theory of publics: Practical applications, methodological challenges, and theoretical horizons. In E. L. Toth (Ed.), *The future of excellence in public relations and communication management: Challenges for the next generation* (pp. 339–355). Routledge.

Ampofo, L., Collister, S., O'Loughlin, B., & Chadwick, A. (2015). Text mining and social media: When quantitative meets qualitative, and software meets humans. In P. Halfpenny & R. Procter (Eds.), *Innovations in digital research methods* (pp. 162–192). Sage.

Barnidge, M., Gunther, A. C., Kim, J., Hong, Y., Perryman, M., Tay, S. K., & Knisely, S. (2020). Politically motivated selective exposure and perceived media bias. *Communication Research*, *47*(1), 82–103. https://doi.org/10.1177/0093650217713066

Bates, S., & Ferri, A. J. (2010). What's entertainment? Notes toward a definition. *Studies in Popular Culture*, *33*(1), 1–20.

Baym, N. K. (1996). Agreements and disagreements in a computer-mediated discussion. *Research on Language and Social Interaction*, *29*(4), 315–345.

Bhalla, N. (2022). Examining the impact of issue salience, issue proximity, situational motivation, and communicative behaviors on environmental CSR outcomes. *Sustainability*, *14*(5), Article 2763.

Bolsover, G., & Howard, P. (2019). Chinese computational propaganda: Automation, algorithms, and the manipulation of information about Chinese politics on Twitter and Weibo. Information, *Communication & Society*, *22*(14), 2063–2080. https://doi.org/10.1080/1369118X.2018.1476576

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Cavalcanti, R. C. T., de Aquino, L. M. P., & de Oliveira, H. C. N. (2021). What is entertainment? Propositions of definitions based on product, experience, culture, and communication perspectives. *Canoas*, *11*(1), 1–18. http://doi.org/10.18316/desenv.v11i1.9465

Chang, H. C. H., Pham, B., & Ferrara, E. (2023). Parasocial diffusion: K-pop fandoms help drive COVID-19 public health messaging on social media. *Online Social Networks and Media*, *37*, Article 100267. https://doi.org/10.1016/j.osnem.2023.100267

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, *6*(2), Article e19273. https://doi.org/10.2196/19273

Chih-Ming, C., & Ying-You, L. (2020). Developing a computer-mediated communication competence forecasting model based on learning behavior features. *Computers and Education: Artificial Intelligence*, *1*, Article 100004. https://doi.org/10.1016/j.caeai.2020.100004

Choi, S. (2020). When digital trace data meet traditional communication theory: Theoretical/methodological directions. *Social Science Computer Review*, *38*(1), 91–107. https://doi.org/10.1177/0894439318788618

Chon, M.-G., & Harrell, J. R. (2024). Building bridges for anti-racism activism: Testing situational theory of problem solving and problem chain recognition effect. *Public Relations Review*, *50*(1), Article 102425.

Chon, M.-G., & Park, H. (2021). Predicting public support for government actions in a public health crisis: Testing fear, organization-public relationship, and behavioral intention in the framework of the situational theory of problem solving. *Health Communication*, *36*(4), 476–486. https://doi.org/10.1080/10410236.2019.1700439

Chon, M.-G., Tam, L., Lee, H., & Kim, J.-N. (2023). Situational theory of problem solving (STOPS): A foundational theory of publics and its behavioral nature in problem solving. In C. Botan & E. Sommerfeldt (Eds.), *Public relations theory III* (pp. 58–76). Routledge.

Chon, M.-G., Xu, L., Kim, J., & Liu, J. (2022). Understanding active communicators on the food safety issue: Conspiratorial thinking, organizational trust, and communicative actions of publics in China. *American Behavioral Scientist*, *69*(2), 168–186. https://doi.org/10.1177/00027642221118284

Dewey, J. (1954). *The public and its problems*. Ohio University Press. (Original work published 1927)

Dozier, D. M., & Ehling, W. P. (2013). Evaluation of public relations programs: What the literature tells us about their effects. In J. E. Grunig (Ed.), *Excellence in public relations and communication management* (pp. 159–184). Routledge.

Dozier, D. M., Grunig, L. A., & Grunig, J. E. (2013). *Manager's guide to excellence in public relations and communication management*. Routledge.

Elberse, A. (2013). *Blockbusters: Hit-making, risk-taking, and the big business of entertainment*. Henry Holt and Company.

Freelon, D., McIlwain, C., & Clark, M. (2016). *Beyond the hashtags: #Ferguson, #BlackLivesMatter, and the online struggle for offline justice* (Research Report). Center for Media & Social Impact. http://cmsimpact.org/blmreport

Gil de Zuniga, H., & Diehl, T. (2017). Citizenship, social media, and big data: Current and future research in the social sciences. *Social Science Computer Review*, *35*(1), 3–9.

Grunig, J. E. (1978). Defining publics in public relations: The case of a suburban hospital. *Journalism Quarterly*, *55*(1), 109–124. https://doi.org/10.1177/107769907805500115

Grunig, J. E. (1989). Sierra Club study shows who become activists. *Public Relations Review*, *15*(3), 3–24.

Grunig, J. E. (1997). A situational theory of publics: Conceptual history, recent challenges and new research. In D. Moss, T. MacManus, & D. Vercic (Eds.), *Public relations research: An international perspective* (pp. 3–46). International Thompson Business Press.

Grunig, J. E. (2013). *Excellence in public relations and communication management*. Routledge.

Grunig, J. E. (2023). Public relations, social inclusion, and social exclusion. *Journalism & Communication Monographs*, *25*(2), 90–108.

Grunig, J. E., & Kim, J.-N. (2017). Publics approaches to segmentation in health and risk messaging. In R. Parrott (Ed.), *Encyclopedia of health and risk message design and processing*. Oxford University Press.

Grunig, L. A., Grunig, J. E., & Dozier, D. M. (2002). *Excellent public relations and effective organizations: A study of communication management in three countries*. Lawrence Erlbaum.

Guo, L., & Vargo, C. (2015). The power of message networks: A big-data analysis of the network agenda setting model and issue ownership. *Mass Communication and Society*, *18*(5), 557–576.

Helles, R., & Ørmen, J. (2020). Big data and explanation: Reflections on the uses of big data in media and communication research. *European Journal of Communication*, *35*(3), 290–300.

Hirsch, T. (2011). More than friends: Social and mobile media for activist organizations. In M. Foth, L. Forlano, C. Satchell, & M. Gibbs (Eds.), *From social butterfly to engaged citizen: Urban informatics, social media, ubiquitous computing, and mobile technology to support citizen engagement* (pp. 135–150). MIT Press.

Holtzhausen, D. R., & Zerfass, A. (2015). Strategic communication: Opportunities and challenges of the research area. In D. R. Holtzhausen & A. Zerfass (Eds.), *The Routledge handbook of strategic communication* (pp. 3–17). Routledge.

Howard, P. N., Duffy, A., Freelon, D., Hussain, M., Mari, W., & Mazaid, M. (2011). *Opening closed regimes: What was the role of social media during the Arab Spring?* (Working Paper 2011.1). Project on Information Technology and Political Islam.

Kent, M. L., & Saffer, A. J. (2014). A Delphi study of the future of new technology research in public relations. *Public Relations Review*, *40*(3), 568–576. https://doi.org/10.1016/j.pubrev.2014.02.008

Kim, H. J., & Hong, H. (2021). Predicting communication behaviors in the Covid-19 pandemic: Integrating the role of emotions and subjective norms into the situational theory of problem solving (STOPS) framework. *Health Communication*, *37*(13), 1640–1649. https://doi.org/10.1080/10410236.2021.1911399

Kim, J.-N. (2006). *Communicant activeness, cognitive entrepreneurship, and a situational theory of problem solving* [Unpublished doctoral dissertation]. University of Maryland.

Kim, J.-N., & Grunig, J. E. (2011). Problem Solving and Communicative Action: A Situational Theory of Problem Solving. *Journal of Communication*, *61*(1), 120–149. https://doi.org/10.1111/j.1460-2466.2010.01529.x

Kim, J.-N., Grunig, J. E., & Ni, L. (2010). Reconceptualizing the communicative action of publics: Acquisition, selection, and transmission of information in problematic situations. *International Journal of Strategic Communication*, *4*(2), 126–154.

Kim, J.-N., & Krishna, A. (2014). Publics and Lay Informatics: A Review of the Situational Theory of Problem Solving. *Annals of the International Communication Association*, *38*(1), 71–105. https://doi.org/10.1080/23808985.2014.11679159

Kim, J.-N., & Lee, S. (2014). Communication and cybercoping: Coping with chronic illness through communicative action in online support networks. *Journal of Health Communication*, *19*(7), 775–794. https://doi.org/10.1080/10810730.2013.864724

Kim, J.-N., Ni, L., Kim, S. H., & Kim, J. R. (2012). What makes people hot? Applying the situational theory of problem-solving to hot-issue publics. *Journal of Public Relations Research*, *24*(2), 144–164. https://doi.org/10.1080/1062726X.2012.626133

Kim, J.-N., Shen, H., & Morgan, S. E. (2011). Information behaviors and problem chain recognition effect: Applying situational theory of problem solving in organ donation issues. *Health Communication*, *26*(2), 171–184. https://doi-org.ezproxy.lib.ou.edu/10.1080/10410236.2010.544282

Kim, S., Kim, J.-N., Tam, L., & Kim, G. T. (2015). Inquiring into activist publics in chronic environmental issues: Using the mutual-gains approach to a deadlock. *Journal of Public Affairs*, *15*(4), 404–422.

Kim, Y. (2016). Understanding publics' perception and behaviors in crisis communication: Effects of crisis news framing and publics' acquisition, selection, and transmission of information in crisis situations. *Journal of Public Relations Research*, *28*(1), 35–50. https://doi.org/10.1080/1062726X.2015.1131697

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th Ed.). Sage. https://doi.org/10.4135/9781071878781

Krishna, A. (2018). Poison or prevention? Understanding the linkages between vaccine-negative individuals' knowledge deficiency, motivations, and active communication behaviors. *Health Communication*, *33*(9), 1088–1096. https://doi.org/10.1080/10410236.2017.1331307

Lee, E. W., & Yee, A. Z. (2020). Toward data sense-making in digital health communication research: Why theory matters in the age of big data. *Frontiers in Communication*, *5*, Article 11. https://doi.org/10.3389/fcomm.2020.00011

Lee, H. L. (2023). Theory-enhanced automation of the digital publics' relationship assessments [Unpublished doctoral dissertation]. The University of Oklahoma. https://shareok.org/items/b53a9b1a-586c-4f8c-a5b2-e0726b824cbe

Lünenborg, M. (2019). Chapter 3. Affective publics: Understanding the dynamic formation of public articulations beyond the public sphere. In A. Fleig & C. von Scheve (Eds.), *Public spheres of resonance* (pp. 29–48). Routledge.

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33.

Murthy, D. (2017). The ontology of tweets: Mixed methods approaches to the study of Twitter. In L. Sloan & A. Quan-Haase (Eds.), *The Sage handbook of social media research methods* (pp. 559–572). Sage.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Sage.

Neuman, W. R., Guggenheim, L., Mo Jang, S., & Bae, S. Y. (2014). The dynamics of public attention: Agenda-setting theory meets big data. *Journal of Communication*, *64*(2), 193–214. https://doi.org/10.1111/jcom.12088

Ni, L., & Kim, J.-N. (2009). Classifying publics: Communication behaviors and problem-solving characteristics in controversial issues. *International Journal of Strategic Communication*, *3*(4), 217–241. https://doi.org/10.1080/15531180903221261

Nussbaum, M. (2013). *Political emotions: Why love matters for justice*. Harvard University Press.

Paget, E. H. (1929). Sudden changes in group opinion. *Social Forces*, *7*(3), 438–444.

Park, D., Lee, H., & Jeong, S. H. (2022). Production and correction of misinformation about fine dust in the Korean news media: a big data analysis of news from 2009 to 2019. *American Behavioral Scientist*, *69*(2), 219–239. https://doi.org/10.1177/00027642221118287

Parks, M. R. (2014). Big data in communication research: Its contents and discontents. *Journal of Communication*, *64*(2), 355–360.

Rains, S. A. (2020). Big data, computational social science, and health communication: A review and agenda for advancing theory. *Health Communication*, *35*(1), 26–34.

Roh, S., & Oh, H. J. (2021). Toward a holistic approach for nuanced public segmentation: Social vigilantism and the situational theory of problem solving (STOPS). *Journal of Public Relations Research*, *33*(2), 106–129. https://doi.org/10.1080/1062726X.2021.2007929

Shen, H., Xu, J., & Wang, Y. (2019). Applying situational theory of problem solving in cancer information seeking: A cross-sectional analysis of 2014 HINTS survey. *Journal of Health Communication*, *24*(2), 165–173. https://doi.org/10.1080/10810730.2019.1587111

Song, L., Li, R. Y. M., & Wareewanich, T. (2023). The cultivation effect of architectural heritage YouTube videos on perceived destination image. *Buildings*, *13*(2), Article 508.

Stein, A., & Evans, B. B. (2009). *An introduction to the entertainment industry*. Peter Lang.

Tilson, D. J. (2016). Entertainment publicity and public relations. In T. Watson (Ed.), *North American perspectives on the development of public relations: Other voices* (pp. 81–96). Palgrave Macmillan.

Vargo, C. J., & Guo, L. (2017). Networks, big data, and intermedia agenda setting: An analysis of traditional, partisan, and emerging online US news. *Journalism & Mass Communication Quarterly*, *94*(4), 1031–1055.

Weiner, M., & Kochhar, S. (2016). *Irreversible: The public relations big data revolution* (IPR White Paper). Institute for Public Relations.

Wiesenberg, M., Zerfass, A., & Moreno, A. (2017). Big data and automation in strategic communication. *International Journal of Strategic Communication*, *11*(2), 95–114.

Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. https://doi.org/10.18608/jla.2015.22.2

Xu, X., Li, H., & Shan, S. (2021). Understanding the health behavior decision-making process with situational theory of problem solving in online health communities: The effects of health beliefs, message source credibility, and communication behaviors on health behavioral intention. *International Journal of Environmental Research and Public Health*, *18*(9), Article 4488. https://doi.org/10.3390/ijerph18094488

About the Authors

Sunha Yeo is a doctoral candidate at Gaylord College and a senior researcher at the Debiasing and Lay Informatics Lab at the University of Oklahoma. She has 11 years of international PR experience and will join the A. Q. Miller School of Kansas State University. Her research focuses on strategic communication and digital publics.

Joohee Kim is a doctoral candidate at the Ulsan Institute of Science and Technology and a senior researcher at the Human–AI Interaction and Visualization Lab. Her research focuses on visual analytics, data visualization, and human–computer interaction, integrating AI-driven techniques to enhance analytical and interactive systems.

Juwon Kim is an MS student at the Graduate School of Artificial Intelligence, Pohang University of Science and Technology (POSTECH). Her research focuses on data analysis for natural language processing and on time series forecasting in the financial domain.

