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

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

Generative artificial intelligence (GenAI) is exacerbating the challenges of Misinformation, Disinformation, and Mal-information (MDM). The quantity and quality of synthetic content requires reconsidering how information is created, disseminated, and consumed. That exploration is crucial for understanding how MDM can impact trust in public institutions and resilience among consumers. We propose a three-tiered interdisciplinary approach to characterize how consumers engage with and perceive GenAI. Recognizing the consumer behavior that shapes MDM consumption, addressing vulnerabilities in the information pipeline, and developing policies that are fit for purpose is essential to safeguarding the integrity of information and maintaining public trust in a digital age.

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

Authors' addresses: Kokil Jaidka, jaidka@nus.edu.sg, Department of Communications and New Media, National University of Singapore; Tsuhan Chen, tsuhan@nus.edu.sg, School of Computing, National University of Singapore; Simon Chesterman, chesterman@nus.edu.sg, Faculty of Law, National University of Singapore; Wynne Hsu, whsu@comp.nus.edu.sg, School of Computing, National University of Singapore; Min-Yen Kan, kanmy@comp.nus.edu.sg, School of Computing, National University of Singapore; Mohan Kankanhalli, mohan@comp.nus.edu.sg, School of Computing, National University of Singapore; Mong Li Lee, leeml@comp.nus.edu.sg, School of Computing, National University of Singapore; Gyula Seres, gyula@nus.edu.sg, Business School, National University of Singapore; Terence Sim, terence.sim@nus.edu.sg, School of Computing, National University of Singapore; Araz Taeihagh, spparaz@nus.edu.sg, Lee Kuan Yew School of Public Policy, National University of Singapore; Anthony Tung, atung@comp.nus.edu.sg, School of Computing, National University of Singapore; Xiaokui Xiao, xkxiao@comp.nus.edu.sg, School of Computing, National University of Singapore; Audrey Yue, audrey.yue@nus.edu.sg, Department of Communications and New Media, National University of Singapore.

ACM 2639-0175/2024/xx-ARTxx
https://doi.org/XXXXXXX.XXXXXXX

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. 2024. Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy. *Digit. Gov. Res. Pract.* xx, xx, Article xx (xx 2024), 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The digital age — characterized by the internet, social media platforms, and the proliferation of mobile devices like smartphones — has transformed how people acquire and share information. Consumers expect new content to be delivered at the swipe of a finger; opinions are formed, and decisions are based solely on algorithms' content and consumed individually. Where the 'news' was once curated by experts, it is now personalized to suit one's interests (and biases). The insatiable demand for content drives an 'economy of digital consumerism,' where supply is usually unregulated. Around the world, this has led to the paradox of digital information: ever more people have access to more information than at any point in human history — yet their trust in the veracity of that information is in decline. The lines between journalism, advertising, and entertainment are increasingly blurred, and motivations behind creating a particular piece of information and the mechanisms that determine why it is served to a consumer are often not immediately evident. The production, consumption, and subsequent dissemination of information with questionable credibility have spurred three major informational crises that are the focus of research and governance efforts worldwide — Misinformation, Disinformation, and Mal-information (MDM). Misinformation refers to misleading information without malicious intent; disinformation uses information deceptively to push an agenda or a false narrative; mal-information aims to inflict societal harm. In reaction to the consequences and implications of MDM, many governments are establishing new laws, as well as adapting established processes, procedures, and provisions to tackle these issues; yet, comprehensive and practical legislation is years away [10, 28].

In this context, Generative Artificial Intelligence (GenAI) now threatens to amplify the portended risks of MDM because of its availability, ease of use, and remarkable sophistication in creating new forms of MDM. GenAI are models learned from data that are capable of creating synthetic multimedia content that simulates the characteristics and sensibilities of content featuring or created by humans [12]. Thousands of user-developed free software and web applications now allow individuals to generate high-quality synthetic portraits and videos, also known as deepfakes [33], that feature politicians, celebrities, and regular citizens saying and doing acts 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. First, GenAI can create high-quality, compelling fake information that is difficult to trace back to a source or creator. Second, GenAI is lowering the threshold for creating and sharing MDM and increasing the difficulty in distinguishing it from authentic sources. Thirdly, the illusory truth effect of GenAI implies more significant media skepticism even towards credible sources [3], thereby sowing distrust and division and undermining the bonds that knit societies together.

