RESEARCH-ARTICLE

Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy

KOKIL JAIDKA, National University of Singapore, Singapore City, Singapore

TSUHAN CHEN, National University of Singapore, Singapore City, Singapore

SIMON CHESTERMAN, National University of Singapore, Singapore City, Singapore

WYNNE HSU, National University of Singapore, Singapore City, Singapore

MIN YEN KAN, National University of Singapore, Singapore City, Singapore

MOHAN KANKANHALLI, National University of Singapore, Singapore City, Singapore

View all

Open Access Support provided by:

National University of Singapore

Lee Kuan Yew School of Public Policy

Published: 12 February 2025
Online AM: 23 August 2024
Accepted: 07 August 2024
Revised: 15 June 2024
Received: 02 February 2024

Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy

KOKIL JAIDKA, Department of Communications and New Media, National University of Singapore, Singapore, Singapore
TSUHAN CHEN, School of Computing, National University of Singapore, Singapore, Singapore
SIMON CHESTERMAN, Faculty of Law, National University of Singapore, Singapore, Singapore
WYNNE HSU, School of Computing, National University of Singapore, Singapore, Singapore
MIN-YEN KAN, School of Computing, National University of Singapore, Singapore, Singapore
MOHAN KANKANHALLI, School of Computing, National University of Singapore, Singapore, Singapore
MONG LI LEE, School of Computing, National University of Singapore, Singapore, Singapore
GYULA SERES, N.1 Institute for Health and Institute for Digital Medicine, National University of Singapore, Singapore, Singapore
TERENCE SIM, School of Computing, National University of Singapore, Singapore, Singapore
ARAZ TAEIHAGH, Lee Kuan Yew School of Public Policy, National University of Singapore, Singapore, Singapore
ANTHONY TUNG, School of Computing, National University of Singapore, Singapore, Singapore
XIAOKUI XIAO, School of Computing, National University of Singapore, Singapore, Singapore
AUDREY YUE, Department of Communications and New Media, National University of Singapore, Singapore, Singapore

The emergence of generative artificial intelligence (GenAI) has exacerbated the challenges of misinformation, disinformation, and mal-information (MDM) within digital ecosystems. These multi-faceted challenges demand a re-evaluation of the

The iGyro Project, hosted at the NUS Centre for Trusted Internet and Community, is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).
Authors' Contact Information: Kokil Jaidka, Department of Communications and New Media, National University of Singapore, Singapore, Singapore; e-mail: kokil.jaidka@gmail.com; Tsuhan Chen, School of Computing, National University of Singapore, Singapore, Singapore; e-mail: tsuhan@nus.edu.sg; Simon Chesterman, Faculty of Law, National University of Singapore, Singapore, Singapore; e-mail: chesterman@nus.edu.sg; Wynne Hsu, School of Computing, National University of Singapore, Singapore, Singapore; e-mail: whsu@comp.nus.edu.sg; Min-Yen Kan, School of Computing, National University of Singapore, Singapore, Singapore; e-mail: kanmy@comp.nus.edu.sg; Mohan Kankanhalli, School of Computing, National University of Singapore, Singapore, Singapore; e-mail: mohan@comp.nus.edu.sg; Mong Li Lee, School of Computing, National University of Singapore, Singapore, Singapore; e-mail: leeml@comp.nus.edu.sg; Gyula Seres, N.1 Institute for Health and Institute for Digital Medicine, National University of Singapore, Singapore, Singapore; e-mail: gyula@nus.edu.sg; Terence Sim, School of Computing, National University of Singapore, Singapore, Singapore; e-mail: terence.sim@nus.edu.sg; Araz Taeihagh, Lee Kuan Yew School of Public Policy, National University of Singapore, Singapore, Singapore; e-mail: spparaz@nus.edu.sg; Anthony Tung, School of Computing, National University of Singapore, Singapore, Singapore; e-mail: atung@comp.nus.edu.sg; Xiaokui Xiao, School of Computing, National University of Singapore, Singapore, Singapore; e-mail: xkxiao@nus.edu.sg; Audrey Yue, Department of Communications and New Media, National University of Singapore, Singapore, Singapore; e-mail: audrey.yue@nus.edu.sg.

ACM 2639-0175/2025/02-ART11
https://doi.org/10.1145/3689372

digital information lifecycle and a deep understanding of its social impact. An interdisciplinary strategy integrating insights from technology, social sciences, and policy analysis is crucial to address these issues effectively. This article introduces a three-tiered framework to scrutinize the lifecycle of GenAI-driven content from creation to consumption, emphasizing the consumer perspective. We examine the dynamics of consumer behavior that drive interactions with MDM, pinpoints vulnerabilities in the information dissemination process, and advocates for adaptive, evidence-based policies. Our interdisciplinary methodology aims to bolster information integrity and fortify public trust, equipping digital societies to manage the complexities of GenAI and proactively address the evolving challenges of digital misinformation. We conclude by discussing how GenAI can be leveraged to combat MDM, thereby creating a reflective cycle of technological advancement and mitigation.

CCS Concepts: • **Computing methodologies → Artificial intelligence**; • **Human-centered computing → Collaborative and social computing**; • **Social and professional topics → Computing/technology policy;**

Additional Key Words and Phrases: Misinformation, disinformation, trust, resilience, generative AI, social media

ACM Reference Format:
Kokil Jaidka, Tsuhan Chen, Simon Chesterman, Wynne Hsu, Min-Yen Kan, Mohan Kankanhalli, Mong Li Lee, Gyula Seres, Terence Sim, Araz Taeihagh, Anthony Tung, Xiaokui Xiao, and Audrey Yue. 2025. Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy. *Digit. Gov. Res. Pract.* 6, 1, Article 11 (February 2025), 15 pages. https://doi.org/10.1145/3689372

1 Introduction

In the era of digital ubiquity characterized by the widespread use of the social media platforms and mobile technology, the ways in which individuals create, consume, and disseminate information have undergone a profound transformation. Today's digital landscape demands immediate delivery of created content, with decisions and opinions increasingly shaped by isolated consumption of algorithmically curated feeds. This shift from traditional expert curation in news dissemination to personalized, user-specific feeds exemplifies a radical change in how information is created, tailored, and consumed, fostering an economy of digital consumerism dominated by an often unregulated flow of information. This transformation has precipitated the digital information paradox: an era where the ability to access created information is unparalleled in human history. Yet, trust in the accuracy of consumed information is notably declining. Blending journalism, advertising, and entertainment complicates the clarity of information creation and the criteria for its dissemination, obscuring the intentions behind content production. Consequently, the proliferation of dubious information has fueled a global crisis in the triad of **misinformation, disinformation, and mal-information (MDM)**: misinformation, involving the unintended sharing of false data; disinformation, reflecting the deliberate spread of misleading content to further specific agendas; and mal-information, designed to cause societal harm by disseminating harmful truths or falsehoods. In response to these pervasive challenges in information creation, consumption, and dissemination, governments worldwide are compelled to implement new regulations and adapt existing frameworks to manage the flow and integrity of information better, although effective legislation remains a critical need on the horizon [8, 28].

