Combating Misinformation in the Era of Generative AI Models

Danni Xu
dannixu@u.nus.edu
National University of Singapore

Shaojing Fan
fanshaojing@nus.edu.sg
National University of Singapore

Mohan Kankanhalli
mohan@comp.nus.edu.sg
National University of Singapore

ABSTRACT

Misinformation has been a persistent and harmful phenomenon affecting our society in various ways, including individuals' physical health and economic stability. With the rise of short video platforms and related applications, the spread of multi-modal misinformation, encompassing images, texts, audios, and videos have exacerbated these concerns. The introduction of generative AI models like ChatGPT and Stable Diffusion has further complicated matters, giving rise to Artificial Intelligence Generated Content (AIGC) and presenting new challenges in detecting and mitigating misinformation. Consequently, traditional approaches to misinformation detection and intervention have become inadequate in this evolving landscape. This paper explores the challenges posed by AIGC in the context of misinformation. It examines the issue from psychological and societal perspectives, and explores the subtle manipulation traces found in AIGC at signal, perceptual, semantic, and human levels. By scrutinizing manipulation traces such as signal manipulation, semantic inconsistencies, logical incoherence, and psychological strategies, our objective is to tackle AI-generated misinformation and provide a conceptual design of systematic explainable solution. Ultimately, we aim for this paper to contribute valuable insights into combating misinformation, particularly in the era of AIGC.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Human-centered computing** → **HCI design and evaluation methods**.

KEYWORDS

Multimodal, Misinformation Detection, Generative AI models, AIGC

ACM Reference Format:
Danni Xu, Shaojing Fan, and Mohan Kankanhalli. 2023. Combating Misinformation in the Era of Generative AI Models. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3581783.3612704

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612704

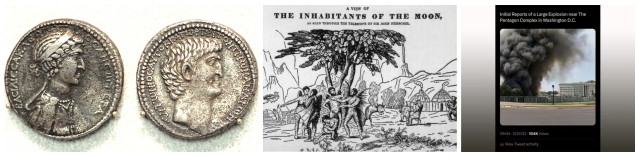


Figure 1: A brief overview of misinformation in human history.[2] Left: Misinformation has existed in human communication since ancient Rome. In 44-30 B. C., Octavian launched a 'fake news' war against Mark Anthony. Slogans written on coins portraying Antony as a womanizer and a drunk, that aided Octavian's victory over Antony. Middle: The New York newspaper *The Sun*'s 'Great Moon Hoax' of 1835 claimed that there was an alien civilization on the moon. Right: On 22 May 2023, an AI-generated image of Pentagon explosion sparked a temporary sell-off in the US stock market.

1 INTRODUCTION

Throughout history, misinformation[1] has had a detrimental impact on individuals and societies (Fig. 1). For example, in South Africa from 2000 to 2005, misinformation denying the link between HIV and AIDS resulted in misguided policies and approximately 330,000 excess deaths [22]; during the presidential election campaign of 2016, an alarming 126 million Americans were exposed to politically-oriented misinformation [51].

The rise of social media and short video platforms has fueled the widespread dissemination of multi-modal misinformation. New generative AI models, such as 'deepfakes', require minimal or no training, reducing the cost of generating multi-modal misinformation [36]. These models enable the manipulation and generation of various digital media content based on specific instructions, *e.g.*, Stable Diffusion generates images from textual prompts [56]. The proliferation of Artificial Intelligence Generated Content (AIGC) poses a significant challenge in combating misinformation [69].

Misinformation detection has been extensively studied, covering text-based, visual, and audio misinformation [49, 77]. Early approaches used text content, writing style, watermarks, and types of manual features. For visual features, artifacts, camera fingerprints, and biological signals have been explored [74]*[3]. Audio features like MFCC [46] and spectrogram [71] have been investigated. Despite advancements, previous models lacked explainability and scalability for general multi-modal misinformation [3]. Machines struggle