Currently, the detection and authentication of MDM is tackled primarily from a computing perspectives, with the onus placed on developers to police and secure their systems. However, while future advances may yield technical improvements, how humans confront and consume new information is expected to remain unchanged. In the face of the proliferation of GenAI tools, there is a need to apply a holistic approach that considers not simply creation or detection but also consumption so that AI governance can effectively harness GenAI and mitigate the risks it poses to stable societies [19]. Accordingly, we offer a consumer behavior perspective 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. We propose that computational approaches should be complemented by understanding consumer motivations, decisions, and responses related to their interactions with digital information. Finally, we propose that these insights should inform regulatory and governance structures on various aspects of the digital information pipeline while recognizing the profound effect of trust in these institutions on public behavior and adherence to guidelines. Trust remains a fundamental component in instilling digital resilience and combating the challenges MDM poses in the digital era.

2 THE LANDSCAPE OF GENERATIVE AI AND INFORMATION INTEGRITY

The arrival of GenAI, with large language models such as ChatGPT and text-to-image generators such as Stable Diffusion, has radically altered content production. The primary advantage of GenAI for a general user is creating engaging and relevant content through simple requests, which now requires minimal effort. A second advantage is synthesizing various sources into a coherent summary or translation. While search engines lowered the barriers to entry for public access to information, GenAI enables a broader audience to create and understand online information, which was previously limited by the need for specialized skills or resources to produce professional-grade media.

Besides the advantages of GenAI to users, GenAI also offers many potential benefits in evidence-based health and medicine [45, 68], policy and public service [9, 42, 69], agriculture and education [1]. However, relying on GenAI content can be problematic, as its answers are based on training data that is often not disclosed or non-representative [50], and can include large amounts of unverified or unverifiable information (The extent to which this source information is protected by copyright is the subject of ongoing litigation in various jurisdictions.) [14]. If GenAI regurgitates such material, it may synthesize these perspectives into inaccurate content bearing a veneer of credibility that the sources lack. Such innocent 'hallucinations' or 'confabulations' are a known feature of the technology, typically including a disclaimer that it should not be relied upon for factual content.

More troublingly, while a large proportion of the internet is indeed truthful and trustworthy, MDM has become rampant across all media platforms. When consumers believe or are influenced by MDM, their subsequent decisions and actions play into the hands of those intent on causing division, promoting alternative agendas, or misleading individuals for personal gain.

For consumers, the risk lies in distinguishing between AI-generated falsehoods and authentic information, underscoring the need for improved digital literacy. GenAI's capacity to tailor content to individual biases increases the likelihood of encountering and engaging with MDM, which calls for critical thinking skills to be a central focus of educational efforts. GenAI exploits dissemination channels, particularly social media, to amplify MDM's reach, exploiting algorithmic biases toward high-engagement content. This issue demands transparent and accountable recommender systems prioritizing factual accuracy over sensationalism.

Addressing the GenAI challenge requires a holistic approach that spans the information life cycle, 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

Given the pervasiveness of GenAI and consequently MDM, it is increasingly necessary to create research approaches that translate across contexts, languages, and cultures while encompassing the digital paradigms of information creation, consumption, and dissemination. Consequently, the role of scientists as the custodians of GenAI is increasingly critical so that policies to improve AI safety can be grounded in evidence-based analyses of technology and user behavior. These considerations have spurred the creation of the iGyro project — a consortium of marketing scientists, computer scientists, social scientists, lawyers, and policymakers all examining the multi-faceted paradigms of MDM within GenAI.