Amid this landscape, the advent of **Generative Artificial Intelligence (GenAI)** threatens to exacerbate the risks associated with MDM. GenAI models are capable of creating synthetic multimedia content that simulates the characteristics and sensibilities of content featuring or created by humans, thereby introducing new dimensions to the creation and spread of MDM [10]. Thousands of user-developed free software and web applications now allow individuals to generate high-quality synthetic portraits and videos, also known as deepfakes [36], that feature politicians, celebrities, and regular citizens saying and doing things that never happened, while others allow the synthesis of coherent and persuasive text in support of any given topic. Consequently, three factors make GenAI especially critical to study in the context of MDM. The ease of access to GenAI tools enables the mass production of deepfakes and persuasive texts, raising three critical concerns. First, GenAI's capacity for generating convincing fake content makes source attribution increasingly challenging. Second, the democratization

of content creation tools lowers the barriers to generating and disseminating MDM, complicating efforts to distinguish genuine from fake sources. Third, the illusory truth effect associated with GenAI fosters a heightened level of media skepticism, even toward reputable sources, eroding trust and societal cohesion [2]. Despite the advancements in detecting and authenticating MDM, the responsibility has largely fallen on developers, with less attention paid to the human aspects of information consumption and response.

It is imperative to adopt a systemic approach that not only addresses the technical creation and detection of MDM but also its consumption and dissemination. A technical strategy should address the technical creation and detection of MDM; yet, its consumption and dissemination still pose risks to stable societies [18]. Trust remains a fundamental component in instilling digital resilience and combating the challenges MDM poses in the digital era. Therefore, what is needed is an approach that leverages insights from multiple disciplines so that even technical perspectives are sensitive to how digital misinformation is consumed and disseminated.

We discuss a three-tier consumer behavior model that seeks to identify *how* existing checks in the digital information pipeline are bypassed in the creation of digital MDM, determine *why* the consumption and dissemination of MDM takes place, and evaluate *where* proposed resilience strategies could mitigate existing vulnerabilities and preempt future ones. This approach underscores the need for a holistic perspective encompassing computational methods and insights into consumer motivations and reactions within the digital information ecosystem. We argue that these findings should guide the development of regulatory and governance frameworks, emphasizing the critical role of trust in fostering digital resilience and countering the threats posed by MDM in the digital age.

To address the identified gaps in understanding the impact of GenAI on MDM and the effectiveness of existing countermeasures, our proposed model addresses three specific research questions:

— How does GenAI facilitate bypassing established checks within the digital information pipeline, contributing to the spread of MDM?
— Why do consumers engage with and disseminate MDM, and what are the underlying motivations and decision-making processes involved?
— Where can resilience strategies be most effectively implemented to mitigate the vulnerabilities exposed by GenAI and prevent future escalations of MDM-related challenges?

Insights from our findings will guide the development of regulatory and governance frameworks that emphasize the critical role of trust in fostering digital resilience and mitigating the threats MDM poses. By integrating behavioral science, technology, and policy analysis, our research paves the way for creating effective strategies that enhance the integrity of information and strengthen public trust. The framework will also set the stage for future research that will adapt to and anticipate the evolving complexities of digital misinformation. Through this ongoing exploration, we aim to contribute significantly to building a more digitally resilient society.

2 The Landscape of GenAI and Information Integrity

GenAI has made content creation accessible to a wider audience through interfaces that require minimal technical skill. The advantages of GenAI extend across various domains, enabling users to generate engaging content, synthesize information from diverse sources, and translate complex data into comprehensible summaries. These capabilities lower the barriers to information access and creation, potentially narrowing the digital divide and enriching public discourse in fields ranging from health and medicine to policy, agriculture, and education [1, 45, 46, 71, 72].

However, the reliance on GenAI for content creation and synthesis also has its pitfalls. The output of GenAI systems is contingent on their training data, which may not be representative or transparent, leading to the generation of content that can unintentionally mislead by presenting unverified or biased information as credible [11, 13, 50]. This issue is compounded by the prevalence of MDM across digital platforms, where the challenge for consumers increasingly lies in distinguishing AI-generated fabrications from authentic information. The sophistication of GenAI in mimicking human-like content amplifies this challenge, necessitating enhanced digital literacy

and critical thinking skills among consumers to navigate the complex digital information landscape. It is therefore essential to consider the MDM implications of GenAI. For instance, while the ability of GenAI to create MDM is unparalleled, so is its ability to tailor MDM to social media users and formats, thereby increasing its likelihood of virulence, consumption, and amplification over time. GenAI allows for interactive chatbots that are tailored to individual biases and preferences and are reportedly more persuasive than humans, based on a recent study [65].

Focusing on the MDM implications of the GenAI ecosystem for its consumers is increasingly imperative. Therefore, our research objectives aim at first dissecting how GenAI contributes to the proliferation of MDM and then exploring strategies for fostering resilience against such threats. Furthermore, given GenAI's immediate implications for the end user, mitigating MDM in GenAI requires a holistic approach that spans the information lifecycle, ensuring content integrity from creation to consumption and fostering a digital landscape resilient to the threats posed by sophisticated AI-driven misinformation campaigns.

3 Consumer Interaction with MDM: A Three-Tiered Approach

Research on GenAI creation, consumption, and dissemination of MDM needs to translate across contexts, languages, and cultures. Policies to improve AI safety also need to be grounded in evidence-based analyses of technology and user behavior. These considerations have spurred the creation of the **Information Gyroscope (iGyro)** project—a consortium of marketing scientists, computer scientists, social scientists, lawyers, and policymakers all examining the multi-faceted paradigms of MDM within GenAI. It draws upon a diverse range of disciplinary insights and real-world applications, epitomizing the kind of systemic, interdisciplinary research approach that is essential for addressing the complex and evolving challenges of GenAI and MDM.

The iGyro project at the National University of Singapore is the first comprehensive approach to address consumer interaction with MDM. Figure 1 shows the conceptual framework comprising three interconnected research spheres, each enhancing resilience and trust in the digital information ecosystem. Our objectives and approaches were developed in deep discussion with policy advisors, technical experts, civil servants, engineers, researchers, social activists, and grassroots volunteers. This ongoing engagement with a broad spectrum of disciplines within the iGyro project is designed to continuously refine and enhance our understanding and responses to the challenges posed by MDM. By integrating new disciplines and expanding our research questions, we aim to keep our approach at the forefront of technological and societal developments, ensuring that our strategies remain relevant and effective in rapid technological change. Furthermore, as the challenges in the digital information ecosystem evolve, keeping up with the state of the art will facilitate the development of more effective, context-sensitive solutions for digital resilience.

At the core of the iGyro framework is Sphere 1, which investigates the motivations and decision-making processes governing how consumers seek, process, and share information. We employ behavioral economics methods such as Bayes' theorem and decision-theoretic modeling to gain insights into consumer behavior.

Sphere 2 is divided into three **technology domains (TDs)**, each representing different stages of the digital information pipeline. TD1 investigates the creation of MDM in text and visual media, focusing on vulnerabilities that allow false information to appear legitimate. TD2 examines the dissemination of MDM, mainly how it exploits consumer biases and belief systems. This includes the study of algorithms in recommender systems and search engines based on consumers' historical or social media behavior and their impact on opinion polarization and the formation of echo chambers. TD3 focuses on consuming digital information, particularly on social media platforms. This domain explores how information is consumed and strategies to mitigate vulnerabilities, emphasizing enhancing consumer reasoning and empowerment.