Sungahn Ko received a doctoral degree in electrical and computer engineering from Purdue University in 2014. He is an associate professor in the School of Computer Science and Engineering, POSTECH, in Pohang, South Korea. His research interests include visual analytics, machine learning, and computational social science.

Media and Communication
2025 • Volume 13 • Article 9677
https://doi.org/10.17645/mac.9677

ARTICLE

Open Access Journal

Ideology and Policy Preferences in Synthetic Data: The Potential of LLMs for Public Opinion Analysis

Keyeun Lee [1], Jaehyuk Park [2], Suh-hee Choi [3], and Changkeun Lee [2]

[1] Department of Communication, Seoul National University, South Korea
[2] KDI School of Public Policy and Management, South Korea
[3] Department of Geography, Kyung Hee University, South Korea

Correspondence: Changkeun Lee (cklee@kdischool.ac.kr)

Submitted: 21 November 2024 **Accepted:** 5 March 2025 **Published:** 22 May 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

This study investigates whether large language models (LLMs) can meaningfully extend or generate synthetic public opinion survey data on labor policy issues in South Korea. Unlike prior work conducted on people's general sociocultural values or specific political topics such as voting intentions, our research examines policy preferences on tangible social and economic topics, offering deeper insights for news media and data analysts. In two key applications, we first explore whether LLMs can predict public sentiment on emerging or rapidly evolving issues using existing survey data. We then assess how LLMs generate synthetic datasets resembling real-world survey distributions. Our findings reveal that while LLMs capture demographic and ideological traits with reasonable accuracy, they tend to overemphasize ideological orientation for politically charged topics—a bias that is more pronounced in fully synthetic data, raising concerns about perpetuating societal stereotypes. Despite these challenges, LLMs hold promise for enhancing data-driven journalism and policy research, particularly in polarized societies. We call for further study into how LLM-based predictions align with human responses in diverse sociopolitical settings, alongside improved tools and guidelines to mitigate embedded biases.

Keywords

AI-generated text; ChatGPT; large language models; news media; policy preferences; public opinions

1. Introduction

With the advancement of large language models (LLMs), there has been growing interest in how well AI-generated factual and opinionated text can resemble human output among academicians and journalists.

AI-generated factual text has begun affecting routine human tasks such as combining and summarizing information (Glickman & Zhang, 2024). A recent focus of academic studies also involves how LLMs can generate text that mirrors the content, style, tone, and grammatical traits of specific demographic groups (Argyle et al., 2023; Gerosa et al., 2024; Harding et al., 2024; Sun et al., 2024). Beyond academic circles, acknowledging the benefits in terms of efficiency and timeliness, journalists are also testing LLMs to gauge public opinion on policy issues. Traditionally, news media have relied on opinion surveys to track public sentiment and shape discourse by presenting "snapshots" of crucial social, political, and economic issues. However, such conventional methods face challenges—they are costly, time-consuming, and often lag behind rapidly changing events. Fast, cost-efficient AI-based approaches are growing in appeal, especially as media outlets strive to capture evolving public attitudes more promptly. For example, *The Atlantic*'s tech reporter experimented by prompting ChatGPT to act as different archetypal voters ("MAGA zealots," "suburban moms," etc.), discovering that a "40-year-old conservative man from rural Ohio" persona produced "vividly partisan" rhetoric (Desai, 2023). This demonstrates how easily the AI can mirror extreme opinions when instructed to adopt a specific viewpoint.

Meanwhile, some news media are considering AI's capacity to generate focus-group-style responses to policy proposals. So far, most experiments remain in a testing phase—not yet having formal "AI-generated polls" in mainstream outlets—and reputable publications like *The New York Times* and BBC have announced cautionary guidelines. The BBC, for instance, explicitly forbids using LLMs (BBC, 2025). Still, the intrigue surrounding AI-based public opinion simulations persists, underscoring the tension between editorial caution and innovative ambition.

Despite extensive scholarly warnings and recent journalism guidelines urging a careful approach, global news outlets increasingly rely on LLMs to produce political comparisons and predictive analyses. For instance, in South Korea, the conservative newspaper *Chosun Ilbo* published an article asking ChatGPT about North Korean leader Kim Jong Un (G. Y. Kim, 2023), while the progressive newspaper *Hankyoreh* published a column inquiring about potential job destruction caused by LLMs (S. Lee, 2023). Such rapidly produced, AI-crafted stories appealed to readers with strong political views—whether supportive or critical of particular leadership styles—yet often lack verifiable data or current context. Scholars emphasize that LLMs cannot replace expert analysis or rigorous reporting; nonetheless, the allure of fast, eye-catching AI-driven content widens the gap between recommended guidelines and actual newsroom practice.

This article examines how LLMs can complement traditional opinion surveys in South Korea, where LLMs have already gained wide traction, particularly on polarized social and political issues. We begin by considering a scenario in which a new topic—or a fresh angle on an existing one—demands timely insights. In this setting, we test whether LLMs, provided individual-level survey data, can effectively predict public sentiment toward that emerging issue. We then turn to whether LLMs can generate synthetic data that closely reflect real-world survey distributions, thus potentially aiding researchers and journalists in predicting trends and outcomes with greater detail and at a lower cost.

In doing so, we pay particular attention to labor policy, where South Korea's left- and right-leaning groups hold notably divergent views. While minimum wage controversies are highly politicized, extending the retirement age typically engenders far less ideological friction. Recent studies such as Rozado (2024) caution that LLMs may carry political biases, but few have examined how these biases vary between politicized and

non-politicized questions. Building on the literature on ideological polarization (Caughey et al., 2019; McCarty et al., 2016), we investigate whether LLM-driven responses amplify partisan rifts on hot-button labor issues like the minimum wage compared to relatively neutral topics such as retirement age. Our preliminary findings reveal that AI-driven outputs can overestimate the effect of users' ideological orientation, underscoring the need for vigilance when employing LLMs to extend or create new data. By highlighting these potential pitfalls, we aim to offer guidance for researchers and media outlets considering AI-based methods to gauge public sentiment—particularly in highly polarized environments.

This article contributes to the literature in several ways. First, unlike many studies that center on general sociocultural topics, including people's values, or specific political perceptual or behavioral topics, such as voting intentions, we focus on specific policy issues. By examining how LLMs capture—or misrepresent—public preferences on concrete policies, our research supports a shift toward policy-oriented journalism rather than surface-level partisan divides.

Second, we situate our analysis within the South Korean context, where political polarization has intensified even though actual differences in public policy preferences often remain modest (Cheong & Haggard, 2023). While political orientation is generally presumed to be a strong predictor of policy preferences, this assumption warrants re-examination in South Korea, where ideological identities dominate public discourse but may not map neatly onto specific policy positions. This fills a critical gap in the predominantly North American- and European-focused literature, providing a fresh perspective on the complex relationship between political identity and policy choices.

Finally, the rising prevalence of AI- and human-generated text in the media has already altered how people consume information (Yang & Menczer, 2024). Some AI-generated content has been flagged as malicious, raising concerns about its distorting effects on the data ecosystem. Blending opinionated comments and news text can confuse audiences—particularly older generations who may be less familiar with AI technologies (Moravec et al., 2024). Furthermore, the problem of AI "hallucination," where models produce inaccurate or misleading content, underscores the need for better training data, greater transparency, and continued ethical oversight (Gerosa et al., 2024; Patel, 2024). We contribute to the extant body of knowledge by showing how these challenges also apply to public opinion data, reinforcing the importance of careful validation and critical scrutiny when using LLM-generated data in this field.

2. Literature Review

2.1. AI in Journalism: Opportunities and Risks

Recent scholarship underscores both the significant promise and potential pitfalls of integrating advanced language models into newsrooms. Caswell (2024) shows how media organizations have moved from simpler automated outputs, such as basic data-driven stories, to generative AI, which has resulted in marked gains in productivity and the possibility of targeted, audience-specific content. However, the author also urges the creation of dedicated "AI editors" who can identify bias and maintain rigorous journalistic standards, pointing out that traditional gatekeeping roles might weaken if automation is not guided by specialized oversight. Pan et al. (2023) highlight the risk of widespread "misinformation pollution," as these models can generate convincing but false or biased statements in large volumes, often surpassing both human and

automated fact-checkers. Fletcher and Nielsen (2024) report public skepticism toward AI-driven stories, with fewer than one in ten people willing to pay extra for content written by machines. Other researchers also doubt if AI systems can consistently meet professional editorial benchmarks, expressing worries about accuracy and transparency. (Binz & Schulz, 2023; Brigham et al., 2024). Brigham et al. (2024) reveal ethical concerns, such as journalists inadvertently exposing confidential or copyrighted materials to LLMs, and describe minimal human revision of AI drafts, which can let factual errors slip through.

Collectively, these findings emphasize that LLMs can streamline production and even personalize news for different readers, but they also risk bias, misinformation, and a gradual decline in editorial integrity when used uncritically. For our study, these dynamics underscore the importance of robust oversight, ethical safeguards, and editorial clarity regarding AI-generated content—principles that likewise apply to the use of LLMs in public opinion research.

2.2. Using AI to Extend Public Opinion Data

Advances in LLMs have stimulated a surge of research on whether AI-generated text can effectively replicate or extend public opinion data. Some applications focus on AI-driven polling, where synthetic survey responses can reduce costs and time relative to conventional polls (Berger et al., 2024). While these pilot initiatives show promise in forecasting policy preferences and voter behavior, they also amplify longstanding concerns about algorithmic bias and the reliability of training data (Berger et al., 2024; Kennedy et al., 2022). Other studies build on this foundation by examining the capacity of LLMs to not only generate textual responses but also to capture more nuanced socioeconomic and political expressions (Argyle et al., 2023; Feng et al., 2023).

A key focus of recent work is conditioning the language model with demographic or ideological cues. Researchers have shown that ChatGPT-based systems can yield textual outputs resembling specific subgroups' ways of expressing political or socioeconomic attitudes (Amirova et al., 2024; Argyle et al., 2023). This approach purports to capture the probabilistic distributions of real human responses under given demographic or ideological constraints. Proponents argue that such "silicon subjects" (Argyle et al., 2023; Sun et al., 2024) offer an avenue to explore large-scale, fine-grained variations in opinion without incurring the high costs and potential sampling biases of traditional surveys (Gerosa et al., 2024; Hutson, 2023; Pachot & Petit, 2024).

Despite these potential benefits, several researchers emphasize the concept of algorithmic fidelity, the extent to which an LLM accurately reflects the attitudes and opinions of a targeted human subgroup (Amirova et al., 2024; Argyle et al., 2023). The premise is that if sufficiently detailed conditioning prompts are provided (covering age, gender, ideology, socioeconomic status, and so forth) the generated text will probabilistically mirror human responses in real-world settings. However, as Amirova et al. (2024) note, there can be inconsistencies or discrepancies between what the model produces and actual subgroup attitudes, especially if the model's training data lack diversity or reflect outdated cultural contexts. A related challenge is persona simulation, where researchers prompt the model to "act as if" it possesses certain cognitive limitations or ideological commitments (Aher et al., 2023; Gerosa et al., 2024; Kotek et al., 2023; Milička et al., 2024). While these techniques may yield lifelike responses, they also risk producing hyper-accuracy distortion (Amirova et al., 2024) or "correct answer" effects (Park et al., 2024). In some cases,

the model's factual knowledge overrides attempts to simulate uncertainty or misinformation, resulting in text that is too well-informed and thus unrepresentative of real human respondents.

A crucial limitation arises from sociopolitical biases embedded in LLMs (Aher et al., 2023; Feng et al., 2023; Kotek et al., 2023; Park et al., 2024). Because models train on data that may overrepresent certain demographic groups or partisan content, they often exhibit skewed outcomes when asked to mimic diverse populations (Aher et al., 2023). Researchers highlight issues such as gender bias (Kotek et al., 2023) and the tendency to produce uniformly "safe" responses that do not capture the full range of ideologically extreme viewpoints (Park et al., 2024). Dillion et al. (2023) and Hutson (2023) note that LLMs cannot replicate the intricate psychological or social processes underlying real human behavior, such as lying, changing opinions over time, or experiencing moral dilemmas. Some studies also observe distortions in the model's adherence to specified personas or instructions. Milička et al. (2024) document instances where advanced ChatGPT-based systems fail to limit or "downplay" their own cognitive abilities when prompted with less-informed personas, thus providing responses that are more coherent and factual than a human subject with similar constraints might. In other words, even a well-tuned model may inadvertently slip into more knowledgeable or accurate modes of response, invalidating efforts to simulate ignorance, confusion, or bias.

Recent literature thus presents a mixed picture: On the one hand, LLMs open up avenues for low-cost, high-volume simulations of public opinion (Argyle et al., 2023; Gerosa et al., 2024); on the other hand, they introduce new methodological and ethical complications. Issues like overly uniform outputs, underestimation of extreme viewpoints, and reliance on training data that reflect outdated or biased cultural contexts limit the reliability of LLM-generated "silicon subjects" (Harding et al., 2024; Park et al., 2024). Harding et al. (2024) argue that language models cannot easily adapt to the rapid cultural or social shifts occurring in real human populations—especially as new moral standards and political events reshape beliefs in ways that may not be reflected in static training corpora. From the perspective of policy analysts, political scientists, and media organizations, AI-based polling and ChatGPT simulations remain attractive for exploring emergent trends or investigating hypothetical scenarios (Sun et al., 2024). However, to fully harness these tools, researchers must actively mitigate biases, continually validate model outputs against real-world data, and disclose the AI's role in the generation process. In doing so, they may foster more nuanced and trustworthy insights while recognizing that LLMs cannot yet replace the complexity and variability inherent in genuine human attitudes (Dillion et al., 2023; Hutson, 2023).

2.3. The South Korean Policy and Media Context

An important underlying assumption in comparing the results from human responses with those from LLM-generated responses is that a refined, particular demographic group with a specific expression of political positioning presents relatively stable attitudes toward social issues such as population policies (Han & Ding, 2024). It is achieved by LLM's connection of the defined profile to the text generation coherent with the trained information (Gerosa et al., 2024).

As most previous studies have been conducted in English-speaking or Western societies, especially the US, where political ideological identities and those of political parties have been relatively stabilized, the input of variables associated with political parties, ideological inclinations, and sociodemographic characteristics can more readily predict individuals' political positions (Argyle et al., 2023). Despite some common notions of

the spectrum of left and right, totalitarian–authoritarian, and libertarian, variation in the way such terms are understood and utilized may occur because the conditions that shape the spectrum consist of more than one criterion and because of region-specific contexts (Gindler, 2021). Similarly, research on US political orientations suggests relatively consistent liberal–conservative divides, reflecting distinct moral intuitions and responses to uncertainty (K. R. Kim & Kang, 2013).