[1]We utilize the term "misinformation" as a broad umbrella encompassing any information that turns out to be false or misleading. We reserve the term "disinformation" specifically for instances where misinformation is intentionally disseminated with the aim to deceive or inflict harm. While the term "mal-information" refers to misinformation that stems from the truth but is often exaggerated in a way that misleads and causes potential harm.
[2]Image sources: Left: "Silver Tetradrachm Portraying Antony and Cleopatra", *www.worldhistory.org*, Mar, 2018. Middle: "Great Astronomical Discoveries Lately Made by Sir John Herschel, L.L.D. F.R.S. &c. at the Cape of Good Hope [From Supplement to the Edinburgh Journal of Science]", *The Sun*, August, 1835. Right: Screenshot on Twitter account @sentdefender, May, 2023.
[3] * represents non-peer reviewed reference from arXiv

to understand deceptive techniques or tricks employed by manipulators, and models trained on limited datasets struggle to assess the semantic veracity of AIGC. Effective combat against future misinformation requires understanding of very sophisticated deception and use of extensive background knowledge.

To address these challenges, this work provides an extensive overview of emerging generative AI models and their potential applications in misinformation. We also explore the future challenges arising from the proliferation of AIGC and the psychological perspective of misinformation drivers and consumer behaviour. Based on the above, we propose a conceptual design of multi-modal misinformation detection architecture tailored for the era of AIGC. Our approach proposes a cascade of detection mechanisms to cover a wide range of fabrication operations, including human-editing, AI-manipulation and human propagation. We categorize the deceptive traces into four layers: signal, perceptual, semantic, and human (psychology), where "signal" and "perceptual" are related to low-level inconsistencies, "semantic" to logical inconsistencies, and "human" to behavioral psychology. The proposed architecture aims to comprehensively detect multi-modal misinformation with an explainable model, providing a robust defense against the multi-faceted challenges posed by misinformation utilizing AIGC. The contribution of this work lies in three folds:

- We investigate the potential misinformation situations that may emerge with the involvement of AIGCs. This includes exploring the potential applications of generative AI models in misinformation generation, analyzing the characteristics of future misinformation, and addressing both the challenges and opportunities of countermeasures in the AIGC era.
- We identify influencing factors of misinformation from a social science perspective, including *confirmation bias* and *social proofing*, and offer valuable insights for combating misinformation through a multi-disciplinary approach.
- Our design proposes a comprehensive multi-modal misinformation detection framework that analyzes manipulation traces – signal and perceptual authenticity, semantic credibility, and behavioral psychology cues — throughout the entire fabrication process, emphasizing explainability and scalability in the era of AIGC.

2 A BRIEF OVERVIEW OF GENERATIVE MODELS IN MISINFORMATION

In this section, we provide an overview of generative models and their current and potential utilization in misinformation, and then review the current misinformation countermeasures.

2.1 AIGC and misinformation generation

2.1.1 AIGC. Generative AI models generate content based on user instructions. The content generated by these models is referred to as AI-Generated Content (AIGC) [18]*. Large generative models, powered by pre-trained networks with diffusion/transformer backbones and extensive datasets, generate high-quality contents. For example, ChatGPT [27], the protoypical generative AI model, can provide coherent responses to language prompts, though occasionally makes mistakes.

Table 1: A non-exhaustive compilation of notable generative AI models since 2022.

Year	Model	I/P	O/P	Arch	Inst.
2022	HTLM [5]	Text	Text	T.	Facebook AI
2022	DQ-BART [45]*	Text	Text	T.	AWS AI Labs (Amazon)
2022	ExT5 [8]	Text	Text	T.	Google Research & DeepMind
2023	LLaMA [75]*	Text	Text	T.	Meta Research
2023	ChatGLM2-6B [4]	Text	Text	T.	Tsinghua Univ.
2023	GPT−4 [52]*	Image, Text	Text	T.	OpenAI
2022	Watson et al. [80]*	Image	Image	D.	Google Research
2022	Cold Diffusion [12]*	Image	Image	D.	Univ. of Maryland & NYU
2022	DiffuseVAE [53]	Image	Image	Mix	UCI, IIT, and Google Research
2023	BLIP-2 [43]*	Image	Text	F. D.	Salesforce Research
2022	Flamingo [7]	Image	Text	F. D.	DeepMind
2023	Grounding [42]*	Image	Text	F. D.	CMU
2022	Stable-Diffusion[5] [56]*	Text	Image	D. D.	Stability AI
2022	Midjourney[6]	Text	Image	-	Midjourney
2022	DALL-E-2 [54]*	Text	Image	D. D.	OpenAI
2022	Imagen [60]*	Text	Image	D. D.	Google Research
2022	GLIGEN [44]*	Text, box	Image	D.	UW–Madison (multi-inst.)
2023	Muse [19]*	Text	Image	T.	Google Research
2022	AdaSpeech4 [84]*	Text	Audio	T.	Microsoft
2023	VALL-E [79]*	Text	Audio	T.	Microsoft
2023	VALL-E X [87]*	Text	Audio	T.	Microsoft
2023	GEN−1,GEN−2[7] [30]*	T.I.V.	video	D.	Runway
2023	Text2LIVE [13]*	T.I.V.	Video	T.	Weizmann Institute of Science; NVIDIA Research