Finally, Sphere 3 studies the potential impact of mitigation strategies and interventions on human and community behavior and considers the role of regulation or policy in deploying these strategies at a population level to nudge consumer behavior. These research spheres build resilience and trust in the digital information lifecycle.

Compared to prior work, centers and labs focusing on one aspect of the MDM problem have reported exemplary work at studying the behavioral, detective, or policy aspects of MDM. However, to our knowledge, no

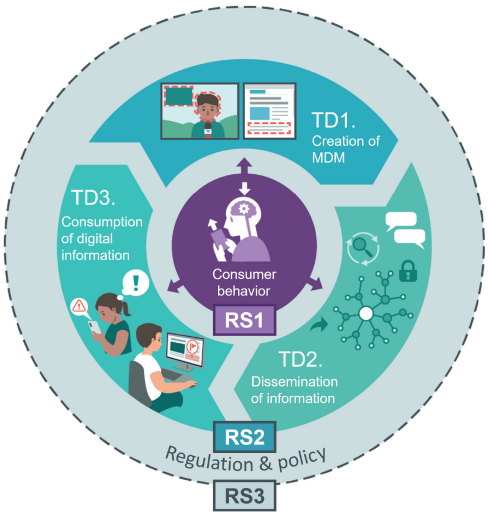

Fig. 1. Conceptual framework of iGyro, showing the three research spheres of consumer behavior, digital information lifecycle, and regulation and policy.

similar interdisciplinary effort spanning behavior, technology, social science, and policy exists anywhere else. iGyro incorporates collaboration across institutions to further embody links among related research spheres. For example, studies of belief formation led by Professor Juanjuan Zhang at the MIT Sloan School of Management and Professor Heidi Larson at the London School of Hygiene & Tropical Medicine focus on information engagement and sharing dynamics, and identifying the levers of trust and distrust. Studies of MDM detection are the expertise of Professor Preslav Nakov at the Mohamed bin Zayed University of Artificial Intelligence, a key collaborator and lead on the research and development of several fact-checking tools. We also note our collaboration with Professor Martin De Jong from the Erasmus School of Law and Professor Jeannie Marie Paterson from the University of Melbourne, in the space of AI and digital ethics and the examination of governance challenges related to digital technology access across vulnerable groups.

4 Sphere 1: Understanding MDM Consumption

Understanding the interplay between consumer behavior and vulnerabilities in the digital information pipeline is essential for fostering long-term resilience against MDM. The first part of the iGyro model is grounded in in behavioral economics principles to analyze how consumers process and act upon information. Consumers navigate the complex digital information landscape under conditions that deviate from the ideal of i.i.d. (independent and identically distributed) information draws. Unlike the printed press, they make a complex set of choices, including consumption and information sharing. Herein lies the necessity to consider information's correlation, duplication, and authenticity, recognizing that information consumption often involves encounters with modified, repeated, or false data. The implications of these non-i.i.d. scenarios on belief formation and information sharing are profound, challenging the rational actor model traditionally assumed in information theory. Therefore, we consider whether people need to adjust or under-adjust when information sources are not independent. Prior research examined how correlated and fake information results in incorrect belief formation using laboratory [20, 21] and observational experiments [60]. In these cases, we strive to understand how people account and adjust for potential conflicts of interest or biases in information sources and determine whether information sharing is affected by observing a conflict of interest. Many reputed information sources contain "both sides of the story." This has been a long-standing influence of integrity [16] that is regaining prominence with the popularity of social media-based influencers and review-style information.

In Sphere 1 of our consumer behavior model, we use the lens of consumer behavior to examine (1) how consumer beliefs are shaped by the type and veracity of the information they receive and (2) how consumers account for the possibility that the information they share might be false and that other people may propagate such incorrect information. In prior work, studies have addressed similar problem statements [37, 44] using observational data to identify exogenous factors, such as online trust and social media fatigue. In contrast, our study allows us to establish causal effects by examining information sharing using high-powered controlled experiments with elicited beliefs about the veracity of information. Randomized trials yield more precise estimates of the effects of the information setting, enabling the comparison of potential policy interventions.

First, our investigation focuses on the factors guiding information consumption preferences and the tradeoff between information diversity and confirmation bias exhibited through information choices. Through cross-national survey experiments, we address the first key question: Do people choose information from multiple independent sources, or do they focus their search on confirming prior sources of information? Insights from this analysis contribute to our understanding of the mechanisms fostering echo chambers and bias reinforcement online. Empirical evidence motivating our work includes findings on social structures and reputation in expert review systems [14], highlighting the intricacies of credibility and trust in digital platforms. As a major stepping stone, we develop a novel interdisciplinary approach to experimentally estimate people's willingness to share information online, using elicitation methods from behavioral economics [4, 7].

A second project in this space explores the motivations underlying the decision to share information, even when its accuracy is questionable. Through controlled experiments designed to elicit beliefs about information's truthfulness, we aim to bridge this gap, drawing on prior work by the team. For instance, in a prior randomized field experiment, we reported insights into behavior modification in response to team dynamics and incentive structures [41]. The study design will also build on the foundational work by Noussair and Seres [51] on the efficiency of information sharing. Based on the insights from Ivanova-Stenzel and Seres [29] on information and bounded consumer rationality, we propose further investigating how economic models can inform the understanding of information credibility assessment in digital environments.

In summary, through Sphere 1, we aim to establish a new framework for understanding MDM that integrates economic incentives, such as financial and reputational gains from information sharing, and social structures, such as community norms and network effects, to understand how consumers engage with misinformation. This integrated approach will also help us craft effective and sustainable solutions to thwart MDM by addressing individual behaviors and the economic and social dynamics that fuel the spread of false information.

5 Sphere 2: Understanding MDM Creation

Sphere 2 is the intermediary, technical sphere that tracks the trajectory of MDM through its creation, dissemination, and consumption. It serves as a bridge between the behavioral insights from Sphere 1 with Sphere 3, preparing the ground for regulatory and policy interventions. It operates at the intersection of MDM's contrast with authentic knowledge and thereby unearths factors influencing information authenticity, lays the groundwork for detecting misrepresentations, and pioneers new algorithms for nuanced news classification and social recommendations that enrich information diversity. Sphere 2's contribution is integral, setting the stage for advanced context mapping via knowledge graphs and enhancing multi-hop reasoning capabilities, all while correlating information styles with users' cognitive faculties. The following paragraphs describe the research scope of Sphere 2, which closely follows the MDM lifecycle.

5.1 Detecting MDM

Our technical approach to MDM detection will integrate behavioral science to better understand consumer responses to detected misinformation. The first TD, TD1, of Sphere 2 focuses on detecting MDM, following two core objectives: understanding the factors influencing the authenticity of information and developing a comprehensive framework for detecting misrepresentation. The generation of MDM and its detection can be thought

of as inverses of each other, with progress in generation technology preceding detection technology. As such, these twin aspects are pitted against each other in an adversarial, co-evolving relationship.

GenAI technologies initially sought to create believable single-modality media: text, images, or others that could pass as natural sources. Subsequently, both legitimate red-teaming researchers and malignant actors harnessed such general-purpose generation technologies to create MDM, especially in high-impact domains such as politics. As a foil, initial MDM detection technologies harnessed data mining perspectives using signals gleaned from knowledge graphs, social communities, and accounting for temporal spread abound [38, 54, 75]. These technologies examine telltale signs of MDM on these specific dimensions, often relying on sophisticated deep learning models to increase efficacy [34, 49, 58, 77].