In contrast, South Korea's party system is characterized by less clearly demarcated ideological boundaries, weakening the representational ties between specific social groups and political organizations (Cheong & Haggard, 2023; Cho et al., 2019). Parties often shift their policy positions or alliances and do not consistently anchor themselves to a stable ideological identity (Cheong & Haggard, 2023). Nevertheless, a trend of polarization has gained traction in recent years, with supporters of the left party embracing more liberal stances and right-leaning constituents identifying as more conservative (Cheong & Haggard, 2023). One illustrative example is the minimum wage debate, which has been highly politicized in South Korea. Although the left typically presents minimum wage hikes as a remedy for inequality and a means to protect vulnerable workers, the right emphasizes concerns about potential burdens on small and medium-sized enterprises, arguing for greater labor market flexibility instead. Such divergences demonstrate how certain issues—like the minimum wage—can become rallying points for entrenched beliefs that map onto left and right orientations, even in a context where overall party identities remain fluid.

Given this complex ideological landscape, assumptions about the stability and predictability of political attitudes in South Korean contexts may not hold to the same extent as in the US or other Western settings. For LLMs, generating coherent profiles based on South Korean demographic and ideological variables can, therefore, be more challenging, potentially yielding more variance in simulated opinions (Cheong & Haggard, 2023). This highlights why cross-national studies of LLM-generated responses need to consider regional political traditions and the strength (or weakness) of partisan identities when interpreting outcomes—particularly on politicized topics such as the minimum wage.

3. Methods

3.1. Data

This study assesses how LLMs, with a particular focus on ChatGPT, can extend existing public opinion survey data and generate synthetic data that resembles the original—ultimately informing policy analysis and news reporting. To achieve our research objectives, we draw on survey data collected by one of the authors in March 2024 for policy analysis (C. Lee et al., 2025). The survey sampled an online panel of 2,000 respondents, designed to be representative of the South Korean population in terms of age, gender, and regional distribution. It covered policy preferences across five domains: macroeconomics, diplomacy, labor, environment, and population policies.

We center our analysis on labor policies, as they represent a critical source of political contention in South Korea—particularly the politicized debate over the minimum wage. During the Moon Jae-in administration (2017–2022), pro-labor measures such as a 14.6% minimum wage increase in the first year, the introduction of 52-hour weekly work limits, and stricter penalties for employers in cases of industrial accidents sparked intense public discourse. Critics, including some economists and conservative politicians, viewed these

reforms as potentially undermining labor market flexibility and economic competitiveness. This charged political environment highlights the broader struggle in balancing workers' rights with market-driven considerations—a tension that continues to define labor policy debates in South Korea.

From the labor-related questions in the survey, we identified 10 items most pertinent to understanding respondents' attitudes on topics such as wage levels, working hours, retirement age, and employment conditions, which capture a range of perspectives on labor policy interventions and outcomes. Each question provided a binary response option (e.g., Yes/No or Option A/B). The 10 items are as follows:

Q1. Do you think the minimum wage increase over the past five years has reduced employment? [Select one]
- A. Yes
- B. No

Q2. Do you think the minimum wage increase over the past five years has raised prices? [Select one]
- A. Yes
- B. No

Q3. Do you think the minimum wage increase over the past five years has increased incomes? [Select one]
- A. Yes
- B. No

Q4. Which would you prefer if you had to choose between two salary systems? [Select one]
- A. A compensation system based on seniority and years of service, such as a seniority-based or grade-based pay system.
- B. A compensation system based on job roles and performance, such as job-based or merit-based pay.

Q5. How should working hours be managed? [Select one]
- A. To provide flexibility in labor management, it should be managed monthly or quarterly rather than weekly.
- B. To prevent overwork, it should be managed strictly every week.

Q6. Should regular and non-regular workers performing the same job at the same intensity receive the same wages? [Select one]
- A. They should receive the same wages.
- B. Their wages can be different.

Q7. What should be the approach to employment and dismissal conditions? [Select one]
- A. They should be eased to facilitate job mobility and create more opportunities for younger workers.
- B. They should be strengthened or at least maintained to ensure job stability.

Q8. What is your opinion on extending the retirement age? [Select one]
- A. It is necessary to address the challenges of population aging.
- B. It is undesirable as it negatively affects the hiring of new workers.

Q9. Which is more effective in reducing poverty? [Select one]
A. Minimum wage increases to supplement the income of low-income households.
B. The Earned Income Tax Credit (EITC) to provide additional income through a work subsidy.

Q10. What should be changed if unemployment benefits need to be reformed to enhance motivation to work? [Select one]
A. The benefit payments should be reduced.
B. The benefit period should be shortened rather than the amount.

Table 1 illustrates the distribution of key demographic variables: age, gender, education, and ideological orientation. To measure ideology, respondents self-identified on a five-point Likert scale from *very liberal* to *very conservative*, reflecting how they perceive their political stance rather than direct party affiliation.

Table 1. Distribution of key demographic variables.

Age Group	*n*	%	Education	*n*	%
20s	334	16.7	Below High School	13	0.7
30s	358	17.9	High School	477	23.9
40s	427	21.4	Community College	309	15.5
50s	471	23.6	College	1,024	51.2
60 and above	410	20.5	Postgraduate	177	8.9
Gender	***n***	**%**	**Ideological Orientation**	***n***	**%**
Male	1,016	50.8	Very liberal	42	2.1
Female	984	49.2	Somewhat liberal	413	20.7
			Moderate	1,094	54.7
			Somewhat conservative	402	20.1
			Very conservative	49	2.5

Note: $N = 2{,}000$.

Table 2 shows the proportion of "Yes" or "A" responses to each labor policy question, broken down by ideological group. We calculated the absolute mean difference in responses between liberal (*very*/*somewhat liberal*) and conservative (*somewhat*/*very conservative*) subsets to gauge the degree of political divide. Notably, minimum wage-related questions (i.e., Q1, Q2, Q3) displayed some of the largest gaps, confirming their salience as a point of ideological contention.

This dataset allows us to test how well ChatGPT-based models can: (a) predict responses for each question using demographic and ideological attributes (age, gender, income level; five-point ideological scale); and (b) generate synthetic data for all questions by using only these demographic and ideological attributes as model inputs, simulating how such data might capture—and potentially distort—real-world patterns.

By comparing human-derived survey responses with LLM-generated synthetic data, we aim to identify which approach introduces more bias and under which conditions that bias is magnified or mitigated. Through this process, we explore two core applications of ChatGPT in extending public opinion surveys for policy analysis and media reporting:

Table 2. Proportion of "Yes" or "A" responses to labor policy questions by ideological orientation (in percentages).

Question	All	Liberal Liberal ($n = 42$)	Somewhat Liberal ($n = 413$)	Moderate ($n = 1{,}094$)	Somewhat Conservative ($n = 402$)	Very Conservative ($n = 49$)	Absolute Mean Difference	Political Divide
Q1	62.2	45.2	47.5	62.3	75.4	87.8	29.5	Big
Q2	74.3	52.4	54.7	78.2	84.3	87.8	30.2	Big
Q3	34.2	40.5	47.7	33.5	22.4	24.5	24.4	Big
Q4	38	45.2	33.7	39.9	36.6	36.7	1.8	Small
Q5	47.5	35.7	31.7	46	67.2	63.3	34.7	Big
Q6	54.6	64.3	64.2	54.8	43.3	53.1	19.8	Small
Q7	47.2	40.5	38.7	48.3	53.5	49	14.1	Small
Q8	77.6	78.6	77.7	78.5	73.4	87.8	2.8	Small
Q9	39.3	47.6	47.2	38.9	31.8	32.7	15.3	Small
Q10	67.6	73.8	84.5	69.7	45.3	55.1	37.1	Big

Note: Big = Absolute mean difference of 20 or more; small = absolute mean difference below 20.

1. Emerging policy issues: When limited or outdated survey data exist, ChatGPT's ability to predict response patterns using only natural language can inform policymakers and journalists about likely public reactions (e.g., on a new minimum wage proposal).
2. Trend prediction: Synthetic datasets that preserve the original distribution of real data may offer deeper insights into how attitudes evolve over time or differ across subgroups. Nevertheless, biases in the training corpus or in the model's assumptions could skew these insights, emphasizing the need for rigorous validation.

In a context as ideologically fluid as South Korea—where party affiliations are less stable, but polarization over certain issues (like the minimum wage) is intensifying—ChatGPT-based approaches risk overstating or oversimplifying divides observed in Western training corpora, or underestimating the rapid shifts that characterize Korean political discourse. By systematically comparing human vs. ChatGPT-predicted responses, we aim to shed light on both the benefits and pitfalls of applying LLMs to public opinion research in dynamic, polarized settings.

In the following methods subsections, we detail how we constructed predictive models, generated synthetic data with ChatGPT, and compared these outputs to actual survey results. Through this comparison, we identify the specific dimensions—such as training data bias, demographic weighting, or topical salience—that drive discrepancies between real and AI-generated public opinion data.

3.2. Predicting Policy Preference Based on Actual Survey Data

Our first analysis focuses on extending existing policy preference data through extrapolation. Although many scientific estimation techniques exist, our primary goal is to determine whether journalists lacking specialized statistical or programming skills can still leverage LLMs to estimate public opinion on emerging issues. We, therefore, use LLMs to explore a user-friendly approach, requiring minimal technical background, to generate plausible insights about shifting policy issues.

For this analysis, we employ the ChatGPT-4o model. After uploading a data file containing variables—age, gender, income level, education, and ideological orientation—with labeled column headers, we input the following command:

> Fit a model using age, gender, income level, and ideological orientation as predictors.
>
> Then, predict which salary system each respondent will choose among the two options.
>
> Assign 1 if the respondent selects a seniority-based or grade-based pay system.
>
> Assign 2 if the respondent selects a job-based or merit-based pay system.
>
> Finally, compare these predicted values to the actual values stored in [original variable].

When training (or "fitting") a ChatGPT-style language model, the statistical method at the core is maximum likelihood estimation—implemented via gradient-based optimization (such as stochastic gradient descent and its variants) to minimize the cross-entropy loss. In other words, the model learns to predict the next token by maximizing the probability (likelihood) of the correct token in every training instance, which mathematically amounts to minimizing the negative log-likelihood (i.e., cross-entropy).

This straightforward prompt illustrates how generative AI can be used with minimal technical setup, suggesting practical value for journalists conducting public opinion research. We repeat the analysis with variations on both the questions and predictors, especially to examine how including or excluding ideological orientation influences ChatGPT's data generation.

When we prompt ChatGPT-4o to predict preferences for compensation systems, the model generates Python code that implements a regression approach. This code treats respondent demographics (age, gender, income level, and ideological orientation) as independent variables predicting the dependent variable (1 = seniority-based, 2 = merit-based). It resembles a typical social science method, pinpointing influential predictors and their respective weights. This automated procedure shows how LLMs might enable journalists to conduct regression analyses without substantial expertise in Python or advanced statistics, potentially streamlining data-driven reporting.

3.3. Creating Synthetic Data for Policy Preference Prediction

Our second analysis explores generating synthetic data with assigned personas based on demographic profiles and testing more advanced LLM use cases in media coverage. Assigning personas involves constructing consistent demographic profiles (age, gender, education, and ideological orientation) and assessing whether AI-generated responses align with how real respondents in those profiles might answer. Recent work (e.g., Argyle et al., 2023) indicates that synthetic datasets can mirror empirical distributions (algorithmic fidelity), provided the underlying models are carefully calibrated. Nevertheless, such studies warn that any inherent biases or simplifications in the model could propagate through the generated dataset.

We create synthetic data as follows. Using the ChatGPT-4o application programming interface, we specify 350 demographic groups: (2 genders × 5 age groups × 7 education levels × 5 ideological orientations) = 350, generating 10 observations per group. We use the following prompt structure (abbreviated as APGE for Age, Political orientation, Gender, and Education):

> {
>
> "id": 5,
>
> "type": "APGE,"
>
> "prompt": "You are {AGE} years old, and your political orientation leans toward {IDEOLOGICAL_ORIENTATION}. As a {GENDER} Korean, your highest level of education is {EDUCATION}."
>
> }

To minimize order bias, we shuffle and randomly select from 24 variations of the prompt format across multiple trials. We then present the AI with the original 10 labor policy questions, collecting only the answers provided by the synthetic personas. Importantly, this approach does not involve retrieving known survey results but rather generating new responses based on persona-based reasoning. Since the survey data used for evaluation was collected after the training data cut-off of ChatGPT-4o, the model could not have been exposed to or memorized these responses. Instead, its outputs reflect inferential reasoning rather than recall, ensuring that our synthetic dataset is not simply an approximation of preexisting distributions. After generating the data, we compare the synthetic dataset to the actual survey data—focusing on response patterns, demographic alignment, and overall consistency. This evaluation helps determine if synthetic data can reliably replicate real-world insights, strengthening the utility of ChatGPT for policy analysis and media applications.

4. Results and Discussions

4.1. Predicting Policy Preference Based on Actual Survey Data

This section reports how two traditional statistical models (one using only demographic variables and one adding ideological orientation) perform on the actual survey data, before exploring any ChatGPT-based (AI) simulations. We begin with minimum wage-related questions, where ideological splits tend to be pronounced, and then turn to labor policy questions that exhibit weaker partisan divides.

4.1.1. Questions With More Political Divide: Minimum Wage-Related Questions

Table 3 compares the observed proportions of "Yes" responses to three questions on the economic impacts of minimum wage increases (i.e., Q1, Q2, Q3) with two model predictions: (a) one that uses only demographic variables (age, gender, education), and (b) another that incorporates a five-point ideological scale (*very liberal* to *very conservative*). The actual data confirm a pronounced ideological divide on the impact of raising the minimum wage on employment, prices, and income. More conservative respondents overwhelmingly believe

Table 3. Comparison of actual and predicted responses to minimum wage-related questions (with big political divide).

Ideological Orientation	Actual	Predicted-Demographics Only	Predicted–Demographics and Ideology
Q1. The minimum wage increase over the past five years has reduced employment.			
Very liberal	45.2	92.9	28.6
Somewhat liberal	47.5	97.3	39.5
Moderate	62.3	96.8	88.8
Somewhat conservative	75.4	97.3	94.5
Very conservative	87.8	98	95.9
Q2. The minimum wage increase over the past five years has raised prices.			
Very liberal	52.4	97.6	38.1
Somewhat liberal	54.7	99.8	62
Moderate	78.2	100	98.8
Somewhat conservative	84.3	100	99.5
Very conservative	87.8	100	95.9
Q3. The minimum wage increase over the past five years has increased income.			
Very liberal	40.5	4.8	33.3
Somewhat liberal	47.7	2.7	55.2
Moderate	33.6	2	2.7
Somewhat conservative	22.4	2	4.2
Very conservative	24.5	2	26.5

Note: The numbers indicate the percentages of respondents who answered "Yes" or "A."

in the negative side effects of minimum wage hikes, while liberal respondents express greater skepticism about adverse outcomes and remain more open to potential benefits.