Arch: Architecture. Inst.: Institution. I/P: Input. O/P: Output. T.: Transformer; D.: Diffusion; D.D.: Diffusion Decoder; F.D.: Frozen Decoder; Mix: Mixed Modeling. T.I.V.:Text or image or video
(Color) Model categorization based on payment and open source status (a crucial factor impacting its potential utilization in misinformation): cyan - open source, yellow - paywall-protected public API (no source code), orange - waitlist-based public API (no source code), light gray - no API or source code.
Readers can refer to [18] for more generative models with both uni- and multi- modal categories.

There are multiple popular models for diverse modality content generation, including text, image, audio, video and multi-modality generation. Table. 1 lists the main generative models available since 2022. Moreover, the above generative models, and other foundation models such as models for image captioning or image classification, have paved the way for the development of multi-round dialogue systems for both uni- and multi-modal tasks. Multimodal-GPT [32] engages users through simultaneous text and image interaction. HuggingGPT [65] goes a step further by integrating foundation models across all modalities, enabling broader multi-modal tasks such as image generation, text-to-speech, and text-to-video. Such expansion includes task planning and model selection to accommodate diverse instructions.

2.1.2 AIGC in misinformation generation. Deepfake models have emerged as a prominent research area. These deep-learning models are utilized to generate or manipulate image, video, and audio content with to mislead viewers into believing the content is authentic [6]. Deepfake models often utilize large generative models to generate high-quality and convincing content. Besides deepfake models, generative models can also be used to create semantically misleading text and multi-modal information. With the increasing accessibility of diverse generative techniques, generated misinformation is poised to dominate future disinformation landscapes. We discuss how generative models can be utilized in misinformation

creation and classify them into three classes — transformation, tampering, and generation — based on the inputs they require.

Transformation means transferring the information from one modality to another. Expansive generative models designed for various modality transformations, such as image captioning, video captioning, and text-to-image/audio/video conversions, can be employed to generate multi-modal information that *vividly* and *persuasively* convey misinformation. There are two primary approaches: 1) Creating images, audio clips, or videos to function as evidence alongside fabricated articles, or for live interactions scripted to deceive. 2) Creating misleading narratives using pre-existing evidence. **Tampering** is editing existing content. Large generative models require only modification prompts to generate desired results, making tampering much easier than traditional photo or audio editing software [19]. This can be applied to both uni- and multi-modal signals [6]. **Generation** refers to creating new misinformation from scratch. It encompasses both random and specific misinformation generation. Random misinformation means random attention-grabbing content created by AI models combining any popular contemporary topics to distract public or cause information disorders [4]. Specific generation is deliberate design targeting specific objectives, often resulting in heightened personal harms. This can be achieved by providing prompts or examples to multiple or multi-modal generative AI model [15, 64]. Notably, even in the absence of malicious cues, large-scale generative models can generate inappropriate content, biases, misinformation, largely as a result of inaccurate or biased training data [92].

2.2 Existing misinformation countermeasures

The primary executors can be used to categorize current countermeasures into user-level, platform-level, and government-level [33]. User-level resources include instructional and fact-checking portals[8]. Platforms like Facebook, Twitter, and YouTube also label, remove, reduce, prebunk, and inform about misinformation.[9] Globally, governments have passed legislation, removed content, and warned against misleading [33]. While all of the above measures have been helpful but the complexity and volume of misinformation makes it difficult to recognize it by human beings.