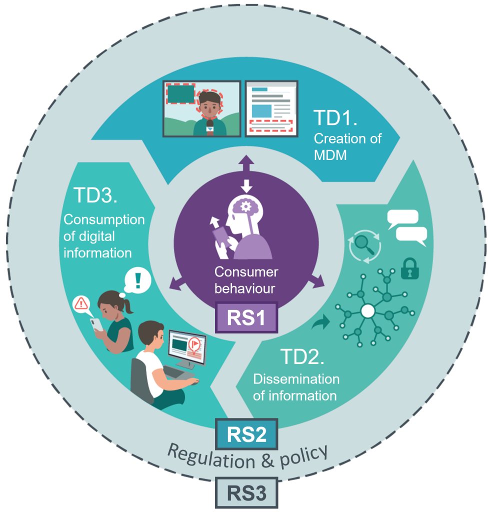

Fig. 1. Conceptual framework of iGyro, showing the three Research Spheres of Consumer Behavior, Digital Information Lifecycle, and Regulation and Policy.

The Information Gyroscope (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 (RS), 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. At its core is Sphere 1, which focuses on investigating the motivations and decision-making processes that govern how consumers seek, process, and share information. Behavioral economics methods such as Bayes' Theorem and decision-theoretic modeling are employed to gain insights into consumer behavior. Understanding the reasons behind vulnerabilities in the digital information landscape helps inform research in RS.

Sphere 2 is divided into three Technology Domains (TD), each representing different stages of the digital information pipeline. TD 1 investigates the creation of MDM in text and visual media, focusing on vulnerabilities that allow false information to appear legitimate. TD 2 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. TD 3 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 life cycle.

4 SPHERE 1: UNDERSTANDING MDM CONSUMPTION

It is imperative to understand the antecedents and consequences of vulnerabilities in the digital information pipeline to achieve long-term digital information resilience at the individual and societal levels. Using the lens of consumer behavior facilitates this, with factors like beliefs and biases influencing the motivations and decisions driving digital media consumption. Understanding MDM consumption should, therefore, examine (1) how consumer beliefs are shaped by the type and veracity of 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.

First, we consider several factors that influence information consumption. For instance, do people choose information from multiple independent sources, or do they focus their search on confirming prior sources of information? Our findings will advance our understanding of bias and the creation of echo chambers in the digital information sphere.

Our approach thrives on understanding the motivations that drive consumer decision-making on information sharing. The extant literature offers limited insights into how consumers account for the fact that the information they share might be false and that others may pass on that wrong information. Studies addressing similar problem statements [34, 41] use observational data to identify exogenous factors such as online trust and social media fatigue. Our study examines information sharing using controlled experiments with elicited beliefs about the veracity of information.

We hypothesize that individual beliefs drive the sharing of misinformation. When considering information consumption in the digital era, three factors violate the traditional assumption that information draws are independent and identically distributed (i.i.d.). First, they are correlated: A piece of information a person sees may be a modified version of a previous draw. Second, false information may drive beliefs away from the actual state of the world. Third, information draws may be duplicated. In this case, identical information from a different source reaches the consumers, but this is not salient to them. We argue that these features induce incorrect beliefs even if these people are rational, i.e., their beliefs follow the statistical principle of conditional expectations [17].

Therefore, we need a framework that considers information access in violation of the i.i.d. assumption. We aim to learn how beliefs on the veracity of information are affected by the relationships of informational draws — whether they are independent if the sources are related (correlated or duplicates), or fake. We consider whether people need to adjust or under-adjust when informational sources are independent. Prior research examined how correlated and fake information results in incorrect belief formation using laboratory experiments [21, 22] and observational [59]. 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.

To model the interplay of these factors, iGyro incorporates recent developments in behavioral economics that investigate how people update their beliefs in ways that are not rational in the textbook sense. Behavioral economics methods of bounded rationality can assess the motivations and decisions determining information consumption. For instance, computer scientists model how people (and machines) update beliefs using Bayes' Theorem, where the reasoning behind how people ascertain the veracity of a claim is tested using informational searches that consist of a series of informational "draws" about whether the claim is true or false.