When relying solely on demographic predictors, the model produces a near-universal agreement that the minimum wage reduces employment and raises prices, thus missing the partisan splits evident in the actual data. By contrast, including ideological orientation substantially improves alignment, especially in distinguishing conservative from liberal viewpoints. Nonetheless, this expanded model occasionally overestimates the divide: For instance, it underpredicts the liberal "Yes" rate for income gains (Q3) and sometimes exaggerates differences between groups. Even so, adding ideology represents a clear improvement over relying on demographics alone.

4.1.2. Questions With Less Political Divide

To test how these models generalize beyond minimum wage debates, Table 4 presents results for three labor policy questions that evoke weaker ideological polarization: preferences for a seniority-based pay system (Q4), easing dismissal conditions (Q7), and extending the retirement age (Q8). For these less contentious issues, the demographics-only model often overstates agreement by predicting near-universal support (sometimes 100%), yet it remains reasonably stable when policy debates do not strongly follow ideological lines. The model that includes ideology typically outperforms the demographics-only approach, as it captures some ideological

Table 4. Comparison of actual and predicted responses to three labor policy questions with small political divide.

Ideological Orientation	Actual	Predicted-Demographics Only	Predicted–Demographics and Ideology
Q4. I prefer a seniority-based pay system to a merit-based pay system.			
Very liberal	45.2	11.9	38.1
Somewhat liberal	33.7	7.8	17.2
Moderate	39.9	12.6	18.5
Somewhat conservative	36.6	12.4	26.1
Very conservative	36.7	20.4	34.7
Q7. Employment and dismissal conditions should be eased to facilitate job mobility and create more opportunities for young workers.			
Very liberal	40.5	42.9	23.8
Somewhat liberal	38.7	28.8	20.6
Moderate	48.3	38	35.6
Somewhat conservative	53.5	36.3	53.7
Very conservative	49	32.7	51
Q8. Extending the retirement age is necessary to address the population aging challenges.			
Very liberal	78.6	100	83.3
Somewhat liberal	77.7	99.3	96.1
Moderate	78.5	99.6	99.9
Somewhat conservative	73.4	97	88.1
Very conservative	87.8	100	93.9

Note: The numbers indicate the percentages of respondents who answered "Yes" or "A."

subtleties—although it may still overemphasize certain subgroup differences. Notably, it accurately reflects cross-ideological backing for extending the retirement age (Q8), showing that a more nuanced model can avoid inflating polarization when the topic itself is less divisive.

4.1.3. Summary of Traditional Statistical Predictions

Overall, the demographics-only model provides a rough baseline that performs acceptably for moderately ideological issues but fails to detect real polarization in strongly politicized contexts. Adding ideology yields predictions that closely match observed responses, although it does occasionally inflate or underestimate certain group preferences. These findings illustrate the practical value of standard statistical methods in environments like newsrooms, where resources may be limited. Still, any misinterpretation, such as inflated ideological rifts, can distort public perceptions or exacerbate partisan tensions. Hence, careful validation against real survey data is essential.

4.2. Evaluating Full Synthetic Data for Policy Preference Prediction

We next explore whether ChatGPT-based synthetic data generation can mitigate or exacerbate these biases. In this subsection, we construct a synthetic dataset of 3,500 observations, incorporating demographic factors (age, gender, income level) and ideological orientation. Rather than merely extending existing data,

we generate a new dataset intended to mirror the underlying patterns from the original survey. We then compare the distributions of 10 labor-policy questions (Q1–Q10) across five ideological groups (very liberal, somewhat liberal, moderate, somewhat conservative, and very conservative) to assess how well the synthetic data aligns with the real data.

We randomly sampled up to 2,000 cases from the synthetic dataset and repeated this process 500 times, conducting a Kolmogorov–Smirnov (KS) test in each iteration. A high KS statistic (accompanied by a low *p*-value) indicates a significant mismatch between the synthetic and actual data, whereas a low KS statistic (with a high *p*-value) suggests a closer alignment between the two distributions.

Table 5 presents the proportion of iterations in which the p-value was below 0.05, signaling meaningful differences between the datasets. A value of 0 suggests that the two datasets exhibit significant differences, while a value of 1 indicates broad alignment across iterations. While many questions yield a value of 1—demonstrating strong overall similarity—certain items, such as Q1, Q3, and Q5, contain multiple instances of 0 for respondents identifying as very liberal or very conservative. This pattern suggests that ideological divergence is most pronounced in these areas, leading to a noticeable bias in the synthetic data. Conversely, questions like Q10, which appear to be less politically charged, exhibit consistently high alignment across all political groups, reinforcing their relative neutrality.

Tables 6 and 7 delve deeper into specific question sets, distinguishing politically salient issues from those deemed less divisive. Table 6 highlights hot-button labor questions, like whether minimum wage hikes reduce employment, raise prices, or increase incomes, and finds that very liberal or very conservative subgroups may jump to near-universal agreement/disagreement in the synthetic data, contrasting with more balanced splits in actual data or the regression model. These results confirm prior observations on minimum wage-related items fueling partisan gaps. Table 7, dealing with seniority-based pay or extending the retirement age, uncovers exaggerated extremes and diminished middle-ground responses, albeit less severe than in Table 6. Nevertheless, the AI-based approach can still reinforce prevalent narratives or stereotypes in the training data, especially when no explicit cultural context is given.

Table 5. KS test results for the actual and synthetic data.

Question	Very Liberal	Somewhat Liberal	Moderate	Somewhat Conservative	Very Conservative
Q1	0	0	1	0	0
Q2	1	1	1	0.951	0
Q3	0	1	0	1	0
Q4	0	0	1	1	0
Q5	0	0.891	0	0.109	0
Q6	0.004	1	0	1	1
Q7	1	1	0.002	0	0
Q8	0.030	1	1	1	1
Q9	0.130	1	1	1	1
Q10	1	1	1	1	1

Note: Numbers indicate the proportion of iterations whose *p*-values were less than 0.05, coded as "1" (adequate similarity) or "0" (significant difference).

Table 6. Comparison of actual and synthetic data-based prediction for questions with more political divide.

Ideological Orientation	Actual	Predicted: Statistical Modelling (Demographics and Ideology)	Predicted: LLM using Synthetic Data (Demographics and Ideology)
Q1. The minimum wage increase over the past five years has reduced employment.			
Very liberal	45.2	28.6	0
Somewhat liberal	47.5	39.5	28.1
Moderate	62.3	88.8	95.3
Somewhat conservative	75.4	94.5	100
Very conservative	87.8	95.9	100
Q2. The minimum wage increase over the past five years has raised prices.			
Very liberal	52.4	38.1	17.7
Somewhat liberal	54.7	62	89.7
Moderate	78.2	98.8	100
Somewhat conservative	84.3	99.5	100
Very conservative	87.8	95.9	100
Q3. The minimum wage increase over the past five years has increased income.			
Very liberal	40.5	33.3	100
Somewhat liberal	47.7	55.2	96.1
Moderate	33.6	2.7	81.3
Somewhat conservative	22.4	4.2	21.4
Very conservative	24.5	26.5	3

Note: The numbers indicate the percentages of respondents who answered "Yes" or "A."

One crucial point is that the model was never informed which questions are more polarizing in South Korea. Journalists, for instance, might prompt AI casually, overlooking local or cultural specifics and presuming the model "just knows." If ChatGPT's training is primarily global or oriented toward English-language contexts, it might apply generalized liberal–conservative frames unsuited to Korean politics—or fail to grasp actual divides in Korean policy debates. In generating our synthetic dataset, we merely instructed ChatGPT to fit a regression-like model using age, gender, income, and ideology, without flagging Q1 or Q3 as politically charged topics. Hence, the model systematically misrepresented extreme-ideology groups, overestimating splits on assumedly salient issues or underestimating them where it lacked context.

Although ChatGPT-based generation can be cost-effective and convenient, the results here underscore the risk of uncritical reliance. Despite the broad KS-based consistency in Table 5, Tables 6 and 7 reveal persistent amplification of extremes for contentious labor policies. Policymakers or media outlets that adopt such synthetic findings without due scrutiny may inadvertently intensify partisan narratives or distort actual sentiment. In countries like Korea, where party identity does not consistently map onto policy stances, ignoring cultural nuance may inflate perceived polarization. Journalists using AI casually, neglecting both policy context and ideological cues, may embed the model's preexisting biases into public discourse. Consequently, transparent disclosures, robust model-tuning, and careful comparison with real survey data are vital to avert misleading conclusions.

Table 7. Comparison of actual and synthetic data-based prediction for questions with less political divide.

Ideological Orientation	Actual	Predicted: Statistical Modelling (Demographics and Ideology)	Predicted: LLM using Synthetic Data (Demographics and Ideology)
Q4. I prefer a seniority-based pay system to a merit-based pay system.			
Very liberal	45.2	38.1	6.1
Somewhat liberal	33.7	17.2	12.4
Moderate	39.9	18.5	12.6
Somewhat conservative	36.6	26.1	53.1
Very conservative	36.7	34.7	70.7
Q7. Employment dismissal conditions should be eased to facilitate job mobility and create more opportunities for young workers.			
Very liberal	40.5	23.8	0.1
Somewhat liberal	38.7	20.6	8.6
Moderate	48.3	35.6	42.3
Somewhat conservative	53.5	53.7	85.1
Very conservative	49.0	51	98.7
Q8. Extending the retirement age is necessary to address the population aging challenges.			
Very liberal	78.6	83.3	99.7
Somewhat liberal	77.7	96.1	97.7
Moderate	78.5	99.9	85
Somewhat conservative	73.4	88.1	8.1
Very conservative	87.8	93.9	0.1

Note. The numbers indicate the % of respondents who answered 'Yes' or 'A.'

4.3. Discussion: Predictive Approaches Compared—Statistical Modeling vs. LLM Inference

This study used two primary approaches to gauge public opinion on labor policies: one that prompts an LLM to leverage regression models built directly from existing survey data using demographic and ideological predictors, and another that employs fully synthetic data generation through LLMs. Each approach caters to different journalistic needs—offering unique advantages, but also distinct vulnerabilities to bias.

When prompting the LLM to build and interpret statistical models, journalists can quickly obtain data-driven insights without advanced programming skills. As demonstrated in Sections 4.1 and 4.2, this method yields reasonably accurate predictions when ideology is included, but it may still oversimplify complex attitudes and occasionally inflate perceived ideological polarization (Tables 1 and 2). Furthermore, it relies on existing survey data, limiting its utility for new policy issues or emerging debates that lack prior information.

In contrast, generating synthetic datasets holds promise for exploring prospective public opinion in scenarios where real data are scarce. This method enables journalists to simulate responses for untested policy proposals and to forecast potential trends. Yet, Section 4.2 shows that these synthetic outputs frequently magnify ideological divisions, particularly on contentious issues such as minimum wage or flexible employment. These distortions stem from latent biases in the LLM's training data—reflecting mainstream or

US-centric assumptions (Shen et al., 2024; Zhang et al., 2023)—and the absence of local context regarding which issues are truly polarizing in Korea. Without explicit cultural labeling, the model may overinterpret or misinterpret certain labor policies, reinforcing stereotypes and inadvertently heightening partisan narratives (Tables 6 and 7).

For media practitioners, the implications are clear. While AI-powered techniques can enhance the speed and scope of data-driven reporting, their uncritical use may lead to misleading conclusions if the underlying biases are not properly addressed. Newsrooms should combine these innovative methods with traditional survey research and maintain robust validation practices. By doing so, they can mitigate the risks of overstatement and ensure that AI-generated outputs are interpreted within the correct cultural and ideological frameworks. Moreover, recent LLM variants, such as "o1," feature improved reasoning capabilities, meaning journalists could ask "why" respondents favor a certain policy and potentially receive more nuanced rationales—though this, too, may produce extreme or rhetorically charged narratives akin to those reported by *The Atlantic* when testing AI-generated partisan sentiment. In any case, close scrutiny and iterative validation remain vital for preventing AI-driven exaggerations of ideological divides.

In summary, our study not only reveals that both approaches underscore the value of LLM-powered methods for policy prediction but also demonstrates the risks of uncritical use. Where sufficient survey data exist, prompting the LLM to generate a regression model offers transparency and decent accuracy—yet it can still oversimplify people's opinions or reinforce the most visible ideological splits. Where data are limited, synthetic samples enable exploratory analysis but risk overstating polarizing trends. Future refinements in prompt design and model calibration are essential to align AI outputs more closely with local realities, ultimately supporting more responsible and nuanced journalistic practices.

5. Conclusion

This article builds on existing scholarship that emphasizes how LLM-generated text often mirrors real-world demographic and ideological biases, whether for summarizing content, filling survey gaps, or simulating entire public opinion datasets. By focusing on labor policy debates in South Korea, an especially compelling case given its fluid party system and persistent ideological polarization, we show that LLMs can replicate key survey patterns yet also overemphasize ideology on contentious issues like the minimum wage. On the one hand, we find that LLMs can approximate demographic and ideological patterns found in real survey data. On the other hand, our results show that the degree of political polarization surrounding a given policy strongly affects the model's performance. For more contentious labor issues (e.g., minimum wage), the model tends to amplify ideological differences or push respondents toward extreme positions. This underscores the need for carefully engineered prompts, domain-specific fine-tuning, and transparent disclosure of AI's role in generating opinion estimates.

Despite these challenges, LLMs hold promise for journalistic and research applications. Newsrooms can harness AI tools to produce cost-effective simulations, quickly testing public responses to new proposals or hypothetical scenarios. By tailoring the model to include balanced demographic profiles, media organizations might reduce biases and foster more inclusive coverage. AI-driven simulations could broaden perspectives in politically polarized environments, but on the condition that they are carefully engineered to avoid reinforcing echo chambers.

At the same time, our results underscore the necessity of cultural contextualization. If journalists rely solely on casual prompts, neglecting to specify which labor policies trigger fierce debates in Korea, the model's pre-trained assumptions about liberal–conservative divisions—often US-centric—may overstate real ideological rifts. Building domain-specific or regionally fine-tuned versions of LLMs could help counterbalance inherent biases and reduce the risk of amplifying polarizing narratives (J. Lee et al., 2024).

In the near future, more advanced models (e.g., those capable of detailed chain-of-thought reasoning) could allow journalists to probe not just "what" the simulated response is but "why" certain demographic or ideological groups endorse one policy over another. These "why" prompts may yield deeper rationales but also risk providing overly confident or partisan-sounding explanations, much like *The Atlantic* experienced when eliciting AI-generated partisan rhetoric. Further research should test how these refined models balance explanatory depth and amplify ideological stereotypes. Expanding experiments beyond Korea could illuminate whether certain societies or cultures are more prone to LLM-induced distortions and how best to mitigate them.