Modern MDM generation is hybridizing, where the veracity of one modality lends credibility to another. MDM detection in iGyro handles (a) text fabrication (falsified headlines with authentic visual content) and (b) misrepresentation (truthful content headline but with irrelevant visual content), alongside (c) complete textual and visual fabrication. Doctoring modalities only at critical points is also common. Our core approach detects such "inconsistencies" which manifest at different levels: signals (e.g., cut & paste boundaries or compression differences), objects (e.g., object or sentence feature differences), and semantics (e.g., differences arising from mixing of two different real-world events). Our approach addresses the text modalities of MDM in ways that leverage decomposing claims into atomic, easily verified claims [55, 56]. For visual content, we examine the physical signal aspects of images and other visual media at the object level. At the same time, for videos, we exploit temporal information, consistency, and constraints and develop image forensics and machine learning based methods to tackle full-body fakes. In recent work, we have examined how synthetic speech for key words can replace an original speech signal [64] and well as how voice authentication systems can be successfully compromised with a single face image [33]. Claims that rely on multiple component facts can turn MDM by falsifying only one part. We have also reported how critical parts of natural images can be replaced with parts from others "pasted in" [82]. The characteristics of the detected misinformation will be applied to design stimuli for field experiments to understand better how consumers engage with and form perceptions of MDM, as discussed in TD2.

Going forward, the interplay between these detections and consumer perceptions will feed into developing more effective, consumer-informed technologies. We plan to develop a suite of techniques that use such signals, analogous to ones in anti-virus software, to inform consumers of the risk that a given media source or item is false or misleading. For instance, we have developed the FANG (FAke News Graph) system, which contextualizes the social and publishing background of news items [49]. Our component work on QA Check addresses the text modalities of MDM in ways that leverage decomposing claims into atomic, easily verified claims [55, 56]. We have also developed Sniffer, a fact-checking system that combines InstructBLIP refinement and GPT-4 instruction data to situate news items within a broader context [59]. Ultimately, these detection tools will be applied to educate consumers about the authenticity of the information they encounter, thus bridging the gap between technical detection and practical application. These efforts aim to empower consumers, thereby enhancing the overall resilience of the information ecosystem.

5.2 How Consumers Engage with MDM

Our research in TD2 goes beyond algorithm development to include a strong focus on user experience and societal impact. The upcoming experimental setups—ranging from dynamic content recommendation systems to social diffusion models—are designed to test not only technological efficiency but also user engagement and trust.

TD2 focuses on how consumers engage with MDM, relating both to the characteristics of trustworthy news and their association with consumer preferences to drive greater engagement and sharing behavior. The quest to refine the flow of digital information and consequently diminish the impact of echo chambers mandates the development of algorithms to improve the diversity of content consumed. Prior work on mitigating echo chambers has examined comparing recommendation algorithms for their ability to supply a diversity of sources and perspectives [23, 30, 62, 67]. However, some challenges still need to be addressed regarding the audit of

recommendation algorithms, where most measures of feed quality treat each piece of information as a single data point. In reality, news items can be understood as the sum of their parts [48]. Furthermore, a news-focused approach is incongruent with typical consumer behavior, as many consumers obtain their news indirectly [12]. In reality, a minority of consumers are interested in the news [47] and are more likely to get information from their network peers.

Under TD2, our first objective centers on creating sophisticated news recommendation algorithms driven by fine-grained content labels and classification. Our project aims to progress beyond static recommendations to dynamic, context-sensitive systems that learn and adapt to the changing information landscape in real time. In prior work, we have reported on the importance of variety in news sources [23]. We have also examined the need to diversity topical coverage [30]; yet, what remains is to test these observational findings through experiments that can establish causality. For instance, in some treatments, we plan to harness algorithms that pair similar headlines with substantively diverse content, thereby personalizing the degree of content variance. This strategy employs reinforcement learning principles to balance the exploration of diverse content with the exploitation of user engagement data, aiming to disrupt the echo chamber effect while respecting individual preferences. In contrast, we will also run experiments that empower users with greater control over their information diet.

There is also a need to examine news consumption behavior on social media in ways that simulate the networked experience of encountering and consuming news. Our second objective will incorporate graph-based methodologies to model the social consumption of information. Treatments can comprise a consensus approach where the aggregate and average properties of the news items requested in the network determine the likelihood of news items being consumed. Alternatively, the social approach [22] can adopt a diffusion model similar to classical epidemiology, where instead of predicting the probability that a virus will infect a person, it predicts for each user the likelihood that they will consume a news item. Our research design is informed by the content feed audits of Robertson et al. [61], which highlight algorithmic curation's role in shaping public discourse. Our ongoing efforts in this space involve exploring the effect of invitation mechanisms on information diffusion on messaging platforms such as WeChat [83] and inferencing networks and influence using scalable continuous time-diffusion methods, in scenarios where the whole graph is not readily available, such as in the social media space [82].

In summary, the experiments of TD2 will develop new experimental web and mobile applications that interface with news and social media to offer different informational and social experiences in news consumption and measure their effects on individual knowledge and trust. These initiatives are crucial in advancing the development of algorithms and policies that better cater to diverse consumer needs and preferences, ultimately fostering more nuanced and effective strategies to counteract echo chambers and misinformation. By integrating these technological innovations with insights from social psychology, such as influence and conformity theories, and consumer behavior studies, such as decision-making processes and group dynamics, we aim to develop a comprehensive understanding of information dynamics on digital platforms.

5.3 Perception of Authenticity and Trustworthiness

TD3 focuses on the consumption aspect of MDM, with the aim to meld technical innovations with insights into human psychology to create tools that are as intuitive as they are accurate. The research projects herein are motivated to understand how individuals process information that may conflict with their preconceptions, recognizing the potential for cognitive dissonance, which may result in either a rejection of new information or an entrenched polarization of views. Our approach transcends speculative research by actively integrating findings such as the role of cognitive ability in false news appraisal [39] and cross-national studies on MDM behavior [2, 5, 43, 81]. These insights will inform the development of applications that map content contextually and enhance user resilience against MDM.

The first objective entails the design of a knowledge graph to contextualize digital content. In prior work, we have developed tools to empower intuitive news search with knowledge graphs [80], which enrich the search

context and link entities inter and intra-news documents. As part of TD3, the development of multi-hop reasoning algorithms and knowledge graphs will be complemented by empirical research into how these tools affect user trust and information processing. TD3 will build on these findings by designing tools to train fact-checkers and general users how to reason through information curated from various sources (e.g., [55]). We see these efforts as necessary to build digital resilience. User-powered interfaces for reasoning can also play an important role to safeguard fact-checking benchmarks against the ideological defaults and social biases evinced even in credible news sources [73, 74] and in LLM-generated outputs [25, 42, 63].

The second objective is to develop algorithms for multi-hop reasoning that can trace the logical connections across various pieces of information. Such algorithms are instrumental in detecting out-of-context misinformation by evaluating the coherence of narratives and their alignment with facts. This objective will leverage on our prior work exploring context-aware outstanding fact mining from knowledge graphs [79], which helps relate a target and context entity, thereby assisting users in making sense of new information. Sniffer's current capabilities of detecting text-image inconsistencies [59] will be enhanced by these algorithms to include reasoning across multiple data points and providing explanations rooted in external knowledge.