Recent detection techniques exploit cues from multi-modal data to identify misinformation. Some techniques consider text content features such as article titles, main content, and descriptions. Stylistic information, such as visual style of the fake websites, and word usage patterns, have also been examined [3]. Social traces, including user interactions (comments and likes), user profiles, and propagation networks, have been investigated [66]. More recently, studies have focused on analyzing the similarity and mismatch between modalities, aiming to improve detection accuracy [85, 88]. Feature fusion approaches for the above features are also being explored — concatenation [68], attention-based fusion [59], generative architecture [37], as well as graph architecture [25, 70]. Some researchers assess multi-dimension indicators. Rubin and Lukoianova [57] explored information veracity across three main dimensions: 1) objectivity/subjectivity, 2) truthfulness/deception, 3)

[8] [1] lists verified fact-checking portals.
[9] https://www.facebook.com/combating-misinfo
https://help.twitter.com/en/resources/addressing-misleading-info

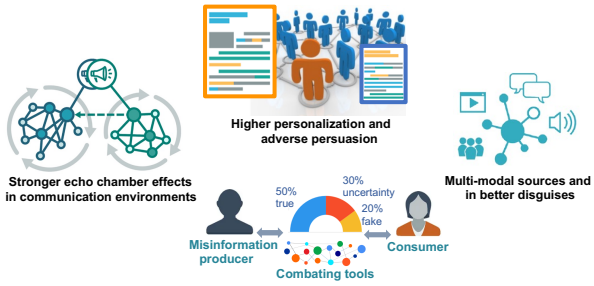


Figure 2: Summary of major challenges in future misinformation detection. The ecosystem of misinformation involves a perpetual loop of battle between the misinformation generators (forgers), the combating tools, and the users who consume the misinformation.

credibility/implausibility. More recently, Rubin [58] proposed layered assessment for credibility, which includes the source, medium, message, receiver and context. One recent work [86] first generates explanations with pre-trained models from evidence library. Some methods focus on early prevention. [61] proposes an image immunization method by injecting imperceptible noise into images to mislead diffusion-based image generation/edition methods to generate unrealistic images. To prevent undesirable generation of LLMs, recent works have also focused on aligning LLMs via various finetuning mechanisms [10]. Despite various proposed models, two ongoing challenges *scalability* and *explainability* have not been resolved. Most methods are trained and tested on limited datasets, making their performance uncertain for new types of misinformation. Additionally, most of current models provide binary verdicts without persuasive evidence to combat misinformation effects.

3 EMERGING THREATS OF AI GENERATED MISINFORMATION

In this section, we will delve into the potential future threats of misinformation (Fig. 2), specifically those introduced by AIGC, rendering many of the current countermeasures ineffective.

3.1 Evolution of misinformation: multimodal and in better disguise

Generative models like chatGPT [40] and Stable Diffusion [56] are gaining popularity. These advanced algorithms can generate highly realistic and convincing content across various modalities, including text, images, audios, and videos. While they have creative applications, they also pose a significant risk for the generation of multimodal misinformation. Malicious actors can exploit the capabilities of generative models to produce fake news articles accompanied by fabricated images or deepfake videos. Situations where deepfakes with face swap and face re-touching to create highly realistic videos of political figures are not uncommon [63, 81]. Artificial arts applications like Midjourney tool has created virtual historical events that look real but have never happened before [2]. Merging of multiple modalities by generative models presents a challenge in combating fake information, making it difficult for both users and algorithms to differentiate between truth and fabrication [31].

3.2 Increased deception: personalized and persuasive misinformation

Generative models usually have been trained on vast amounts of data to learn the underlying patterns and relationships between different types of information, such as text, images, and videos. By combining these learned representations, generative models can generate highly persuasive and coherent multi-modal content that aligns with specific narratives or outlines [34]. Furthermore, these models can also create personalized information by leveraging user data and capturing individual preferences and characteristics [11, 41]. For example, the study by Baird and colleagues work on personalised emotional audio generation based on neural network-based generative models [11]. While this has benefits, generative models can also be exploited by malicious attackers to generate personalized misinformation. By understanding user preferences, browsing history, and online behavior, generative models can generate personalized news articles, product recommendations, or even social media posts that cater specifically to an individual's interests. This personalized information can be highly persuasive, increasing engagement and the likelihood of dissemination [39], presenting significant challenges for combating misinformation.