Acknowledgments

We thank the journal editors and reviewers for their invaluable feedback, which greatly strengthened this manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) Grant through the Korean Government [Ministry of Science and ICT (MSIT)] under Grant RS-2022-NR070854 and the KDI School of Public Policy and Management's financial support.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability

Data supporting this study may be made available upon reasonable request.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning*, *202*, 337–371. https://proceedings.mlr.press/v202/aher23a.html

Amirova, A., Fteropoulli, T., Ahmed, N., Cowie, M. R., & Leibo, J. Z. (2024). Framework-based qualitative analysis of free responses of large language models: Algorithmic fidelity. *PLoS ONE*, *19*(3), Article e0300024. https://doi.org/10.1371/journal.pone.0300024

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

BBC. (2025). BBC AI transparency. https://www.bbc.co.uk/supplying/working-with-us/ai-transparency

Berger, A., Schneier, B., Gong, E., & Sanders, N. (2024, June 7). *Using AI for political polling: Will AI-assisted polls soon replace more traditional techniques?* Ash Center for Democratic Governance. https://ash.harvard.edu/articles/using-ai-for-political-polling

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(6), Article e2218523120. https://doi.org/10.1073/pnas.2218523120

Brigham, N. G., Gao, C., Kohno, T., Roesner, F., & Mireshghallah, N. (2024). *Breaking news: Case studies of generative AI's use in journalism*. arXiv. https://arxiv.org/html/2406.13706v1

Caswell, D. (2024). Audiences, automation, and AI: From structured news to language models. *AI Magazine*, *45*(2), 174–186.

Caughey, D., O'Grady, T. O. M., & Warshaw, C. (2019). Policy ideology in European mass publics, 1981–2016. *American Political Science Review*, *113*(3), 674–693. https://doi.org/10.1017/S0003055419000157

Cheong, Y., & Haggard, S. (2023). Political polarization in Korea. *Democratization*, *30*(7), 1215–1239. https://doi.org/10.1080/13510347.2023.2217762

Cho, Y., Kim, M.-s., & Kim, Y. C. (2019). Cultural foundations of contentious democracy in South Korea: What type of democracy do Korean citizens prefer? *Asian Survey*, *59*(2), 272–294. https://doi.org/10.1525/as.2019.59.2.272

Desai, S. (2023, April 3). Return of the people machine. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/04/polls-data-ai-chatbots-us-politics/673610

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

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*. arXiv. https://doi.org/10.48550/arXiv.2305.08283

Fletcher, R., & Nielsen, R. K. (2024). *What does the public in six countries think of generative AI in news?* Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/what-does-public-six-countries-think-generative-ai-news

Gerosa, M., Trinkenreich, B., Steinmacher, I., & Sarma, A. (2024). Can AI serve as a substitute for human subjects in software engineering research? *Automated Software Engineering*, *31*(1), Article 13. https://doi.org/10.1007/s10515-023-00409-6

Gindler, A. (2021). The theory of the political spectrum. *Journal of Libertarian Studies*, *24*(2), 240–271.

Glickman, M., & Zhang, Y. (2024). AI and generative AI for research discovery and summarization. *Harvard Data Science Review*, *6*(2). https://doi.org/10.1162/99608f92.7f9220ff

Han, H., & Ding, J. (2024). Measures to overcome population decline and regional extinction crises of pan-tourism aspects in Korea: An exploratory study using ChatGPT. *Global Business & Finance Review*, *29*(7), 78–92. https://doi.org/10.17549/gbfr.2024.29.7.78

Harding, J., D'Alessandro, W., Laskowski, N. G., & Long, R. (2024). AI language models cannot replace human research participants. *AI & Society*, *39*(5), 2603–2605. https://doi.org/10.1007/s00146-023-01725-x

Hutson, M. (2023). Guinea pigbots. *Science*, *381*(6654), 121–123. https://doi.org/10.1126/science.adj6791

Kennedy, C., Mercer, A., Hatley, N., & Lau, A. (2022, September 21). Does public opinion polling about issues still work? *Pew Research Center*. https://www.pewresearch.org/short-reads/2022/09/21/does-public-opinion-polling-about-issues-still-work

Kim, G. Y. (2023, February 8). "Kim Jong Un eo-tteoh-ge saeng-gag-ha-nya" ChatGPT-e mul-eoss-deo-ni…dol-a-on dae-dab-eun. *Chosun Ilbo*. https://www.chosun.com/economy/tech_it/2023/02/08/S4A25QYUKJEFJO4ZVQZIS3ACDM

Kim, K. R., & Kang, J.-S. (2013). Liberal–conservative self-identification in Korea: A cross-cultural explanation. *Korean Social Science Journal*, *40*(2), 113–120. https://doi.org/10.1007/s40483-013-0009-7

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In M. Bernstein, S. Savage, & A. Bozzon (Eds.), *CI '23: Proceedings of The ACM Collective Intelligence Conference* (pp. 12–24). ACM. https://doi.org/10.1145/3582269.3615599

Lee, C., Kim, Y., & Park, J. (2024). *Korean policy choice survey (2024–2025).* [Survey data]. KDI School of Public Policy and Management.

Lee, J., Kim, M., Kim, S., Kim, J., Won, S., Lee, H., & Choi, E. (2024). KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In C. Gardent (Ed.), *Findings of the Association for Computational Linguistics ACL 2024* (pp. 11177–11213). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.666

Lee, S. (2023, March 28). ChatGPT-ey mwul-ess-ta…ne-nun wu-li il-ca-li-lul pha-koy-ha-ni? *Hankyoreh*. https://www.hani.co.kr/arti/opinion/column/1085554.html

McCarty, N., Poole, K. T., & Rosenthal, H. (2016). *Polarized America: The dance of ideology and unequal riches*. MIT Press.

Milička, J., Marklová, A., VanSlambrouck, K., Pospíšilová, E., Šimsová, J., Harvan, S., & Drobil, O. (2024). Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLoS ONE*, *19*(3), Article e0298522. https://doi.org/10.1371/journal.pone.0298522

Moravec, V., Hynek, N., Skare, M., Gavurova, B., & Kubak, M. (2024). Human or machine? The perception of artificial intelligence in journalism, its socio-economic conditions, and technological developments toward the digital future. *Technological Forecasting and Social Change*, *200*, Article 123162. https://doi.org/10.1016/j.techfore.2023.123162

Pachot, A., & Petit, T. (2024). *Can large language models accurately predict public opinion? A review* (HAL Working Paper hal-04688498). HAL Open Science. https://hal.science/hal-04688498

Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M. Y., & Wang, W. (2023). On the risk of misinformation pollution with large language models. In Y. Matsumoto (Ed.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1389–1403). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.97

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, *56*(6), 5754–5770. https://doi.org/10.3758/s13428-023-02307-x

Patel, H. (2024). *Bane and boon of hallucinations in context of generative AI*. TechRxiv. https://doi.org/10.36227/techrxiv.171198062.20183635/v1

Rozado, D. (2024). *The political preferences of LLMs*. arXiv. https://doi.org/10.48550/arXiv.2402.01789

Shen, S., Logeswaran, L., Lee, M., Lee, H., Poria, S., & Mihalcea, R. (2024). Understanding the capabilities and limitations of large language models for cultural commonsense. In K. Erk (Ed.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 5668–5680). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.316

Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B. J., & Kim, J. H. (2024). *Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information*. arXiv. https://doi.org/10.48550/arXiv.2402.18144

Yang, K.-C., & Menczer, F. (2024). Anatomy of an AI-powered malicious social botnet. *Journal of Quantitative Description: Digital Media*, *4*(2024), 1–36. https://doi.org/10.51685/jqd.2024.icwsm.7

Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G. (2023). Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In Y. Matsumoto (Ed.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7915–7927). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.491

About the Authors

Keyeun Lee is a master's student researching AI communication.

Jaehyuk Park is an assistant professor and chair of data science at KDI School of Public Policy and Management. He holds a PhD in informatics from Indiana University and was a postdoctoral fellow at Northwestern University's Kellogg School of Management. His research explores AI and data science applications in labor markets and public policy.

Suh-hee Choi is an associate professor at Kyung Hee University. She researches tourism mobilities, migrants' work and life, cultural tourism, tourist experiences, public diplomacy, and community-based tourism. She teaches courses on the mobilities paradigm and human geography, integrating service learning and generative AI into her project-based and regional geography courses.

Changkeun Lee is an associate professor of economics at the KDI School of Public Policy and Management. His research explores the intersection of technology, labor markets, and institutions, covering topics in modern American and Korean history as well as contemporary policy issues in Korea.

Media and Communication
2025 • Volume 13 • Article 9623
https://doi.org/10.17645/mac.9623

cogitatio

ARTICLE

Open Access Journal

Unmasking Machine Learning With Tensor Decomposition: An Illustrative Example for Media and Communication Researchers

Yu Won Oh [1] and Chong Hyun Park [2]

[1] School of Digital Media, Myongji University, Republic of Korea
[2] School of Business, Sungkyunkwan University, Republic of Korea

Correspondence: Chong Hyun Park (chypark@skku.edu)

Submitted: 15 November 2024 **Accepted:** 5 February 2025 **Published:** 24 April 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

As online communication data continues to grow, manual content analysis, which is frequently employed in media studies within the social sciences, faces challenges in terms of scalability, efficiency, and coding scope. Automated machine learning can address these issues, but it often functions as a black box, offering little insight into the features driving its predictions. This lack of interpretability limits its application in advancing social science communication research and fostering practical outcomes. Here, explainable AI offers a solution that balances high prediction accuracy with interpretability. However, its adoption in social science communication studies remains limited. This study illustrates tensor decomposition—specifically, PARAFAC2—for media scholars as an interpretable machine learning method for analyzing high-dimensional communication data. By transforming complex datasets into simpler components, tensor decomposition reveals the nuanced relationships among linguistic features. Using a labeled spam review dataset as an illustrative example, this study demonstrates how the proposed approach uncovers patterns overlooked by traditional methods and enhances insights into language use. This framework bridges the gap between accuracy and explainability, offering a robust tool for future social science communication research.

Keywords

automated content analysis; explainable AI; machine learning; PARAFAC2; tensor decomposition

1. Introduction

As vast amounts of communication data accumulate online, manual content analysis, which is widely employed in media studies within the social sciences, faces limitations in terms of coding scope, effort, and efficiency

(Kroon et al., 2024). This explains why automated approaches—despite suspicions of being "*incorrect* models of language" (Grimmer & Stewart, 2013, p. 268, emphasis in original)—are increasingly being tried in digital corpus analysis. Among these, automated methods combined with machine learning draw on the utility of the technique in classification and prediction and have been applied to detect, for instance, opinion spam (e.g., Oh & Park, 2021), incivility (e.g., Burnap & Williams, 2015), and misinformation (e.g., Tanvir et al., 2019).

Typically, machine learning approaches operate on training data—whether labeled or unlabeled—to predict the outcomes of the test data. The algorithms developed through this process show great promise for efficient analysis of large-scale online media and communication data. In the advancement of machine-learning algorithms that demonstrate satisfactory performance in the field of communication, one unsettling aspect is that the underlying mechanism behind the final prediction remains a *black box*. Most machine-learning models are designed with a primary focus on achieving high accuracy in decisions, and the development of such models is often a significant accomplishment in itself. At the same time, however, it also remains true that one cannot understand which features or variables—or combinations of them—drive the predictions. In other words, they are uninterpretable and unexplainable (Rudin & Radin, 2019).

This black-box nature is particularly unfortunate in communication research within social science disciplines. For example, although a study proposed an algorithm capable of distinguishing deceptive comments written by paid commenters from genuine ones with nearly 81% accuracy (Oh & Park, 2021), it did not indicate which features of the comments should raise suspicion. A kind of dilemma—models can be accurate but cannot be understood—makes it challenging to apply research findings from state-of-the-art methods to media literacy education or guidelines and, more fundamentally, to deepen our understanding of human communicative acts.

In this regard, explainable AI, which has been actively explored in other fields, has attracted attention. Explainable AI aims to communicate the meaning from resulting models without significantly compromising the performance advantages of machine learning in solving complex problems (Ali et al., 2023). It offers a way to improve the trustworthiness and transparency of models that ensure high prediction accuracy (Rai, 2020). However, despite the clear potential benefits its application could bring to the analysis of large online media and communication datasets, to date, few related attempts have been made in social science communication research (cf. Dobbrick et al., 2022).

As a proactive and forward-looking response, this study provides media scholars with a guide for digital content analysis using interpretable machine learning methods. We focused on tensor decomposition—specifically, PARAFAC2—among the several techniques worth considering. In this study, we illustrate this method and demonstrate how media and communication researchers can employ it to analyze online corpora and interpret the results. To explain it, we rely on a review dataset constructed by Ott et al. (2011, 2013).

2. Literature Review

2.1. Text Analysis Method

The evolution of text analysis methods began with the basic bag of words (BoW) approach, which progressively developed into more sophisticated techniques. Grimmer and Stewart (2013) highlighted the limitations of BoW

models, noting that they analyze solely based on word frequency while ignoring word order and contextual information, thus failing to capture the structural meaning within texts. Despite these limitations, BoW remains widely used because of its computational efficiency and straightforward structure. Boumans and Trilling (2016) emphasized the efficiency of dictionary-based approaches such as BoW, explaining that the advancement of automated methodologies is essential for handling the increasing demand for data processing in text analysis.

Among dictionary-based methods, Linguistic Inquiry and Word Count (LIWC) provides an in-depth linguistic analysis by categorizing words into psychological, social, and linguistic domains. Van Atteveldt et al. (2021) noted that LIWC extends beyond BoW's simple analysis to assess psychological and emotional elements, although it still struggles to capture subtle contextual nuances. Recently, large language models such as BERT and GPT-x have demonstrated impressive contextual understanding, and are increasingly being applied to content analysis. Rogers et al. (2020) analyzed the internal mechanisms of BERT, underscoring the model's complex and difficult-to-interpret learning process. Similarly, Zini and Awad (2022) discussed how large language models such as GPT-x exhibit substantial generative capabilities from extensive data training but retain a black-box quality, posing challenges in predictability and explainability. While efforts to improve the explainability of such models are ongoing, the complexity of large language models remains a key concern.

In this context, our study adopted a dictionary-based LIWC-supported BoWs approach, which is better suited for social science analysis, where interpretability and explainability are paramount. LIWC allows for a clear interpretation of analysis results and provides explanations based on specific linguistic characteristics, aiding researchers in understanding psychological and linguistic patterns in communication data. Additionally, our analysis utilized only the linguistic dimensions of LIWC, not to exclude its psychological and social features, but rather to clarify the study's focus on presenting a methodological approach to corpus analysis.

2.2. Machine Learning Challenges in High Dimensions

Supervised machine learning models focus on prediction and classification tasks using labeled data and employ algorithms such as linear regression, logistic regression, support vector machine, and neural networks to learn the correct output for given inputs. These models are intuitive and predictive due to the presence of clear answers. However, they struggle to effectively handle complex nonlinear relationships or multidimensional interactions in high-dimensional data. Bishop (2006) highlighted these limitations, stressing the need for more sophisticated models to address the complexity of high-dimensional relationships. As data complexity increases, classic machine-learning techniques face difficulties in adequately capturing intricate connections and learning nonlinear patterns (Hastie et al., 2009). Addressing these challenges requires advanced analytical methods, emphasizing the growing importance of techniques capable of high-dimensional data analysis.