The third objective explores how individual differences influence the perceived accuracy of online information, which previous studies have demonstrated can vary widely across individuals [3]. To better understand perceptions of information authenticity and trustworthiness, iGyro will conduct studies focusing on how trust is developed through technological affordances and perceived through a consumer-focused lens. We plan to build a deeper understanding of trust and MDM resilience mechanisms through large-scale surveys and experiments that reveal how individuals encounter, compare, and contrast information. Based on survey insights, we will run experiments that test the effectiveness of the previously developed information vignettes for different demographic groups to disentangle the various effects of multiple simultaneously operating cues driving sense-making behavior. Next, we will focus on tailoring these technologies to enhance user engagement and trust, assessing their impact across different demographics to ensure inclusiveness and effectiveness.

In summary, TD3's focus on the consumption of MDM aims to integrate technical innovations with insights into human psychology to develop intuitive and accurate tools. This objective drives the creation of systems like knowledge graphs and multi-hop reasoning algorithms that enhance the contextual understanding and evaluation of digital content, thus empowering users in recognizing and avoiding misinformation. By incorporating empirical research on how such tools influence user trust and information processing, we refine our approaches to align with real-world user interactions and cognitive behaviors. Additionally, our studies on individual differences in perception of information accuracy and trustworthiness will guide the development of more personalized and effective user engagement strategies. Through large-scale surveys and targeted experiments, iGyro aims to map out varied consumer responses to information, tailoring technology to build resilience and trust across diverse demographics. This comprehensive approach pushes forward our technical capabilities and enriches our understanding of misinformation's social and psychological dynamics, propelling us toward a more informed and resilient digital society.

6 Sphere 3: Regulation and Policy

In the iGyro project, Sphere 3 focuses on synthesizing the informational and behavioral insights developed through research spheres 1 and 2 into three well-defined objectives: conducting a desk study to map global regulatory approaches, creating a framework for digital information resilience, and executing a responsive regulation study of Singapore and four other Asian cities.

However, a few factors need to be considered when developing GenAI policies. First, regulating digital information needs to keep evolving with the technologies it seeks to govern. Second, clarifying the available objectives, tools, and levers is necessary. "Regulation" includes rules, standards, and less formal forms of supervised self-regulation [6]. Policy interventions are still broader, including educational and social policies intended to build consumer resilience. Third, it is necessary to consider the stakeholder's scope of focus. On the one hand, it

is interesting to compare whether and where government regulations apply to pieces of content, their creators, their sharers, or the platforms where these exchanges occur. On the other hand, the role of platforms is increasingly important in policies around GenAI, as we discuss in recent work [35]. Fourth and finally, the focus should be on balance so that regulators can address the perceived harms of GenAI while not unduly limiting innovation or driving it elsewhere. For the most part, however, the harm is in the information's impact on other users and society. In addition to punishing those who intend harm such as fraud, hate speech, or defamation, much attention has focused on the responsibility of platforms that host and facilitate access. Spreading malicious content is already the subject of regulation in many jurisdictions. Although there is wariness about unnecessary limits on freedom of speech, even in broadly libertarian jurisdictions like the United States, one cannot yell "Fire!" in a crowded theatre. Key questions to resolve include whether the tools to generate content should be regulated. Our societies do not usually restrict private activity—a hateful lie written in a diary is not a crime, for example; nor do we punish word processing software for the threats typed on it. A notable exception is that many jurisdictions make it an offense to create or possess child pornography, including synthetic images in which no actual child was harmed, even if the photos are not shared.

The first objective of Sphere 3 is to conduct a comprehensive survey of the existing regulatory frameworks from a global perspective. Previous findings have underscored the need for balance in regulation, avoiding undue constraints on innovation while mitigating harms associated with GenAI. We will systematically examine the landscape of digital regulation, contrasting varied approaches from the libertarian policies in the United States to the more assertive measures adopted in countries like China. For instance, in the United States, this would require a review of Section 230 of the 1996 Communications Decency Act, which absolves Internet platforms of responsibility for the content posted on them. Australia released a draft bill in 2023 on combatting misinformation and disinformation [52] that has been hotly debated [61]—including its fair share of fake news. Around the same time, the EU's Digital Services Act [15] came into force, while Britain passed a new Online Safety Act [40]. All struggle with the problem of how to deal with "legal but harmful" content online. Australia's bill would have granted its media regulator more power to question platforms on their efforts to combat misinformation. The backlash against GenAI's perceived threats to free speech led the government to postpone its introduction to Parliament until later 2024, with promises to "improve the bill" [68] The EU legislation avoids defining disinformation but limits measures on socially harmful (as opposed to "illegal") [78] content to "very large online platforms" and "very large online search engines"—in essence, big tech companies like Google and Meta. Ofcom [53], the body tasked with enforcing the new UK law, states that it is "not responsible for removing online content" but will help ensure that firms have effective systems to prevent harm. Such gentle measures may be contrasted with China's more robust approach, where over-inclusion often characterizes the "great firewall" [26]. Some years ago, Winnie the Pooh was briefly blocked [70] because of memes comparing him to President Xi Jinping; earlier efforts to limit discussion of the "Jasmine Revolution" unfolding across the Arab world in 2011 led to a real-world impact on online sales of jasmine tea [19].

However, regulations by platforms to curtail MDM apply mainly to users, where correcting or shadowbanning them may be one way they address the problem, as we explored in prior work [31]. Limiting the speed with which false information can be transmitted is another option, analogous to the circuit breakers that protect stock exchanges from high-frequency trading algorithms sending prices spiraling. In India in 2018, WhatsApp began limiting the ability to forward messages [57] after lynch mobs killed several people following rumors circulated on the platform. A study based on data collected from India, Brazil, and Indonesia showed that such methods can delay the spread of information [17] but are ineffective in blocking the propagation of disinformation campaigns in public groups. Another platform-based approach is to be more transparent about the provenance of information. Several now promise to label synthetic content, although the ease of creation makes this a challenging game of catch-up. Tellingly, the U.S. tech companies that agreed to voluntary watermarking [66] last year limited those commitments to images and video, echoed in the Biden Administration's October 2023 exec-

utive order [27]. Synthetic text is nearly impossible to label consistently, and as it becomes easier for GenAI to generate, multimedia, images, and video will likely go the same way.

The second objective of Sphere 3 is to develop a framework for digital information resilience. It motivates users to take responsibility for what they consume and share. Concerns about the public's information diet are as old as democracy itself. Some months before the U.S. Constitution was drafted in 1787, Thomas Jefferson pondered whether it would be better to have a government without newspapers or newspapers without a government [69]. "I should not hesitate a moment to prefer the latter," he concluded, making clear that he meant that all citizens should receive those papers and be capable of reading them. Those who grew up watching curated nightly news or scanning a physical newspaper may be mystified by a generation that learns about current events from social media feeds and the following video on TikTok.