5 SPHERE 2: UNDERSTANDING MDM CREATION

MDM and its contrast with authentic and legitimate forms of knowledge are the focus of Sphere 2, where they interact with behavioral studies on consumption (Sphere 1) and regulatory and policy interventions (Sphere 3). This Sphere's structure follows the lifecycle of MDM: its creation, dissemination, and consumption.

5.1 Detecting MDM

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 [35, 54, 70]. These technologies examine telltale signs of MDM on these specific dimensions, often relying on sophisticated deep learning models to increase efficacy [31, 49, 58, 72].

Yet, as generation technologies diversify, iGyro must pursue corresponding aspects in detection. Modern MDM generation is hybridising, 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. Synthetic speech for key words can replace an original speech signal. Claims that rely on multiple component facts can turn MDM by falsifying only one part. Critical parts of natural images can be replaced with parts from others "pasted in."

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; while for videos, we exploit temporal information, consistency, and constraints and develop image forensics and machine learning-based methods to tackle full-body fakes.

These will feed into integrative technologies for practical use [49], perhaps as a Veracity Meter, analogous to ones in anti-virus software, to inform consumers of the risk that a given media source or item is false, or misleading. Finally, iGyro's MDM detection efforts blend these aspects with analyses of the MDM lifecycle: how consumers engage with and form perceptions of MDM, discussed next.

5.2 How consumers engage with MDM

Navigating the complexities of digital information flow presents a unique set of challenges, especially when addressing the nuanced balance between information accessibility and the potential for polarizing echo chambers. Central to this issue is the role of search engines and social networks, which, driven by sophisticated algorithms, often inadvertently perpetuate a cycle of repetitive content delivery based on user history and biased search inputs. Herein lies the need to thoroughly comprehend and scrutinize these algorithms, aiming to offer a more equitable news landscape and counteract the pervasive influence of echo chambers.

Prior work on mitigating echo chambers has examined comparing recommendation algorithms for their ability to supply a diversity of sources and perspectives [24, 29, 61, 64]. 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 [47]. Furthermore, a news-focused approach is incongruent with typical consumer behavior, as many consumers obtain their news indirectly [13]. In reality, a minority of consumers are interested in the news [46] and are more likely to get information from their network peers.

We propose research designs focusing on news content and social network peer interactions. First, to better understand and analyze the content of news items, we plan to explore a diversity of algorithms that suggest or retrieve articles with similar headlines but diverse content. The degree of dissimilarity may be calibrated for each user through an exploration–exploitation approach often used in reinforcement learning. Next, we will explore the dynamics of online social networks, particularly in understanding how news spreads within and across different echo chambers. Using social network analysis, we will identify, understand, and predict the impact of these clusters on the propagation of digital information. This will enable the designing and implementation of interventions and policy recommendations to lessen the potentially harmful consequences of echo chambers. Alternative graph-based methods can predict how information items diffuse among users, even without complete network data. For example, a consensus approach may be adopted 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 [23] adopts 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 probability that he or she will consume a news item. Both approaches will be pursued in this project to explore the dissemination of information on social media and messaging platforms.

5.3 Perception of authenticity and trustworthiness

When individuals are exposed to information that challenges their assumptions, there is a danger of triggering cognitive dissonance, wherein individuals cannot reconcile the new information with their existing beliefs [23]. The consequences of cognitive dissonance can be a detachment from further information consumption or sharing [63], or even a backlash effect which further isolates and polarizes an individual against new perspectives [6]. Therefore, while new algorithms will be imperative to detect and arrest MDM, there remains a need to understand better how consumers develop perceptions of authenticity and trustworthiness, anticipate the effectiveness of digital literacy interventions, and engineer digital resilience technologies for the future.