Meanwhile, unsupervised machine learning models have evolved to identify patterns and structures in unlabeled data. Algorithms such as K-means, density-based spatial clustering of applications with noise, principal component analysis, and autoencoders excel at extracting features and exploring patterns, particularly in clustering and dimensionality reduction. However, the results are often difficult to interpret. To overcome this problem, tensor decomposition methods have gained attention as powerful tools for analyzing multidimensional interactions in high-dimensional data. Shin and Woo (2022) emphasized the

effectiveness of tensor decomposition for extracting significant patterns from complex datasets and uncovering hidden structures. Shi et al. (2018) also demonstrated its utility in capturing crucial features from multidimensional data for better understanding.

Our study applies the PARAFAC2 algorithm, which is typically utilized in unsupervised learning to detect essential patterns in multidimensional data, in the unique context of analyzing labeled data. Although PARAFAC2 is known for its effectiveness in identifying interactions within high-dimensional datasets (Sidiropoulos et al., 2017), our approach extends its conventional usage. By leveraging tensor decomposition in this manner, we aim to make the complex relationships within the labeled data more comprehensible. Kolda and Bader (2009) emphasized the broader implications of tensor methods in uncovering intricate data structures, supporting the application of this technique to provide interpretable insights.

2.3. Applications of Tensor Decomposition in Communication-Related Topics

Tensor decomposition methods have been in use for several decades (Carroll & Chang, 1970; Harshman, 1970; Tucker, 1966) and have seen widespread application across various fields, including chemometrics (Smilde et al., 2004), signal processing (Sidiropoulos et al., 2000), computer vision (Vasilescu & Terzopoulos, 2002), numerical analysis (Beylkin & Mohlenkamp, 2005), graph analysis (Kolda et al., 2005), and web search personalization, where query terms or anchor text serve as the third dimension (Kolda et al., 2005; Sun et al., 2005). Building on these diverse applications, another research direction is centered on improving the performance of tensor decomposition techniques. For example, Kolda and Sun (2008) explored tensor decompositions in multi-aspect data mining and optimized these methods for high-dimensional and sparse data. For a comprehensive overview of tensor decompositions, see Kolda and Bader (2009), which provides an in-depth discussion of the mathematical foundations, various decomposition models, and their applications across multiple fields.

Although the potential of tensor decomposition in social science communication research is gradually becoming more apparent, it unfortunately remains largely unfamiliar to social science media and communication researchers. Most studies applying tensor decomposition to communication data have been conducted in the fields of science and engineering, the so-called STEM. In the context of text analysis, tensor decomposition methods have proven particularly valuable for improving the interpretability of machine learning models because they allow for the extraction of underlying patterns and structures from high-dimensional text data. For instance, Acar et al. (2005) explored the use of tensor decomposition techniques across various types of data, including texts. Specifically, they used tensor decompositions of (user × keyword × time) data to distinguish conversation threads in chatroom data. This approach is highly beneficial for handling the complexity of text data compared with traditional, simpler analytical methods. Similarly, PARAFAC has been applied to email communications, as in Bader et al. (2008), where it was used to track discussions in the Enron email corpus. Papalexakis et al. (2016) further underscored the significance of tensor decomposition in data-mining applications, such as topic modeling and sentiment analysis, providing an efficient framework for managing high-dimensional data. Saha and Sindhwani (2012) introduced a method using dynamic tensor decomposition to analyze temporal topic evolution and user interaction patterns on social media, offering valuable insights for addressing time-series text data. Subsequent research continued to adopt tensor decomposition to analyze the structural intricacies of text data and social interactions, progressively broadening its scope of application.

Nevertheless, challenges remain in the broader adoption of high-dimensional data analysis techniques such as tensor decomposition in communication research. Stoll et al. (2023) identified the limitations of existing methods in their study on detecting incivility in German online discussions, highlighting the difficulties in analyzing the multidimensional nature of text data. To address these challenges, recent studies, such as Schuld et al. (2023), explored advanced techniques using machine learning to deepen the analysis of opinion discourse and enrich the field of text data analysis. Our research demonstrates how tensor decomposition methods, specifically PARAFAC2, can be used to enhance the analysis of text data in social science communication research. By leveraging this technique, researchers can extract complex multidimensional relationships, identify critical patterns, advance the precision of text analysis, and address the unresolved complexities in existing methodologies.

3. Methods

This section describes the data analysis method using tensor decomposition. Tensor decomposition is a technique that represents data as a multidimensional array and divides it into various components to extract hidden patterns. To illustrate the application of tensor decomposition methods in social science communication research, we selected an online spam reviews dataset as an illustrative example. We first used the LIWC tool to extract linguistic features from online reviews, then transformed the data into a tensor form, and applied a tensor decomposition algorithm to identify significant patterns. Furthermore, to demonstrate the effectiveness of tensor decomposition in extracting valuable information from high-dimensional data, we applied it to an analysis of deceptive reviews.

3.1. Overview of Tensor Decomposition

Here, for social science communication researchers, we explain the fundamental concepts necessary for understanding tensor decomposition. We introduce the definition and structure of tensors and provide a simple example to illustrate the key principles and methods of tensor decomposition. This content serves as a foundation for understanding the tensor decomposition algorithms discussed in subsequent sections.

3.1.1. Tensor

A tensor is a mathematical concept that represents data as a multidimensional array, allowing for a structured representation across multiple dimensions. For example, a scalar is a 0th-order tensor, a vector is a 1st-order tensor, and a matrix is a 2nd-order tensor. Arrays with dimensions higher than these are referred to as 3rd-order, 4th-order tensors, and so on, depending on the number of dimensions. A 3rd-order tensor X can be approximated as the outer product of three vectors. The outer-product operation combines two or more vectors to create a higher-dimensional object. For example, a 3rd-order tensor X $\times \times$ can be approximately expressed using three vectors , , and . Mathematically, this is represented as:

$$X =$$

Here, denotes the outer product. The outer-product operation combines the elements of the two vectors to form a matrix. For instance, the outer product of vectors and ^{n} is defined as:

$$= \begin{matrix} {}_{1}{}_{1} & {}_{1}{}_{2} & & {}_{1}{}_{n} \\ {}_{2}{}_{1} & {}_{2}{}_{2} & & {}_{2}{}_{n} \\ & & & \\ {}_{1} & {}_{2} & & {}_{n} \end{matrix}$$

This results in an $\times n$ matrix. The operation multiplies each element of with each element of , combining the results into a matrix. Consider a 3rd-order tensor $\mathcal{X}$ $^{2\times2\times2}$ as an example. This tensor consists of two matrices stacked along the third dimension:

$$\mathcal{X} = \begin{matrix} 1 & 2 \\ 3 & 4 \end{matrix}, \begin{matrix} 5 & 6 \\ 7 & 8 \end{matrix}$$

This structure can be approximated using three vectors, , , and , which are defined as = 1 2 , = 1 3 , and = 1 4 . The outer product of these vectors approximates $\mathcal{X}$, computed as:

$$= \begin{matrix} 1\times1\times1 & 1\times1\times4 \\ 1\times3\times1 & 1\times3\times4 \end{matrix}, \begin{matrix} 2\times1\times1 & 2\times1\times4 \\ 2\times3\times1 & 2\times3\times4 \end{matrix}$$

This yields:

$$= \begin{matrix} 1 & 4 \\ 3 & 12 \end{matrix}, \begin{matrix} 2 & 8 \\ 6 & 24 \end{matrix}$$

Although this approximation generates a structure similar to $\mathcal{X}$, it may not match exactly. Techniques such as matrix factorization and tensor decomposition have been used to achieve more accurate approximations.

3.1.2. Tensor Decomposition

Tensor decomposition approximates a tensor as the sum of multiple rank-1 tensors. The more rank-1 tensors are used, the more accurately $\mathcal{X}$ can be approximated. CANDECOMP/PARAFAC (CP) decomposition is one such method that approximates a 3rd-order tensor as the sum of the outer products of vectors for each dimension. Figure 1 illustrates that a given 3rd-order tensor X $^{\times\times}$ can be approximated as a sum of these outer products. The CP decomposition is mathematically expressed as:

$$X \quad {}_{i=1}$$

Here, $_{i}$, $_{i}$, and $_{i}$ are the i-th component vectors corresponding to each dimension, and represents the number of rank-1 tensors needed for the approximation.

Tensor decomposition is a highly effective unsupervised learning method used to extract features and patterns from high-dimensional data. It is well-known for its capability to classify data even in the absence of labeled training examples (Kolda & Bader, 2009). To illustrate the efficiency of tensor decomposition in identifying important features from multidimensional data, we analyzed a small example of social media user interaction data. In this example, suppose we aim to classify user groups based on their interaction times with various content types such as images and videos (Acar et al., 2005). The data can be represented as a three-dimensional tensor based on the user, content type, and time of day. Suppose that tensor $\mathcal{X}$ comprises

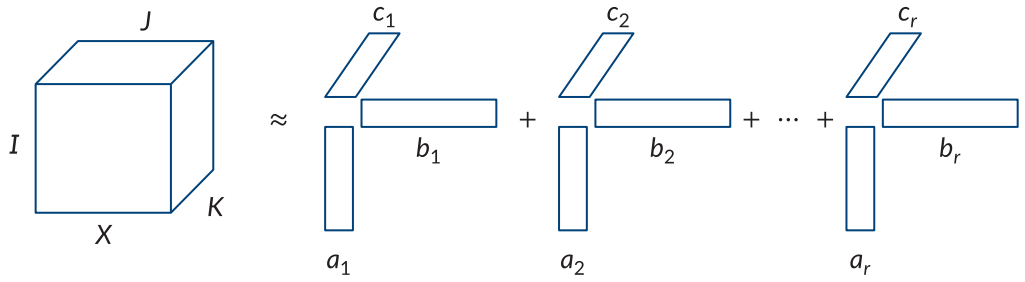


Figure 1. Tensor decomposition method.

interactions among three users, two content types (image and video), and two time periods (day and night). Each value in $\mathcal{X}$ represents the number of times a user interacted with a given content type during a specific time period. For instance, the first user interacted with image content five times during the day and twice at night and with video content once during the day and three times at night. The interaction data can be organized as follows:

$$X = \begin{matrix} 5 & 2 \\ 1 & 3 \end{matrix}, \begin{matrix} 4 & 3 \\ 2 & 1 \end{matrix}, \begin{matrix} 1 & 0 \\ 0 & 5 \end{matrix}$$

Here, the first matrix represents interactions of user 1, the second matrix those of user 2, and the third matrix those of user 3. Using CP decomposition, we approximate $\mathcal{X}$ as the sum of two rank-1 tensors:

$$X \quad {}_1 \; {}_1 \quad {}_1 \quad {}_1 + {}_2 \; {}_2 \quad {}_2 \quad {}_2$$

Here $_1$ and $_2$ are scalar values, $_1$, $_2$ are user feature vectors, $_1$, $_2$ are content-type feature vectors, and $_1$, $_2$ are time-period feature vectors. Suppose that the first rank-1 tensor is given by:

$$_1 = 5, \quad _1 = \begin{matrix} 1.0 \\ 0.8 \\ 0.2 \end{matrix}, \quad _1 = \begin{matrix} 0.95 \\ 0.4 \end{matrix}, \quad _1 = \begin{matrix} 0.7 \\ 0.6 \end{matrix}$$

The outer product of these vectors generates the following three-dimensional array:

$$_1 \; {}_1 \quad {}_1 \quad {}_1 = \begin{matrix} 3.325 & 2.85 \\ 1.4 & 1.2 \end{matrix}, \begin{matrix} 6.65 & 5.7 \\ 2.8 & 2.4 \end{matrix}, \begin{matrix} 0.95 & 0.9 \\ 0.4 & 0.35 \end{matrix}$$

Now, assume the second rank-1 tensor is defined as:

$$_2 = 3, \quad _2 = \begin{matrix} 0.3 \\ 0.6 \\ 0.9 \end{matrix}, \quad _1 = \begin{matrix} 0.1 \\ 0.8 \end{matrix}, \quad _1 = \begin{matrix} 0.5 \\ 0.9 \end{matrix}$$

The outer product of these vectors is computed similarly, and the sum of both rank-1 tensors approximates X:

$$_2 \; {}_2 \quad {}_2 \quad {}_2 = \begin{matrix} 0.045 & 0.081 \\ 0.36 & 0.648 \end{matrix}, \begin{matrix} 0.09 & 0.162 \\ 0.72 & 1.296 \end{matrix}, \begin{matrix} 0.135 & 0.243 \\ 1.08 & 1.944 \end{matrix}$$

The final approximation is obtained by adding the two rank-1 tensors:

$$X \quad \begin{matrix} 3.37 & 2.931 \\ 1.76 & 1.848 \end{matrix}, \begin{matrix} 6.74 & 5.862 \\ 3.52 & 3.696 \end{matrix}, \begin{matrix} 1.085 & 1.143 \\ 1.48 & 2.294 \end{matrix}$$

Through this method, tensor decomposition serves as a tool for analyzing high-dimensional data. After CP decomposition, the feature matrices for users, content types, and time periods can be extracted. For instance, the user feature matrix A is:

$$A = {}_1 ; {}_2 = \begin{matrix} 1.0 & 0.3 \\ 0.8 & 0.6 \\ 0.2 & 0.9 \end{matrix}$$

The users can be grouped using these feature matrices. For example, by applying k-means clustering, we might group users with similar interaction patterns. Users 1 and 2 could belong to a group interacting mainly with content during the day, whereas user 3 might be grouped based on nighttime interactions with images. This analysis demonstrates that tensor decomposition effectively summarizes the key patterns in user behavior, enabling the grouping of similar users and the provision of personalized services (Wang et al., 2023).

3.2. Tensor Decomposition-Based Method

This section describes the method for analyzing the linguistic features of opinion spam using a tensor decomposition-based algorithm. It explains the process of transforming text data into a multidimensional tensor to extract linguistic patterns, which are then used to differentiate between fake and genuine reviews.

3.2.1. Dataset

For illustration and demonstration, we used the Deceptive Opinion Spam Corpus v1.4, developed by Ott et al. (2011, 2013). This reliable labeled opinion dataset consists of 1,600 reviews about 20 hotels located in Chicago, with 800 genuine and 800 fake reviews:

- Genuine reviews: A total of 400 positive reviews were collected from TripAdvisor. These reviews were based on actual lodging experiences and were sampled by Ott et al. (2011), excluding non-English and short reviews to ensure a matching length distribution among five-star reviews for the 20 hotels. Four hundred negative reviews written by travelers with genuinely negative experiences were collected from various travel websites, such as Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp (Ott et al., 2013).
- Fake reviews: The 400 positive fake reviews were written by Amazon Mechanical Turk workers who were instructed to create positive reviews promoting specific hotels. The reviews had to be realistic and persuasive. The workers were US-based with a past approval rate of over 90%. All reviews were manually screened to ensure quality (Ott et al., 2011). A total of 400 negative fake reviews were also generated by Amazon Mechanical Turk workers who were tasked with writing reviews that portrayed competing hotels negatively (Ott et al., 2013).