The final objective of Sphere 3 is to take a pan-Asian perspective in exploring digital resilience approaches to GenAI and MDM, focusing on Singapore and four other Asian cities. This responsive regulation study will examine the effectiveness of local and regional regulations, drawing from a variety of cultural and legal contexts to enhance our understanding of digital governance. Using the POFMA (Protection from Online Falsehoods and Manipulation Act) in Singapore, for example, offers valuable lessons on the balance between regulatory enforcement and maintaining public trust in digital spaces. Although Singapore was criticized [76] when it adopted POFMA in 2019 [32], governments around the world are considering similar legislation to deal with the problem of fake news [9, 24]. Therefore, drawing on the enactment and application of POFMA and other Asian cities identified through the first objective, this study will yield insights into the relationship between regulation, technological adoption, and societal norms, contributing to a more nuanced understanding of digital governance. The value of the approach adopted in the iGyro project is that such regulatory debates are connected to the technology that provides the medium for creation and dissemination and to the underlying consumer behavior that drives and feeds it.

In summary, through the synthesis of evidence-based insights gained in Spheres 1 and 2—ranging from behavioral analysis, such as understanding consumer decision-making processes and cognitive biases, to technological assessments, such as evaluating the effectiveness of multi-hop reasoning algorithms—Sphere 3 will propose tailored, actionable strategies that account for each region's unique societal and technological landscapes. A principles-based approach that emphasizes the human dynamic, such as focusing on trust-building and enhancing digital literacy, may also address a key challenge in regulation: delays in adoption relegate policymakers to fighting battles in which the technology itself has become obsolete by the time of adoption. This comprehensive approach will inform better regulatory frameworks and support the development of more resilient digital environments that can adapt to the rapid evolution of misinformation threats and technological advancements.

7 Conclusion and Outlook

In a landscape where GenAI is reshaping the creation and the very fabric of digital information exchange, the challenge for governments and regulatory bodies is to identify the aspects of GenAI that require oversight and implement adaptive regulations that stay relevant amidst rapid technological advancements. Dynamic policies, supported by a robust monitoring system for GenAI innovations and initiatives to enhance public understanding of AI outputs, are crucial for engaging critically with AI-generated content. The iGyro initiative exemplifies a systemic, interdisciplinary research approach that fosters an in-depth understanding of the interplay between technological advancements and societal impacts, thereby addressing digital misinformation challenges effectively.

The future of *trustworthy* GenAI demands a better understanding of its consumers; therefore, we are looking to diversify the disciplines represented within the iGyro project, specifically exploring the psychological implications of synthetic media on individuals and societies. The iGyro project's pan-South Asia goal also involves developing international collaborations that include policy and technological discussions, and scholarship and cultural exchange to understand and devise solutions tailored to different societal norms and values. Additionally,

we will plan to establish partnerships with educational institutions to research and potentially integrate curriculum on media literacy that reflects the latest findings from GenAI's role in combating misinformation.

We portend that the journey to trustworthy GenAI will be challenging and tumultuous. The current focus on copyright infringements around GenAI may overlook broader societal impacts, such as how GenAI alters perception and interaction in our digital world. With the proliferation of synthetic media, methods like labeling human-created content and watermarking images to help users identify their origins are becoming more feasible, although user engagement in verifying such information remains low. For instance, "read before you share" prompts, aimed at encouraging more thoughtful sharing of news, have improved conscious news sharing but may not have reversed trends toward platform toxicity.

As we envision the evolving landscape of digital information, the role of GenAI can extend beyond current applications. For instance, GenAI systems can actively participate in the remediation and correction of MDM in real time, through informational, infrastructural, or algorithmic interventions. Exploring the development of autonomous AI agents capable of detecting and interacting within digital ecosystems to restore informational integrity is also a priority. The impact of GenAI on societal structures warrants deeper investigation, particularly how synthetic media influences group dynamics and political discourse, and shapes collective behaviors and societal norms. Additionally, as GenAI technologies permeate global markets, there is a compelling need to establish cohesive standards and regulatory frameworks that address ethical, privacy, and transparency issues. These efforts should aim to harmonize approaches to GenAI governance, ensuring equitable and ethical utilization across different cultural and political landscapes.

These initiatives represent just a fraction of the potential pathways for leveraging GenAI to strengthen digital ecosystems against MDM. By pushing the boundaries of current research and exploring these new domains, we can harness GenAI's capabilities more effectively, ensuring that advancements in AI contribute positively to societal resilience and the integrity of our information environments.

Acknowledgement

The authors would like to thank Ms. Nur Insyirah Binte Imam Mujtahid for her help in the literature review, copyediting, formatting, and quality assurance for this work.

References

[1] Rachel Adams, Ayantola Alayande, Zameer Brey, Brantley Browning, Michael Gastrow, Jerry John Kponyo, Dona Mathew, Moremi Nkosi, Henry Nunoo-Mensah, Diana Nyakundi, Victor Odumuyiwa, Olubunmi Okunowo, Philipp Olbrich, Nawal Omar, Kemi Omotubora, Paul Plantinga, Gabriella Razzano, Zara Schroeder, Andrew Selasi Agbemenu, Araba Sey, Kristophina Shilongo, Shreya Shirude, Matthew Smith, Eric Tutu Tchao, and Davy K. Uwizera. 2023. A new research agenda for African generative AI. *Nature Human Behaviour* 7, 11 (2023), 1839–1841.

[2] Saifuddin Ahmed. 2023. Examining public perception and cognitive biases in the presumed influence of deepfakes threat: Empirical evidence of third person perception from three studies. *Asian Journal of Communication* 33, 3 (2023), 308–331.

[3] Saifuddin Ahmed and Han Wei Tan. 2022. Personality and perspicacity: Role of personality traits and cognitive ability in political misinformation discernment and sharing behavior. *Personality and Individual Differences* 196 (2022), 111747.

[4] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The welfare effects of social media. *American Economic Review* 110, 3 (2020), 629–676.

[5] Antonio A. Arechar, Jennifer Allen, Adam J. Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G. Lu, Robert M. Ross, Michael N. Stagnaro, Yunhao Zhang, Gordon Pennycook, and David G. Rand. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour* 7, 9 (2023), 1502–1513.

[6] Robert Baldwin, Martin Cave, and Martin Lodge. 2011. *Understanding Regulation 2E P: Theory, Strategy, and Practice*. Oxford University Press.

[7] Gordon M. Becker, Morris H. DeGroot, and Jacob Marschak. 1964. Measuring utility by a single-response sequential method. *Behavioral Science* 9, 3 (1964), 226–232.

[8] Claudi L. Bockting, Eva A. M. van Dis, Robert van Rooij, Willem Zuidema, and Johan Bollen. 2023. Living guidelines for generative AI—Why scientists must oversee its use. *Nature* 622, 7984 (2023), 693–696.

[9] Lee C. Bollinger and Geoffrey R. Stone. 2022. *Social Media, Freedom of Speech, and the Future of our Democracy*. Oxford University Press.