Much of prior work on news trustworthiness has examined the role of surface and content cues [44, 47]. However, a sociotechnical focus still needs to be added on understanding news trustworthiness, with few studies, if any, exploring the role of platforms and interface cues in determining perceptions of trust [2]. Furthermore, the literature on trust perception is also disconnected from prior work that has reported individual differences in the perceived accuracy of online claims [4]. We anticipate that consumers may be influenced in their decision whether to trust a piece of news by various social factors, such as the environment they grew up in and their latent predispositions, which could implicate that they are more knowledgeable or less biased in some areas than others. Furthermore, depending on the style of heuristic processing activated and the duration of news exposure, consumers may trust differently for the same piece of content. Much of prior work has focused on cognitive ability as an antecedent of false news appraisal [36], and some studies offer a cross-national exploration of MDM behavior [3, 5, 40, 74]. However, only some studies have considered these factors in interplay with the content characteristics of MDM. We also plan to extend prior work beyond deepfakes to semantic information and beyond news to understand the antecedents of sense-making for a broader range of topics in health, local politics, world politics, science, and social affairs so that we can test the generalizability of our methods beyond individual events to the broader information contexts.

In order to better understand perceptions of information authenticity and trustworthiness, the iGyro project will conduct studies focusing on how trust is developed through technological affordances and perceived through a consumer-focused lens. We plan to develop a deeper understanding of trust and MDM resilience mechanisms through large-scale surveys and experiments that would 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.

6 SPHERE 3: REGULATION AND POLICY

[$_{Min}$ Section Word Count: 1056 as of 27 Jan 2024]

Efforts to regulate any aspect of the digital information pipeline face challenges, particularly if it limits access to information through censorship. One of the essential virtues of the Web is the ability to access data from around the world. In the context of larger debates over the governance of AI, regulators across the globe are struggling to address perceived harms associated with GenAI while not unduly limiting innovation or driving it elsewhere. The starting point is to be clear about the available objectives, tools, and levers. "Regulation" includes rules, standards, and less formal forms of supervised self-regulation[7]. Policy interventions are still broader, including educational and social policies intended to build consumer resilience.

Spreading malicious content is already the subject of regulation in many jurisdictions. Though 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. We do not normally regulate 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 images are not shared.

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. 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.

Singapore adopted the Protection from Online Falsehoods and Manipulation Act (POFMA) [53], which empowers ministers to make correction orders for false statements of fact if it is in the public interest to do so. Though Singapore was criticised [71] when it adopted POFMA in 2019 [30], governments around the world are considering similar legislation to deal with the problem of fake news [11, 25]. Australia released a draft bill last year on Combatting Misinformation and Disinformation [51] that has been hotly debated [60] — 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 [38]. 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 this year, with promises to "improve the bill" [65] The EU legislation avoids defining disinformation but limits measures on socially harmful (as opposed to "illegal") [73] content to "very large online platforms" and "very large online search engines" — in essence, big tech companies like Google, Meta, and the like. Ofcom, the body tasked with enforcing the new UK law, states [52] that it is "not responsible for removing online content" but will help ensure that firms have effective systems in place 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 [67] 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 [20].

Correcting or blocking content is one of many means of addressing the problem. 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 [18] 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, though the ease of creation makes this a challenging game of catch-up. Tellingly, the US tech companies that agreed to voluntary watermarking [62] last year limited those commitments to images and video, echoed in the Biden Administration's October 2023 executive order [27]. Synthetic text is nearly impossible to label consistently; as it becomes easier to generate multimedia, images and video will likely go the same way.

As synthetic media becomes more common, it may be easier to label human content rather than AI. Trusted organizations may also watermark images so that users can identify where a photo originates. The problem here is that tracking such data requires effort, and many users demonstrate little interest in verifying whether information is true. Twitter (prior to its acquisition by Elon Musk) introduced a "read before you retweet" [32] prompt, which was intended to stop knee-jerk sharing of news based solely on the headline. It appeared to have a positive impact [39] but was not enough to stop the slide into toxicity post-Musk.