3.3 Unleashing the torrent of misinformation: amplified scale and accelerated velocity

Generative AI models have revolutionized misinformation creation, enabling its generation at an unprecedented scale and speed. First, the user-friendly nature of generative AI tools, like ChatGPT, has democratized misinformation creation, allowing anyone to contribute to its proliferation. Second, with hardware advancements and optimization techniques, these models can quickly produce deceptive narratives, manipulated images, and deepfake videos, leading to rapid spread across online platforms and social networks [31]*. Furthermore, social media users' exposure to information that challenges their worldviews can be limited when communication environments foster confirmation of previous beliefs, known as echo chambers [14, 26]. In this case, the larger scale and prevalent misinformation would potentially reinforce users' false beliefs.

4 THE PSYCHOLOGICAL AND SOCIETAL DRIVERS OF MISINFORMATION

Technology alone is insufficient to combat misinformation. While advancements have provided tools for analysis and mitigation, the core issue lies with human consumers. Therefore, adopting a human-centered perspective is essential. Insights from social science help understand the causes and behavioral patterns driving misinformation. In this section, we discuss the social science aspects of misinformation. We also explore their implications on the design of detection models.

4.1 Psychological theories of misinformation

The psychology and history of misinformation cannot be fully grasped without taking into account the *information deficit model*, which suggests that public misconceptions or lack of knowledge about a particular topic can be attributed to a deficit or lack of information [28, 48]. This highlights the significance of data collection and processing within misinformation detection algorithms, emphasizing the key role of equipping computer models with sufficient, reliable, and responsible data during the design phase [89].

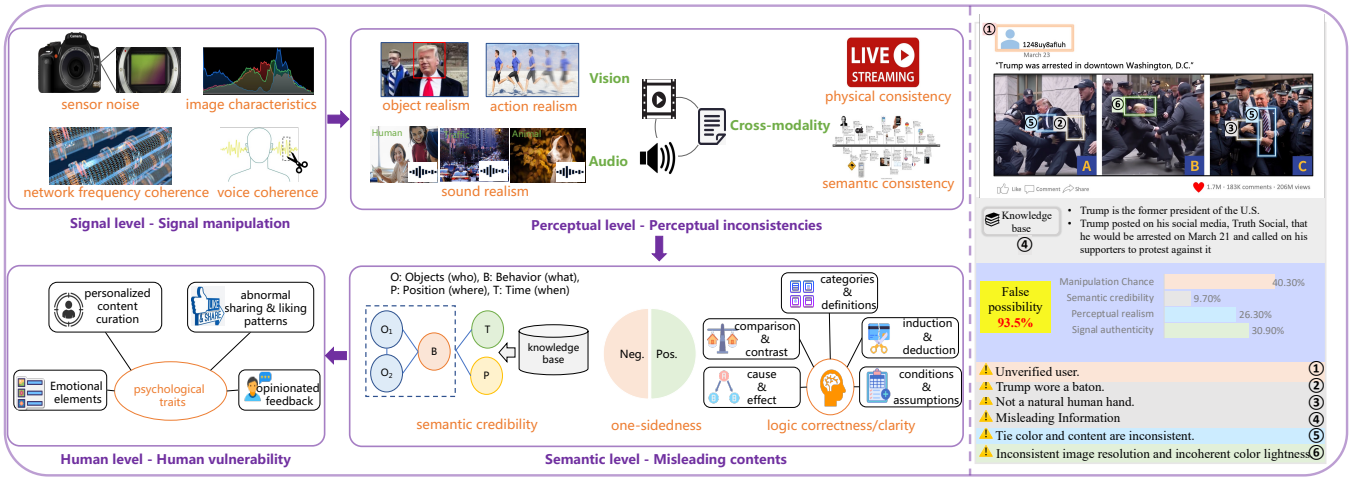
Though the information deficit model has been intensively studied and applied, the drivers of misinformation are multifold. One key factor is *confirmation bias*, where individuals seek and interpret information that aligns with their existing beliefs while disregarding contradictory evidence [90]. For example, a study in the United States [76] have shown that liberals may label conservative outlets (*e.g.*, Fox News) as 'fake news', while well-known liberal outlets (*e.g.*, CNN) are described as 'fake news' by conservatives [76]. Additionally, the proliferation of digital technologies has significantly contributed to the phenomenon of *information glut*, which refers to the overwhelming abundance of information available to individuals that surpasses their capacity to effectively process and make sense of it [73]. Information glut can lead to difficulties in decision-making, and to rely on easily accessible or vivid examples rather than considering broader evidence [17]. Moreover, affective elements, such as mood and emotions, play a significant role. Emotional content influences the formation of false beliefs and the spread of misinformation [29, 35, 72]. Negative emotions have been found to promote misinformation sharing and impact attitudes and behaviors [72, 83]. Studies have shown that the use of emotional language in tweets about political issues increases retweets by 20% [16]. Understanding the affective aspects is crucial for effective misinformation detection. Furthermore, misinformation can exert a lasting influence on individuals' thinking, even after they receive and accept a correction. This phenomenon is known as the *continued influence effect (CIE)* [38, 78]. The CIE occurs when people selectively retrieve the misinformation or fail to retrieve the correction. If a correction is not sufficiently encoded and integrated with the misinformation in memory, individuals may continue to rely on the outdated misinformation in their reasoning [62]. Enhancing the explainability of misinformation detection could help mitigate the impact of the CIE.