[10] Antonio Carnevale, Claudia Falchi Delgado, and Piercosma Bisconti. 2023. Hybrid ethics for generative AI: Some philosophical inquiries on GANs. *HUMANA.MENTE Journal of Philosophical Studies* 16, 44 (2023), 33–56.
[11] Sanjay Chawla, Preslav Nakov, Ahmed Ali, Wendy Hall, Issa Khalil, Xiaosong Ma, Husrev Taha Sencar, Ingmar Weber, Michael Wooldridge, and Ting Yu. 2023. Ten years after ImageNet: A 360° perspective on artificial intelligence. *Royal Society Open Science* 10, 3 (2023), 221414.
[12] Zhe Chen, Aixin Sun, and Xiaokui Xiao. 2021. Incremental community detection on large complex attributed network. *ACM Transactions on Knowledge Discovery from Data* 15, 6 (2021), 1–20.
[13] Simon Chesterman. 2024. Good models borrow, great models steal: intellectual property rights and generative AI. *Policy and Society* (2024), puae006.
[14] Kevin Chung, Keehyung Kim, and Noah Lim. 2020. Social structures and reputation in expert review systems. *Management Science* 66, 7 (2020), 3249–3276.
[15] European Commission. n.d. The Digital Services Act Package. Retrieved August 28, 2024 from https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package
[16] Gregory Currie. 2007. Both sides of the story: Explaining events in a narrative. *Philosophical Studies* 135 (2007), 49–63.
[17] Philipe de Freitas Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro O. S. Vaz de Melo, and Fabrício Benevenuto. 2020. Can WhatsApp counter misinformation by limiting message forwarding? In *Complex Networks and Their Applications VIII (COMPLEX NETWORKS 2019)*. Studies in Computational Intelligence, Vol. 881. Springer, 372–384.
[18] Cristina Godoy B. de Oliveira, Fabio G. Cozman, and João Paulo C. Veiga. 2023. This hot AI summer will impact Brazil's democracy. *Nature Human Behaviour* 7, 11 (2023), 1842–1844.
[19] Bruce J. Dickson. 2011. No "jasmine" for China. *China and East Asia* 110, 737 (2011), 211–216. https://www.jstor.org/stable/45319730
[20] Benjamin Enke and Florian Zimmermann. 2019. Correlation neglect in belief formation. *Review of Economic Studies* 86, 1 (2019), 313–332.
[21] Erik Eyster and Georg Weizsacker. 2010. *Correlation Neglect in Financial Decision-Making*. Discussion Papers of DIW Berlin 1104. DIW Berlin, German Institute for Economic Research.
[22] Leon Festinger. 1962. Cognitive dissonance. *Scientific American* 207, 4 (1962), 93.
[23] Sean Fischer, Kokil Jaidka, and Yphtach Lelkes. 2020. Auditing local news presence on Google News. *Nature Human Behaviour* 4, 12 (2020), 1236–1244.
[24] Serena Giusti and Elisa Piras. 2020. *Democracy and Fake News: Information Manipulation and Post-Truth Politics*. Routledge.
[25] Lucas Gover. 2023. Political bias in large language models. *Commons: Puget Sound Journal of Politics* 4, 1 (2023), 2.
[26] James Griffiths. 2021. *The Great Firewall of China: How to Build and Control an Alternative Version of the Internet*. Bloomsbury Publishing.
[27] The White House. 2023. FACT SHEET: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence. *The White House*. Retrieved August 28, 2024 from https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/
[28] Matthew Hutson. 2023. Rules to keep AI in check: Nations carve different paths for tech regulation. *Nature* 620, 7973 (2023), 260–263.
[29] Radosveta Ivanova-Stenzel and Gyula Seres. 2021. Are strategies anchored? *European Economic Review* 135 (2021), 103725.
[30] Kokil Jaidka, Sean Fischer, Yphtach Lelkes, and Yifei Wang. 2023. News nationalization in a digital age: An examination of how local protests are covered and curated online. *Annals of the American Academy of Political and Social Science* 707, 1 (2023), 189–207.
[31] Kokil Jaidka, Subhayan Mukerjee, and Yphtach Lelkes. 2023. Silenced on social media: The gatekeeping functions of shadowbans in the American Twitterverse. *Journal of Communication* 73, 2 (2023), 163–178.
[32] Shashi Jayakumar, Benjamin Ang, and Nur Diyanah Anwar. 2021. Fake news and disinformation: Singapore perspectives. In *Disinformation and Fake News*. Springer, 137–158.
[33] Nan Jiang, Bangjie Sun, Terence Sim, and Jun Han. 2024. Can I hear your face? Pervasive attack on voice authentication systems with a single face image. In *Proceedings of the 33rd USENIX Security Symposium*.
[34] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications* 80 (2021), 11765–11788. https://api.semanticscholar.org/CorpusID:230800534
[35] Shaleen Khanal, Hongzhou Zhang, and Araz Taeihagh. 2024. Why and how is the power of Big Tech increasing in the policy process? The case of generative AI. *Policy and Society* 2024 (2024), puae012.
[36] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. 2020. Deepfakes: Trick or treat? *Business Horizons* 63, 2 (2020), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006
[37] Samuli Laato, A. K. M. Najmul Islam, Muhammad Nazrul Islam, and Eoin Whelan. 2020. What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems* 29, 3 (2020), 288–305.
[38] Laks V. S. Lakshmanan, Michael Simpson, and Saravanan Thirumuruganathan. 2019. Combating fake news: A data management and mining perspective. *Proceedings of the VLDB Endowment* 12 (2019), 1990–1993. https://api.semanticscholar.org/CorpusID:201653243
[39] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[40] Legislation.gov.uk. 2023. Online Safety aAct 2023. Retrieved August 28, 2024 from https://www.legislation.gov.uk/ukpga/2023/50/contents/enacted
[41] Jia Li, Noah Lim, and Hua Chen. 2020. Examining salesperson effort allocation in teams: A randomized field experiment. *Marketing Science* 39, 6 (2020), 1122–1141.
[42] Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *Proceedings of the 12th International Conference on Learning Representations*.
[43] Dani Madrid-Morales, Herman Wasserman, Gregory Gondwe, Khulekani Ndlovu, Etse Sikanku, Melissa Tully, Emeka Umejei, and Chikezie Uzuegbunam. 2021. Motivations for sharing misinformation: A comparative study in six Sub-Saharan African countries. *International Journal of Communication* 15, 2021 (2021), 1200–1219.
[44] Aqdas Malik, Amandeep Dhir, Puneet Kaur, and Aditya Johri. 2020. Correlates of social media fatigue and academic performance decrement: A large cross-sectional study. *Information Technology & People* 34, 2 (2020), 557–580.
[45] Helen Margetts and Cosmina Dorobantu. 2019. Rethink government with AI. *Nature* 568, 7751 (2019), 163–165.
[46] Marissa Mock, Suzanne Edavettal, Christopher Langmead, and Alan Russell. 2023. AI can help to speed up drug discovery—But only if we give it the right data. *Nature* 621, 7979 (2023), 467–470.
[47] Subhayan Mukerjee, Kokil Jaidka, and Yphtach Lelkes. 2022. The political landscape of the US Twitterverse. *Political Communication* 39, 5 (2022), 565–588.
[48] Subhayan Mukerjee and Tian Yang. 2021. Choosing to avoid? A conjoint experimental study to understand selective exposure and avoidance on social media. *Political Communication* 38, 3 (2021), 222–240.
[49] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. https://api.semanticscholar.org/CorpusID:221150991
[50] Linda Nordling. 2019. A fairer way forward for AI in health care. *Nature* 573, 7775 (2019), S103–S103.
[51] Charles N. Noussair and Gyula Seres. 2020. The effect of collusion on efficiency in experimental auctions. *Games and Economic Behavior* 119 (2020), 267–287.
[52] Parliament of the Commonwealth of Australia. 2023. Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023. Retrieved August 28, 2024 from https://www.infrastructure.gov.au/sites/default/files/documents/communications-legislation-amendment-combatting-misinformation-and-disinformation-bill2023-june2023.pdf
[53] Ofcom. 2023. Online Safety—What Is Ofcom's Role, and What Does It Mean for You? Retrieved August 28, 2024 from https://www.ofcom.org.uk/news-centre/2023/online-safety-ofcom-role-and-what-it-means-for-you
[54] Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *Proceedings of the International Workshop on the Semantic Web*. https://api.semanticscholar.org/CorpusID:52900831
[55] Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. 2023. QACheck: A demonstration system for question-guided multi-hop fact-checking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 264–273. https://doi.org/10.18653/v1/2023.emnlp-demo.23
[56] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6981–7004. https://doi.org/10.18653/v1/2023.acl-long.386
[57] Sankalp Phartiyal and Krishna V. Kurup. 2018. WhatsApp curbs message forwarding in bid to deter India lynch mobs. *Reuters*. Retrieved August 28, 2024 from https://www.reuters.com/article/idUSKBN1KB026/
[58] Francesco Pierri, Carlo Piccardi, and Stefano Ceri. 2019. Topology comparison of Twitter diffusion networks effectively reveals misleading information. *Scientific Reports* 10 (2019), 1372. https://api.semanticscholar.org/CorpusID:147703897
[59] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the 2024 Conference on Computer Vision and Pattern Recognition*.
[60] Alex Rees-Jones and Dmitry Taubinsky. 2020. Measuring "schmeduling." *Review of Economic Studies* 87, 5 (2020), 2399–2438.
[61] Amy Remeikis. 2023. Why is Labor's bill on combatting disinformation so controversial? *The Guardian* Retrieved August 28, 2024 from https://www.theguardian.com/australia-news/2023/oct/01/why-is-labors-bill-on-combatting-disinformation-so-controversial
[62] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within Google search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.
[63] David Rozado. 2023. The political biases of ChatGPT. *Social Sciences* 12, 3 (2023), 148.
[64] Sanjay Saha, Rashindrie Perera, Sachith Seneviratne, Tamasha Malepathirana, Sanka Rasnayaka, Deshani Geethika, Terence Sim, and Saman Halgamuge. 2023. Undercover deepfakes: Detecting fake segments in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 415–425.
[65] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. On the conversational persuasiveness of large language models: A randomized controlled trial. *arXiv preprint arXiv:2403.14380* (2024).