The ideal, of course, is for users to take responsibility for what they consume and share. 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. Nevertheless, concerns about the information diet of the public are as old as democracy itself. Some months before the US Constitution was drafted in 1787, Thomas Jefferson pondered whether it would be better to have a government without newspapers or newspapers without a government [66]. "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.

7 CONCLUSION AND OUTLOOK

[$_{Min}$ Section Word Count: 279 as of 27 Jan 2024]

The world over, as governments begin to regulate GenAI, it is still being determined whether legal restrictions or self-regulation will be more effective in the long run [28], as regulations are often outdated by the time they become policy. One of the reasons is also that governments and citizens still dispute precisely what aspects of AI need reining in and where the risks reside [28]. Lawsuits focusing on the copyright infringements around GenAI may miss the forest for the trees, as GenAI is increasingly affecting and altering how people confront and perceive their world [43]. Geographic borders do not bind the problems and consequences of MDM [37]; therefore, while the iGyro project will act as the epicenter of MDM technology and policy research in Asia, it will benefi from and leverage international collaborations toward the general goal of mitigating the risks of GenAI.

Even in the face of evolving adversarial technologies such as GenAI, human nature remains the critical driver of information diffusion on social media, as consumers continue to accept and even demand information rife with falsity. Without the tools to discern truth from falsity, vulnerable citizens will be influenced, misled, and possibly fall prey to those seeking personal gain. Moreover, without sufficient guardrails in place for GenAI by big tech [8], societies worldwide are in danger of descending into misinformation, disinformation, and malinformation (MDM) that threaten to disrupt the precarious balance that allows different identities, ideologies, and communities to coexist mutually and thrive [48]. iGyro's goals are to develop new technologies that reinforce the digital information pipeline, tools that empower consumers, and policies that enervate governments to apply a prophylactic approach to AI governance. Together, our three spheres of research will build a sustainable and flexible approach to digital information resilience.

8 ACKNOWLEDGEMENTS

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).

We want 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, et al. 2023. A new research agenda for African generative AI. *Nature Human Behaviour* 7, 11 (2023), 1839–1841.

[2] Saifuddin Ahmed. 2021. Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences* 182 (2021), 111074.

[3] 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.

[4] 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.

[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, et al. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour* 7, 9 (2023), 1502–1513.

[6] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[7] Robert Baldwin, Martin Cave, and Martin Lodge. 2011. *UNDERSTANDING REGULATION 2E P: Theory, Strategy, and Practice*. Oxford university press.

[8] Yochai Benkler. 2019. Don't let industry write the rules for AI. *Nature* 569, 7754 (2019), 161–162.

[9] Devis Bianchini, Carlo Bono, Alessandro Campi, Cinzia Cappiello, Stefano Ceri, Francesca De Luzi, Massimo Mecella, Barbara Pernici, and Pierluigi Plebani. [n. d.]. Challenges in AI-supported process analysis in the Italian judicial system: what after digitalization? Commentary paper. *Digital Government: Research and Practice* ([n. d.]).

[10] Claudi L Bockting, Eva AM 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.

[11] Lee C Bollinger and Geoffrey R Stone. 2022. *Social media, freedom of speech, and the future of our democracy*. Oxford University Press.

[12] 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.

[13] Zhe Chen, Aixin Sun, and Xiaokui Xiao. 2021. Incremental community detection on large complex attributed network. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 6 (2021), 1–20.

[14] Simon Chesterman. Forthcoming. Good Models Borrow, Great Models Steal: Intellectual Property Rights and Generative AI. *Policy & Society* (Forthcoming).

[15] European Commission. [n. d.]. The Digital Services Act package. *European Commission* ([n. d.]). 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] Richard Michael Cyert and Morris H DeGroot. 1987. *Bayesian analysis and uncertainty in economic theory*. Rowman & Littlefield.