4.2 Societal and interpersonal factors of misinformation

Social networks and interpersonal communication is a crucial factor. Misinformation can quickly disseminate through social networks, facilitated by the rapid sharing on platforms and the influence of opinion leaders. Chen [20] found that people's sharing of misinformation on social media is mainly influenced by their personalities and motivations. *social interaction* is found to be the most common purpose of using social media [24, 82]. Chen and colleagues [21] reported that social interaction overwhelmed truthfulness in misinformation sharing decisions in their experiments.

The concept of *social proof*, where individuals look to others' actions and opinions to guide their own behavior, can further amplify the spread of misinformation when people observe others accepting or endorsing false information [50]. For example, Colliander [23] found that others' negative comments critical of a fake news article lowered users' intention to share the fake news, and vice versa.

The aforementioned research provides insights for designing a misinformation detection framework. It suggests employing user cognitive background to address confirmation bias and information



(a) Framework (b) Prototype User Interface

Figure 3: (a) Illustration of the proposed conceptual framework for multi-modal misinformation detection. The framework extracts fabrication traces from four levels – signal level, perceptual level, semantic level, and human level. (b) Conceptual Prototype Interface. 4-color boxes in the pictures correspond to 4 levels of misinformation characteristics; It displays the overall false possibility score, detailed scores of the 4-level misinformation characteristics, explanations, and the reference knowledge

overload. Explainability in detection models can help mitigate the CIE. Additionally, integrating sentiment analysis could potentially improve detection considering affective factors. In the next section, we propose a conceptual multi-modal framework based on these analyses.

5 A DESIGN OF MULTI-MODAL FRAMEWORK FOR CURBING MISINFORMATION IN AIGC ERA

5.1 Motivation

Sections 2 and 3 underscore the challenges introduced by AIGC in combating misinformation and emphasize the significance of scalability and explainability in misinformation detection. In this section, we propose a *conceptual framework* composed of 4 levels, signal, perceptual, semantic and human, for comprehensive and explainable multi-modal misinformation detection (Fig. 3(a)).

During the process of misinformation generation and dissemination, we identify three common elements: techniques, misleading contents, and propagation processes, each of which leaves behind discernible fabrication traces. 1) Techniques: Despite the technical development of generative models, AIGCs still exhibit inherent technical flaws. From Sec. 2.1, current large generative models rely on data training on collected digitized samples, lacking features of physical signals in the recording and transmitting process. Besides, the digital samples are less complex than real world, with limited views, dimensions and length. Therefore, AI lack of overall understanding of objects from the multiple perspectives, *e.g.*, it is not easy for AI to learn physical laws from these simplified samples. Therefore, AIGC often have perceptual artifacts in realism and naturalism. Hence, we propose analyzing the above technique flaws at two levels: the *signal level* and the *perceptual level*. 2) Misleading contents: Diverse misleading contents require the understanding of both content and background knowledge. Therefore, we propose *semantic level* for misleading contents. 3) Propagation operations: Our analyses in Sec. 4 highlight four psychological features of users, which offer novel insights into propagation operations. Building upon these findings, we propose psychological cues as *human-level* features.