[66] Sabrina Siddiqui and Deepa Seetharaman. 2023. White House says Amazon, Google, Meta, Microsoft agree to AI safeguards. *The Wall Street Journal*. Retrieved August 28, 2024 from https://www.wsj.com/articles/white-house-says-amazon-google-meta-microsoft-agree-to-ai-safeguards-eabe3680

[67] Kazunari Sugiyama and Min-Yen Kan. 2015. "Towards higher relevance and serendipity in scholarly paper recommendation" by Kazunari Sugiyama and Min-Yen Kan with Martin Vesely as coordinator. *ACM SIGWEB Newsletter* 2015, Winter (Feb. 2015), Article 4, 16 pages. https://doi.org/10.1145/2719943.2719947

[68] Josh Taylor. 2023. Labor to overhaul misinformation bill after objections over freedom of speech. *The Guardian*. Retrieved August 28, 2024 from https://www.theguardian.com/australia-news/2023/nov/13/labor-misinformation-bill-objections-freedom-of-speech-religious-freedom

[69] Jefferson Thomas. 1787. Jefferson's preference for "newspapers without government" over "government without newspapers." *Online Library of Liberty*. Retrieved August 28, 2024 from https://oll.libertyfund.org/quote/jefferson-s-preference-for-newspapers-without-government-over-government-without-newspapers-1787

[70] The Straits Times. 2017. 'Oh, bother': Chinese censors can't bear Winnie the Pooh. *The Straits Times*. Retrieved August 28, 2024 from https://www.straitstimes.com/asia/east-asia/oh-bother-chinese-censors-cant-bear-winnie-the-pooh

[71] Augustin Toma, Senthujan Senkaiahliyan, Patrick R. Lawler, Barry Rubin, and Bo Wang. 2023. Generative AI could revolutionize health care—But not if control is ceded to big tech. *Nature* 624, 7990 (2023), 36–38.

[72] Chris Tyler, K. L. Akerlof, Alessandro Allegra, Zachary Arnold, Henriette Canino, Marius A. Doornenbal, Josh A. Goldstein, David Budtz Pedersen, and William J. Sutherland. 2023. AI tools as science policy advisers? The potential and the pitfalls. *Nature* 622, 7981 (2023), 27–30.

[73] Francielle Vargas, Kokil Jaidka, Thiago A. S. Pardo, and Fabrício Benevenuto. 2023. Predicting sentence-level factuality of news and bias of media outlets. *arXiv preprint arXiv:2301.11850* (2023).

[74] Preetika Verma, Hansin Ahuja, and Kokil Jaidka. 2023. Quantify the political bias in news edits: Experiments with few-shot learners (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 16354–16355.

[75] Christian von der Weth, Ashraf Abdul, Shaojing Fan, and Mohan Kankanhalli. 2020. Helping users tackle algorithmic threats on social media: A multimedia research agenda. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*. ACM, New York, NY, USA, 4425–4434. https://doi.org/10.1145/3394171.3414692

[76] Human Rights Watch. 2021. Singapore: 'Fake news' law curtails speech. *Human Rights Watch*. Retrieved August 28, 2024 from https://www.hrw.org/news/2021/01/13/singapore-fake-news-law-curtails-speech

[77] Nick Wingfield, Mike Isaac, and Katie Benner. 2016. Google and Facebook take aim at fake news sites. *The New York Times*. Retrieved August 28, 2024 from https://www.nytimes.com/2016/11/15/technology/google-will-ban-websites-that-host-fake-news-from-using-its-ad-service.html

[78] Konarski Xawery. 2023. The Digital Services Act (DSA) and Combating Disinformation—10 Key Takeaways. Retrieved August 28, 2024 from https://www.traple.pl/en/the-digital-services-act-dsa-and-combating-disinformation-10-key-takeaways/

[79] Yueji Yang, Yuchen Li, Panagiotis Karras, and Anthony K. H. Tung. 2021. Context-aware outstanding fact mining from knowledge graphs. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2006–2016.

[80] Yueji Yang, Yuchen Li, and Anthony K. H. Tung. 2021. NewsLink: Empowering intuitive news search with knowledge graphs. In *Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE'21)*. IEEE, 876–887.

[81] Jing Zeng and Chung-Hong Chan. 2021. A cross-national diagnosis of infodemics: Comparing the topical and temporal features of misinformation around COVID-19 in China, India, the US, Germany and France. *Online Information Review* 45, 4 (2021), 709–728.

[82] Bowen Zhang and Terence Sim. 2022. Localizing fake segments in speech. In *Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR'22)*. IEEE, 3224–3230.

[83] Shiqi Zhang, Jiachen Sun, Wenqing Lin, Xiaokui Xiao, Yiqian Huang, and Bo Tang. 2024. Information diffusion meets invitation mechanism. In *Companion Proceedings of the ACM Web Conference 2024 (WWW'24)*. 383–392.

Received 2 February 2024; revised 15 June 2024; accepted 7 August 2024