[18] Philipe de Freitas Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro OS Vaz de Melo, and Fabrício Benevenuto. 2020. Can WhatsApp counter misinformation by limiting message forwarding?. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*. Springer, 372–384.

[19] 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* (2023), 1–3.

[20] Bruce J Dickson. [n. d.]. No "Jasmine" for China. *Current History* ([n. d.]). https://www.jstor.org/stable/45319730

[21] Benjamin Enke and Florian Zimmermann. 2019. Correlation neglect in belief formation. *The Review of Economic Studies* 86, 1 (2019), 313–332.

[22] Erik Eyster and Georg Weizsacker. 2010. Correlation neglect in financial decision-making. (2010).

[23] Leon Festinger. 1962. A Theory of Cognitive Dissonance (Evanston: Row Peterson, 1957).'Cognitive Dissonance'. *Scientific American* 207 (1962).

[24] Sean Fischer, Kokil Jaidka, and Yphtach Lelkes. 2020. Auditing local news presence on Google News. *Nature Human Behaviour* 4, 12 (2020), 1236–1244.

[25] Serena Giusti and Elisa Piras. 2020. *Democracy and fake news: information manipulation and post-truth politics*. Routledge.

[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. [n. d.]. FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. *The White House* ([n. d.]). 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] 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. *Jaidka, K., Fischer, S., Lelkes, Y., Wang, Y.(In Press). News Nationalization in a Digital Age: An examination of how local protests are covered and curated online. Forthcoming in the Annals of the American Academy of Political and Social Science* (2023).

[30] Shashi Jayakumar, Benjamin Ang, and Nur Diyanah Anwar. 2021. Fake News and disinformation: Singapore perspectives. *Disinformation and Fake News* (2021), 137–158.

[31] 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

[32] Michael Kan. [n. d.]. Twitter: Our 'Read Before You Retweet' Function Actually Works. *PCMag* ([n. d.]). https://www.pcmag.com/news/twitter-our-read-before-you-retweet-function-actually-works

[33] 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 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING.

[34] Samuli Laato, AKM 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.

[35] Laks V. S. Lakshmanan, Michael Simpson, and Saravanan Thirumuruganathan. 2019. Combating Fake News: A Data Management and Mining Perspective. *Proc. VLDB Endow.* 12 (2019), 1990–1993. https://api.semanticscholar.org/CorpusID:201653243

[36] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[37] Roy Ka-Wei Lee and Lynnette Hui Xian Ng. 2023. TikTok's Project Texas-Social Media Data Governance Across Geopolitical Lines. *Digital Government: Research and Practice* 4, 4 (2023), 1–5.

[38] Legislation.gov.uk. [n. d.]. Online Safety Act 2023. ([n. d.]). https://www.legislation.gov.uk/ukpga/2023/50/contents/enacted

[39] Rebecca Polly Leider. [n. d.]. Q: Do questions like "Do you want to share this article without clicking it first" help prevent the spread of misinformation? *News Literacy Matters* ([n. d.]). https://newsliteracymatters.com/2023/04/26/q-do-questions-like-do-you-want-to-share-this-article-without-clicking-it-first-help-prevent-the-spread-of-misinformation/

[40] 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.

[41] 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.

[42] Helen Margetts and Cosmina Dorobantu. 2019. Rethink government with AI. *Nature* 568, 7751 (2019), 163–165.

[43] Bernard Marr. [n. d.]. Is Google's Reign Over? ChatGPT Emerges As A Serious Competitor. *Forbes* ([n. d.]). https://www.forbes.com/sites/bernardmarr/2023/02/20/is-googles-reign-over-chatgpt-emerges-as-a-serious-competitor/?sh=4ab00b571072

[44] Solomon Messing and Sean J Westwood. 2014. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication research* 41, 8 (2014), 1042–1063.

[45] 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.

[46] Subhayan Mukerjee, Kokil Jaidka, and Yphtach Lelkes. 2022. The political landscape of the US Twitterverse. *Political Communication* 39, 5 (2022), 565–588.

[47] 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.