3.2.2. Linguistic Feature Generation

Extracting linguistic features from review texts in our dataset using the LIWC program may hint at the characteristics of spam opinions. LIWC is a dictionary-based text analysis tool that connects word usage to various linguistic categories (Boyd et al., 2022). The LIWC analysis process involves several steps. First, LIWC examines each word in the text to determine whether it belongs to predefined linguistic categories,

such as pronouns, conjunctions, or interjections, which reflect the structural characteristics of the text. LIWC then calculates the total frequency of words in each category and converts it into a percentage relative to the total word count. For example, if pronouns constitute 10% of the words in a given text, LIWC assigns a value of 10 to that category. This approach enables objective measurement of the linguistic features present in the text.

From the opinion spam reviews dataset, we extracted over 18 features, focusing on the linguistic dimensions categories of LIWC. The extracted features were represented using a BoW model based on the LIWC dictionary approach. The BoW model analyzes word frequency while ignoring word order, thus offering simplicity and high explainability (Kroon et al., 2024). Although this model may have limitations such as the loss of contextual information, it is widely used in tasks such as spam opinion analysis due to its effectiveness. Through a linguistic feature analysis, the linguistic pattern characteristics of spam opinions were explored.

Before constructing a tensor to analyze the linguistic features of spam opinions, we preprocessed the data. The 18 features utilized in this study had values distributed across different ranges and scales. These differences arise because the frequency of words belonging to various categories in LIWC analyses varies significantly. For instance, some categories may have high frequencies, resulting in large values, while others may have low frequencies, leading to lower values. To enhance data consistency and analytical accuracy, we applied min-max normalization. This technique transforms the values of each feature to fall within the range of 0 to 1, adjusting the minimum value to 0 and the maximum value to 1.

The formula for min-max normalization is as follows:

$$Z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, Z represents the normalized value, x is the original value, and $\min(x)$ and $\max(x)$ denote the minimum and maximum values of the feature, respectively. By applying this normalization process, we ensured that all features were on the same scale, thus preventing any single feature from disproportionately affecting the analysis. This transformation uniformly distributes the data, allowing for effective comparisons between different values and minimizing the risk of skewed results during tensor construction. Such preprocessing forms a reliable foundation for accurately exploring the linguistic patterns of spam opinions.

The data generated in the previous preprocessing step are represented in the form of a tensor. Specifically, the preprocessed datasets for genuine and fake reviews consisted of 800 reviews and 18 features. Thus, both the genuine and fake review groups have dimensions of 800×18, where 18 represents the number of linguistic features extracted for each review group. Using these data, we constructed a three-dimensional tensor. The tensor structure is illustrated on the left in Figure 2. The final tensor size was defined as [2] × × , which translated to $(800, 800) \times 18 \times 2$. Here, each dimension represented reviews ([2]), linguistic features (= 18), and review types (= 2). The review dimension encompassed both genuine and fake review groups, the feature dimension consisted of the 18 linguistic features extracted using LIWC, and the review type dimension distinguished between genuine and fake categories. This tensor structure was utilized to analyze the complex interactions between the linguistic features of genuine and fake reviews and to uncover hidden patterns and relationships using tensor decomposition techniques.

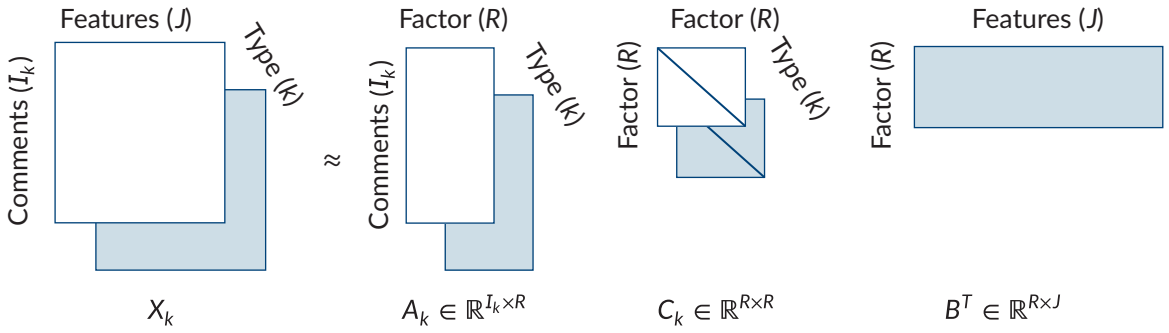


Figure 2. PARAFAC2 decomposition method.

We used the PARAFAC2 algorithm to decompose the constructed three-dimensional review dataset into component matrices *A*, *B*, and . PARAFAC2 is a variant of CP that can be applied to a collection of matrices with the same number of columns but different numbers of rows, that is, when the matrices do not form a true tensor. One of the key advantages of PARAFAC2 is its ability to handle matrices with varying sizes in one mode, such as datasets with the same column dimensions but different row sizes. This flexibility makes PARAFAC2 well-suited for cases where the dimensions vary across slices, for example, in datasets with different numbers of genuine and fake reviews. The PARAFAC2 algorithm decomposes the tensor into multiple rank-1 components, thereby enabling the effective discovery of hidden patterns. Specifically, the three-dimensional tensor is divided into frontal slices, where each slice is represented as a two-dimensional matrix. For the -th frontal slice X $^{\times \times}$, the PARAFAC2 decomposition is mathematically expressed as:

$$X \quad A \bullet \quad \bullet B$$

Here, = 1, 2, , , A $^{\times R}$, B $^{\times R}$, $^{R \times R}$ denote the component matrices. is the -th diagonal matrix, whose diagonal elements explain the relationships among factors. Through this decomposition, we obtain the component matrices *A*, *B*, and . Matrix *A* represents the factor dependencies for each review group, and is composed of submatrices (800×) that explain how each review is influenced by specific factors.

The value of , which represents the number of such factors in the tensor decomposition, needs to be determined during the analysis. In tensor decomposition, selecting an appropriate rank is crucial to ensure the quality and interpretability of the decomposition. The rank represents the number of components or factors used to decompose the tensor. Choosing an optimal helps balance the complexity and accuracy of the model. However, there is no straightforward algorithm to select the optimal , and determining the best rank requires a balance between approximation accuracy and model complexity (Bader et al., 2008). If the rank is excessively small, the decomposition may fail to capture the underlying patterns in the data, leading to a poor approximation. On the other hand, if the rank is too large, the model may overfit the data and capture noise rather than meaningful features.

A widely used method for selecting the optimal rank is by minimizing the reconstruction error R_{error}, which is calculated using the Euclidean norm (also known as the Frobenius norm) between the original tensor $^{1\times}{}^{2\times}{}^{3}$ and the approximated tensor $\hat{}$. The Euclidean norm of the error is expressed as:

$$R_{\text{error}} = \quad - \hat{} \quad {}_2$$

Here, $\| \cdot \|_2$ denotes the Euclidean norm of the tensor. For a tensor , the Euclidean norm is computed as:

$$\| \|_2 = \sqrt{\sum_{i_1=1}^{1} \sum_{i_2=1}^{2} \sum_{i_3=1}^{3} {}^{2}_{i_1 i_2 i_3}}$$

In practice, as increases, the reconstruction error typically decreases because a larger rank allows the model to approximate the original data more accurately. However, continuing to increase beyond a certain value results in diminishing improvements, and at some point, the error reduction becomes negligible. Therefore, selecting the optimal rank involves finding a balance between the reduction in the reconstruction error and the model's complexity.

In addition, the variance of the reconstruction ratio R^2_{ratio} is computed as the ratio of the squared norm of the approximated tensor $\hat{}$ to the squared norm of the original tensor X:

$$R^2_{\text{ratio}} = \frac{\hat{}^{2}}{\| \|_2}$$

This ratio indicates how well the decomposition approximates the original tensor. The optimal rank can often be determined by observing both the reconstruction error and the variance of the reconstruction ratio for different values of r and identifying the point where further increases in lead to marginal or no improvement. This balance is crucial to avoid overfitting while still capturing meaningful data patterns. In practice, the elbow method is commonly used to identify this optimal point. This method involves plotting both the reconstruction error and variance of the reconstruction ratio against rank and selecting the rank where further increases result in diminishing returns in terms of reconstruction error reduction. In doing so, we ensure that the selected rank achieves a good trade-off between model complexity and the ability to explain the underlying data structure. In this study, we determined the optimal rank r by running PARAFAC2 for each rank, starting with ten random initializations, and selecting the rank with the lowest R_{error}. The corresponding R^2_{ratio} for the selected R_{error} was then used to identify the optimal rank.

In addition, B is the feature-factor matrix ($18 \times$) that describes the influence of linguistic features on the factors. Finally, is the diagonal matrix ($\times$) that shows the relationship between review types and factors, indicating the effect of each factor on genuine and fake reviews. Our algorithms were written in Python, using the PARAFAC2 function from the tensorly.decomposition module. All tests were performed on a computer with an 11th Gen Intel(R) Core(TM) i7–11700@2.50 GHz processor and 8.00 GB of RAM.

In the next section, we demonstrate how the results of this decomposition can be used to analyze hidden linguistic patterns in the text and gain deeper insights into the characteristics of spam opinions.

3.2.3. Strength Distance Matrix

Based on the component matrices A, B, and obtained through tensor decomposition, we calculated a distance matrix to better understand the relationships between linguistic and psychological features and review types. The main objective of this approach was to identify how strongly each feature influenced genuine or fake reviews. Component matrix B described the impact of each factor on specific linguistic and psychological features, while explained how each factor contributed to different review types (genuine or fake). To calculate these distances, we used the Euclidean distance, which previous studies

(Shin & Woo, 2019, 2022) have proposed as an effective method for identifying the relationships between components derived from tensor decomposition and target categories, aiding in the interpretability of high-dimensional data.

We generated a distance vector by computing the Euclidean distance between each feature vector and the review-type vector. Specifically, the feature vector $_{f_n}$ consisted of the factor values from matrix *B*, and the review-type vector comprised the diagonal component values from matrix . Mathematically, these vectors are defined as:

$$f_n = \quad 1,n, \quad 2,n, \quad , \quad ,n$$

$$= \quad ,1,1, \quad ,2,2, \quad , \quad , ,$$

Here, *n* denotes the number of features (18), represents the review type (genuine or fake), and is the rank selected during tensor decomposition. We then computed the Euclidean distance between each feature vector and the review-type vector to form the distance vector D_{f_n}:

$$D_{f_n} = \quad _1 - f_n \quad ^2, \quad _2 - f_n \quad ^2$$

This distance vector helped determine the review type in which each feature had a stronger influence. A smaller distance indicates that a feature has a stronger influence on the corresponding review type. The distance matrix is constructed by vertically combining all the distance vectors, where smaller distances highlight features with a greater influence on a particular review type. This analysis provided a clearer understanding of the linguistic patterns that distinguished genuine from fake reviews and quantitatively evaluated the importance of each feature.

Figure 3 visually explains the process of calculating the distance matrix, showing how the relationships between features and review types are determined using the component matrices from tensor decomposition. In this tensor decomposition, we set the rank to 4 (= 4), explaining the data structure using four factors. Matrix is shown at the top left of Figure 3, where the original matrix has been simplified to a two-dimensional form by isolating the diagonal elements. Matrix captures the influence of each factor on review strength. As shown, the genuine review group is influenced by Factor 1 with a value of 3.6436 and by Factor 2 with a value of 1.4957, whereas the fake review group is influenced by Factors 1 (3.9033) and 2 (1.4894). These values quantitatively describe the relationship between review strength groups and each factor.

The bottom left of Figure 3 presents matrix *B* , originally of dimensions $\times R$, where only 10 of the 18 features are displayed for clarity. Matrix *B* explains the relationship between the four factors and review features. In tensor decomposition, understanding the relative relationships among features is often more important than interpreting the exact values of the factors (Acar et al., 2005). This approach enables a comparison of the influences of different factors on each feature.

We then use the Euclidean distance to calculate the strength distance matrix and analyze how close each feature is to the genuine and fake review groups. The right side of Figure 3 displays the strength distance matrix for 10 features, with the red dashed lines indicating the vectors from and $B_{\quad _1}$, f_1 used to calculate the proximity of the first feature to the genuine review group. This distance matrix measures how close each

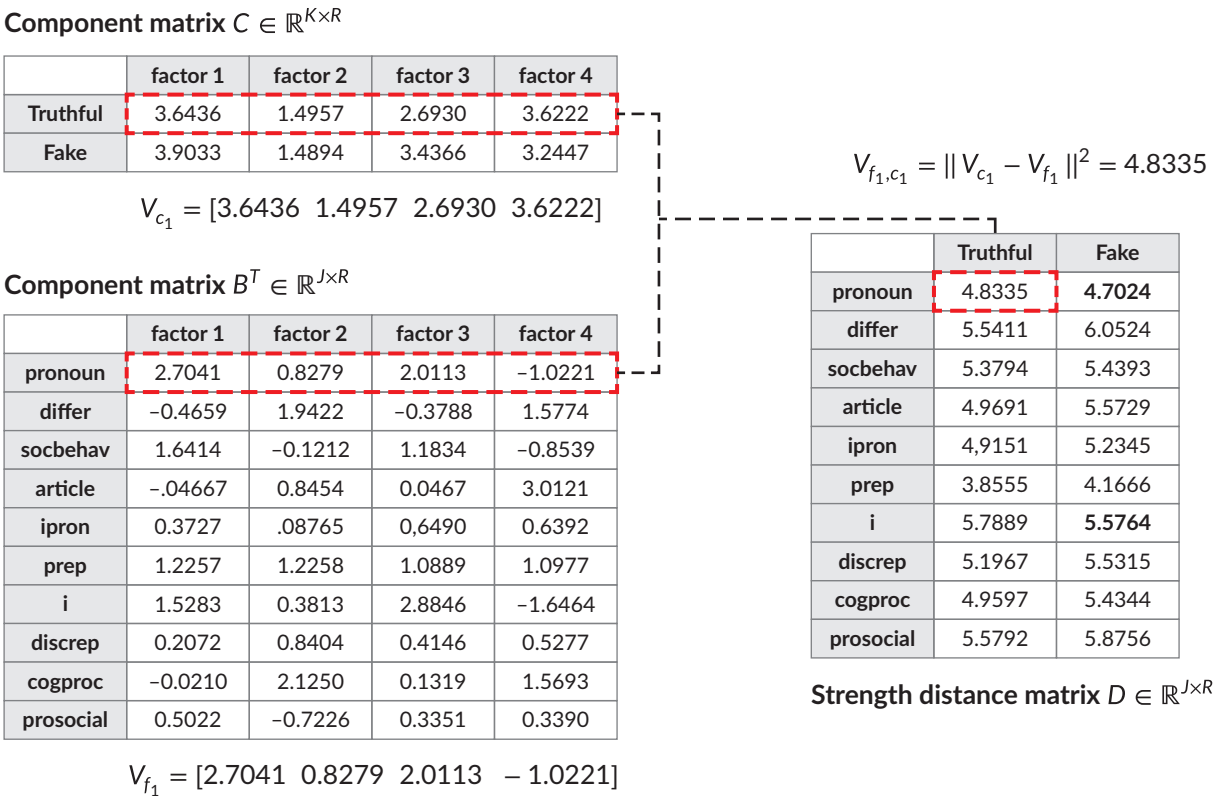


Figure 3. Calculation of the strength distance matrix.

feature vector $_f$ is to the vectors of the two review strength groups, demonstrating that smaller distances indicate a stronger influence of that feature on the review strength group.