[48] Casey Newton. [n. d.]. The Taylor Swift deepfakes are a warning. *Platformer* ([n. d.]). https://www.platformer.news/taylor-swift-deepfake-nudes-x/

[49] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (2020). 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] Parliament of the Commonwealth of Australia. [n. d.]. Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023. ([n. d.]). https://www.infrastructure.gov.au/sites/default/files/documents/communications-legislation-amendment-combatting-misinformation-and-disinformation-bill2023-june2023.pdf

[52] Ofcom. [n. d.]. Online safety - what is Ofcom's role, and what does it mean for you? ([n. d.]). https://www.ofcom.org.uk/news-centre/2023/online-safety-ofcom-role-and-what-it-means-for-you

[53] Pofma Office. [n. d.]. Protection from Online Falsehoods and Manipulation Act. *Pofma Office* ([n. d.]). https://www.pofmaoffice.gov.sg/regulations/protection-from-online-falsehoods-and-manipulation-act/

[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 *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*, Yansong Feng and Els Lefever (Eds.). Association for Computational Linguistics, Singapore, 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 6981–7004. https://doi.org/10.18653/v1/2023.acl-long.386

[57] Sankalp Phartiyal and Krishna V Kurup. [n. d.]. WhatsApp curbs message forwarding in bid to deter India lynch mobs. *Reuters* ([n. d.]). 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). https://api.semanticscholar.org/CorpusID:147703897

[59] Alex Rees-Jones and Dmitry Taubinsky. 2020. Measuring "schmeduling". *The Review of Economic Studies* 87, 5 (2020), 2399–2438.

[60] Amy Remeikis. [n. d.]. Why is Labor's bill on combatting disinformation so controversial? *The Guardian* ([n. d.]). https://www.theguardian.com/australia-news/2023/oct/01/why-is-labors-bill-on-combatting-disinformation-so-controversial

[61] 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.

[62] Sabrina Siddiqui and Deepa Seetharaman. [n. d.]. White House Says Amazon, Google, Meta, Microsoft Agree to AI Safeguards. *The Wall Street Journal* ([n. d.]). https://www.wsj.com/articles/white-house-says-amazon-google-meta-microsoft-agree-to-ai-safeguards-eabe3680

[63] Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business information review* 34, 3 (2017), 150–160.

[64] 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. *SIGWEB Newsl.* 2015, Winter, Article 4 (feb 2015), 16 pages. https://doi.org/10.1145/2719943.2719947

[65] Josh Taylor. [n. d.]. Labor to overhaul misinformation bill after objections over freedom of speech. *The Guardian* ([n. d.]). https://www.theguardian.com/australia-news/2023/nov/13/labor-misinformation-bill-objections-freedom-of-speech-religious-freedom

[66] Jefferson Thomas. 1787. Jefferson's preference for "newspapers without government" over "government without newspapers". *Online Library of Liberty* (1787). https://oll.libertyfund.org/quote/jefferson-s-preference-for-newspapers-without-government-over-government-without-newspapers-1787

[67] The Straits Times. [n. d.]. 'Oh, bother': Chinese censors can't bear Winnie the Pooh. *The Straits Times* ([n. d.]). https://www.straitstimes.com/asia/east-asia/oh-bother-chinese-censors-cant-bear-winnie-the-pooh

[68] 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.

[69] Chris Tyler, KL 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.

[70] 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* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 4425–4434. https://doi.org/10.1145/3394171.3414692

[71] Human Rights Watch. [n. d.]. Singapore: 'Fake News' Law Curtails Speech. *Human Rights Watch* ([n. d.]). https://www.hrw.org/news/2021/01/13/singapore-fake-news-law-curtails-speech

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

[73] Konarski Xawery. [n. d.]. The Digital Services Act (DSA) and combating disinformation - 10 key takeaways. ([n. d.]). https://www.traple.pl/en/the-digital-services-act-dsa-and-combating-disinformation-10-key-takeaways/

[74] 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.