3.2.4. Interpreting the Results

In the previous section, we presented a subset of the distance matrix results for illustration. Here, we provide complete results for all 18 linguistic features (Table 1). Table 1 quantitatively displays the Euclidean distances between each feature and the two review groups (truthful and fake). Each row contains the distance values for a specific feature. The final column of Table 1 presents the differences in distances for each feature between the genuine (truthful) and fake (fake) review groups. This allows for a quick assessment of the review group on which each feature has the strongest influence. The values highlighted with an asterisk indicate the minimum distance, signifying that the feature had the strongest influence on the corresponding review group.

To interpret, the analysis revealed that features such as pronouns (9.5034), personal pronouns (Personal pronouns; 9.5862), and first-person singular pronouns (First person singular; 9.7629) are prominent indicators of fake reviews. This suggests that fake reviews often emphasize personal and self-centered language, possibly reflecting an attempt to narrow the psychological distance with readers and appear more credible (c.f., Hancock et al., 2010; Newman et al., 2003).

Table 1. Distance matrix between 18 features and review type.

Features	Truthful	Fake	Difference
Impersonal pronouns	9.7766*	9.9235	−0.1469
Common adjectives	9.4235*	9.7097	−0.2862
Conjunctions	9.3392*	9.5170	−0.1778
Determiners	8.8372*	9.0319	−0.1947
Personal pronouns	9.6539	9.5862*	0.0677
Total pronouns	9.5272	9.5034*	0.0238
Adverbs	9.7837*	9.9726	−0.1889
First person singular (I)	9.8060	9.7629*	0.0431
First person plural (we)	10.2994*	10.4288	−0.1294
Third person plural (they)	10.0790*	10.1654	−0.0864
Articles	8.9768*	9.2555	−0.2787
Auxiliary verbs	9.7301*	9.9725	−0.2424
Quantities	9.2445*	9.4659	−0.2214
Verbs	9.5149*	9.6614	−0.1465
Second person (you)	10.0196*	10.1890	−0.1694
Negations	10.0051*	10.1517	−0.1466
Prepositions	8.9532*	9.0083	−0.0552
Total function words	9.1715*	9.3350	−0.1635

Note: * = The shorter distance between each feature and either truthful or fake.

4. Discussion

This study aims to offer a guide for social science communication researchers on how to apply a tensor-decomposition-based machine learning approach to the analysis of high-dimensional data, providing interpretable results. For illustration purposes, we systematically examined the linguistic features of fake reviews using a large-scale reviews dataset. Initially, we extracted the linguistic features using the LIWC tool. Following data normalization to ensure consistency, the tensor was decomposed using the PARAFAC2 algorithm. We then compared the influence of each feature on genuine and fake reviews using the Euclidean distance analysis. This comprehensive approach allowed us to quantify and understand the complex relationships between reviews, revealing that linguistic features such as pronouns, personal pronouns, and first-person singular pronouns were prominent in fake reviews. These features might be strategically used to foster intimacy with the reader or enhance emotional appeal. The insights gained from such interpretable results can be used to understand persuasive strategies, communication patterns, and other related aspects.

The versatility of the PARAFAC2 model should be particularly emphasized. This model offers significant flexibility in two ways. First, tensor-based representations can be applied to diverse domains. For instance, Acar et al. (2005) analyzed online chatroom data using a three-dimensional tensor (user–keyword–time) to track the evolution of social groups. Such tensor-based approaches are valuable for analyzing complex relationships in social media, news recommendation systems, and real-time communication networks, and have broad applicability in communication and data science (Bader & Kolda, 2006). Second, PARAFAC2 can handle tensors with varying dimensions along one mode, making it suitable for datasets of different sizes.

We used this model to conduct a nuanced analysis of the linguistic features in reviews, which can be applied to multidimensional studies, such as evaluating review credibility and analyzing advertising effectiveness in multi-label contexts.

This study focused on exploring the application of tensor decomposition to improve the interpretability of machine-learning models, particularly in the analysis of linguistic features in text data. By demonstrating the effectiveness of this approach, we highlighted its potential to improve the interpretability of detection algorithms and provide insights into complex data patterns. Our methodology emphasizes the importance of explainability in machine learning, offering a framework that can be adapted to various applications that require transparent and comprehensible analysis. Future research can further refine this approach by exploring how tensor decomposition models can be optimized to balance accuracy with interpretability across diverse data environments.

From a methodological perspective, determining the optimal rank for tensor decomposition is crucial, as it significantly affects the performance and accuracy of the algorithm. Although Section 3 in this article does not delve deeply into rank optimization, selecting an appropriate rank is essential to accurately represent data patterns and prevent overfitting. The Frobenius norm is commonly used to minimize reconstruction error (Kolda & Bader, 2009), and in our research, we adopted the elbow method to identify the optimal rank, which provides an intuitive and straightforward guideline by pinpointing inflection points on residual plots. For more detailed techniques of rank selection, readers may refer to Kolda and Bader (2009) and Cichocki et al. (2015).

Based on this study, there are several promising avenues for future research. While the Euclidean distance was chosen for its interpretability, further studies may investigate alternative measures for distance matrix computation, such as the dot product between component matrices B and , which could provide additional insights. In addition, we implemented the standard PARAFAC2 algorithm, but some aspects can be applied more precisely. Specifically, the Python implementation that we used does not guarantee a unique solution, and there are various other ways to address this issue. For example, methods such as those proposed by Kiers et al. (1999) could be explored to improve the uniqueness of the results. Furthermore, to enhance both predictive power and interpretability, future studies could investigate the interactions between linguistic features particularly in the context of addressing the curse of dimensionality, which arises due to the exponential growth of parameters (e.g., Govindarajan et al., 2022; Novikov et al., 2017). Another promising direction could involve exploring the use of tensors to assess the linguistic similarity between fake and real texts without relying on decomposition, and utilizing faster and more numerically stable algorithms (Van Eeghem & De Lathauwer, 2020).

Conflict of Interests

The authors declare no conflicts of interests.

Data Availability

The codes and complete experimental results are available upon request from interested researchers. This ensures transparency and facilitates reproducibility while maintaining ethical considerations regarding data sharing and implementation.

References

Acar, E., Çamtepe, S. A., Krishnamoorthy, M. S., & Yener, B. (2005). Modeling and multiway analysis of chatroom tensors. In P. Kantor, G. Muresan, F. Roberts, D. D. Zeng, F.-Y. Wang, H. Chen, & R. C. Merkle (Eds.), *Intelligence and security informatics* (pp. 256–268). Springer.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Ser, J. D., Diaz-Rodriguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, Article 101805. https://doi.org/10.1016/j.inffus.2023.101805

Bader, B. W., Berry, M. W., & Browne, M. (2008). Discussion tracking in Enron email using PARAFAC. In M. W. Berry & M. Castellanos (Eds.), *Survey of text mining II: Clustering, classification, and retrieval* (pp. 147–163). Springer.

Bader, B. W., & Kolda, T. G. (2006). Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software (TOMS)*, *32*(4), 635–653.

Beylkin, G., & Mohlenkamp, M. J. (2005). Algorithms for numerical analysis in high dimensions. *SIAM Journal on Scientific Computing*, *26*(6), 2133–2159.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. University of Texas at Austin. https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, *35*(3), 283–319.

Cichocki, A., Lee, N., Oseledets, I., Phan, A. H., Zhao, Q., & Mandic, D. P. (2015). Tensor networks for dimensionality reduction and large-scale optimization: Part 2 applications and future perspectives. *Foundations and Trends in Machine Learning*, *9*(6), 431–673.

Dobbrick, T., Jakob, J., Chan, C. H., & Wessler, H. (2022). Enhancing theory-informed dictionary approaches with "glass-box" machine learning: The case of integrative complexity in social media comments. *Communication Methods and Measures*, *16*(4), 303–320.

Govindarajan, N., Vervliet, N., & De Lathauwer, L. (2022). Regression and classification with spline-based separable expansions. *Frontiers in big Data*, *5*, Article 688496.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2010). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, *45*(1), 1–23.

Harshman, R. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, *16*, 1–84.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Kiers, H. A., Ten Berge, J. M., & Bro, R. (1999). PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics*, *13*(3/4), 275–294.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500.

Kolda, T. G., Bader, B. W., & Kenny, J. P. (2005). Higher-order web link analysis using multilinear algebra. In J. Han, B. W. Wah, V. Raghavan, X. Wu, & R. Rastogi (Eds.), *Proceedings of the 5th IEEE International Conference on Data Mining* (pp. 242–249). IEEE.

Kolda, T. G., & Sun, J. (2008). Scalable tensor decompositions for multi-aspect data mining. In F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan, & X. Wu (Eds.), *2008 Eighth IEEE International Conference on Data Mining* (pp. 363–372). IEEE.

Kroon, A., Welbers, K., Trilling, D., & van Atteveldt, W. (2024). Advancing automated content analysis for a new era of media effects research: The key role of transfer learning. *Communication Methods and Measures*, *18*(2), 142–162. https://doi.org/10.1080/19312458.2023.2261372

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665–675.

Novikov, A., Trofimov, M., & Oseledets, I. (2017). *Exponential machines*. arXiv. https://doi.org/10.48550/arXiv.1605.03795

Oh, Y. W., & Park, C. H. (2021). Machine cleaning of online opinion spam: Developing a machine-learning algorithm for detecting deceptive comments. *American Behavioral Scientist*, *65*(2), 389–403. https://doi.org/10.1177/0002764219878238

Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam. In L. Vanderwende (Ed.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 497–501). Association for Computational Linguistics.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In D. Lin (Ed.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 309–319). Association for Computational Linguistics.

Papalexakis, E. E., Faloutsos, C., & Sidiropoulos, N. D. (2016). Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(2), 1–44.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*, 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Saha, A., & Sindhwani, V. (2012). Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization. In E. Adar & J. Teevan (Eds.), *WSDM '12: Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 693–702). ACM.

Schuld, M., Durrheim, K., & Mafunda, M. (2023). Speaker landscapes: Machine learning opens a window on the everyday language of opinion. *Communication Methods and Measures*, 18(4), 315–331.

Shi, Q., Cheung, Y. M., Zhao, Q., & Lu, H. (2018). Feature extraction for incomplete data via low-rank tensor decomposition with feature regularization. *IEEE transactions on neural networks and learning systems*, *30*(6), 1803–1817.

Shin, Y., & Woo, S. S. (2019). What is in your password? Analyzing memorable and secure passwords using a tensor decomposition. In L. Liu & R. White (Eds.), *WWW '19: The World Wide Web Conference* (pp. 3230–3236). ACM.

Shin, Y., & Woo, S. S. (2022). PasswordTensor: Analyzing and explaining password strength using tensor decomposition. *Computers & Security*, *116*, Article 102634.

Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., & Faloutsos, C. (2017). Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, *65*(13), 3551–3582.

Sidiropoulos, N. D., Giannakis, G. B., & Bro, R. (2000). Blind PARAFAC receivers for DS-CDMA systems. *IEEE Transactions on Signal Processing*, *48*(3), 810–823.

Smilde, A. K., Bro, R., & Geladi, P. (2004). *Multi-way analysis: Applications in the chemical sciences*. Wiley.

Stoll, A., Wilms, L., & Ziegele, M. (2023). Developing an incivility dictionary for German online discussions—A semi-automated approach combining human and artificial knowledge. *Communication Methods and Measures*, *17*(2), 131–149.

Sun, J.-T., Zeng, H.-J., Liu, H., Lu, Y., & Chen, Z. (2005). CubeSVD: A novel approach to personalized Web search. In A. Ellis & T. Hagino (Eds.), *WWW '05: Proceedings of the 14th international conference on World Wide Web* (pp. 382–390). ACM.

Tanvir, A. A., Mahir, E. M., Akhter, S., & Huq, M. R. (2019). Detecting fake news using machine learning and deep learning algorithms. In L. Gopal, D. Pradhan, & A. Singh (Eds.), *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICSCC.2019.8843612

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*(3), 279–311.

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140.

Van Eeghem, F., & De Lathauwer, L. (2020). Tensor similarity in chemometrics. In S. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive chemometrics: Chemical and biochemical data analysis* (pp. 337–354). Elsevier.

Vasilescu, M. A. O., & Terzopoulos, D. (2002). Multilinear analysis of image ensembles: Tensorfaces. In A. Heyden, G. Sparr, M. Nielsen, & P. Johansen (Eds.), *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision, Part I* (pp. 447–460). Springer.

Wang, J., Zhang, S., Li, H., & Chen, Y. (2023). A tensor factorization approach for personalized recommendation with contextual information. *Journal of Information Science*, *49*(1), 85–98.

Zini, J. E., & Awad, M. (2022). On the explainability of natural language processing deep models. *ACM Computing Surveys*, *55*(5), 1–31.

About the Authors

Yu Won Oh (PhD, University of Michigan, 2015) is an associate professor in the School of Digital Media at Myongji University in Korea and the associate director of the Debiasing and Lay Informatics Lab in Oklahoma, USA. Her current research interests include the intersection of new media and civic communication with an emphasis on opinion expression, misinformation, issue development, and big data analytics.

Chong Hyun Park (PhD, Purdue University, 2016) is an associate professor in the School of Business at Sungkyunkwan University. His research interests include mathematical programming and machine learning modeling. He has been conducting interdisciplinary research that attempts to solve various social problems. He recently developed a machine learning algorithm that can detect manipulated opinion spam in the comment section. He has published research articles in *Production and Operations Management*, *European Journal of Operations Research*, *American Behavioral Scientists*, *Journalism and Mass Communication Quarterly*, etc.

MEDIA AND COMMUNICATION
ISSN: 2183-2439

Media and Communication is an international, peer-reviewed open access journal dedicated to a wide variety of basic and applied research in communication and its related fields. It aims at providing a research forum on the social and cultural relevance of media and communication processes.

The journal is concerned with the social development and contemporary transformation of media and communication and critically reflects on their interdependence with global, individual, media, digital, economic and visual processes of change and innovation.

cogitatio

www.cogitatiopress.com/mediaandcommunication