5.2 Signal-level

In the process of collecting, digitally transforming, and transmitting authentic visual and audio content, various types of traces are formed, such as camera sensor noise [47] and electrical network frequency [55]. These traces are absent in AIGCs. Additionally, natural contents exhibit certain signal characteristics, such as continuous lightness and sharpness in images, and coherent MFCC, tone, and sound quality in voices (both main sounds and background sounds). However, in AIGCs or manipulated contents, these natural characteristics may be absent or distorted. Therefore, these changes in signal characteristics can be utilized to represent signal authenticity — one key attribute of misinformation.

5.3 Perceptual-level

Authentic content should be easily comprehensible to humans. Conversely, content that is unrecognizable or misleading is often the result of a potential AI generation. At the perceptual level, the assessment of content realism extends to both uni- and multi-modal contexts. On uni-modality, visual realism involves the realism of object appearance and movement in both spatial and temporal dimensions, audio realism entails that each audio composition corresponds a physical sound in the aspects of the physical properties (*e.g.*, 'timbre') and semantic properties (*e.g.*, language). On cross-modality, realism refers to the cross-modal consistency. For narrative multi-modal content, it means semantic consistency, e.g., an image showing a 'bird' is described as 'plane' in text. While for

live videos it also includes physical consistency, which means the sound and vision should belong to the same subjects. For example, the mouth shapes and utterances should be synchronous, and different speaker should have voice different timbre. In summary, the perceptual level rules out the unrealistic contents in physical world.

5.4 Semantic-level

Authenticity of multimedia content at the signal and perceptual levels is a key challenge. However, the manipulation of content remains the primary differentiating factor between authentic information and misinformation. Manipulators can fabricate events or combine existing events in misleading ways. Consequently, detecting fabricated events and illogical connections becomes crucial. Additionally, the filtration and selective presentation of information for the benefit of certain groups can lead to misinformation by promoting one-sidedness. At the semantic level, we assess the credibility of each piece of information using a human knowledge base and evaluate the logical coherence and contextual reasonability. Furthermore, we evaluate the overall one-sidedness of the information based on subjective expression and breadth of coverage.

5.5 Human-level

In Section 4, we illustrated how individuals' information consumption is shaped by their cognitive characteristics and social background. It is crucial to recognize that these human-related features (*e.g.,* confirmation bias [90], social proof [50]) could potentially be exploited by malicious attackers for the generation and dissemination of misinformation. At the human level, our framework aims to identify psychological traits associated with conspiracy beliefs and persuasive behavior. These traits encompass elements that evoke strong emotions, feedback with highly opinionated attitudes, abnormal sharing and liking patterns, as well as content that clearly caters to the specific interests and needs of individual users. By incorporating these psychological traits, we aim to equip the framework with insights into the mindset of misinformation distributors. This allows for the information analysis from the standpoint of those who generate and spread misinformation, thereby contributing to the detection of misinformation from a higher and more comprehensive human perspective.

5.6 Explainability and Scalability

Explainability plays a significant role in mitigating the CIE discussed in Sec 4.1. Attaining explainability involves constructing more interpretable models, creating effective user interfaces, and understanding the psychological requirements for interpretation [9]. Existing models, although effective in highlighting important text features [49], lack the ability to explain the overall fabrication traces, and leave a gap on multi-modal misinformation. This framework focuses on general misinformation, encompassing both uni- and multi-modal formats, human- and AI-generated misinformation. To provide users with comprehensive explanations, highlighting the locations and methods of counterfeiting, we propose its prototype interface (Fig. 3(b)), with 4-dimension veracity and explanations provided. 1) Signal authenticity: the possibility and proofs that the signals are from actual recording devices. 2) Perceptual authenticity: the possibility and proofs that the multi-modal content is not generated or edited by AI or any software. 3) Semantic credibility: The credibility of overall semantics with logic inference and its explanations with knowledge base. 4) Manipulation likelihood: The abnormal content, publisher, propagation pattern and potential psychology tricks used for manipulating public.

5.7 Challenges and opportunities in AIGC era

Malicious adversaries will continuously enhance their malicious endeavors and create large volume of new misinformation with large AI models, making it imperative for protectors to devise **robust** and **real-time** countermeasures. We foresee the following specific challenges during the framework development. At the signal-level, signal degrades during compression and transmission. One potential solution is identifying key signal features robust to compression and transmission. At the perceptual-level, cross-modal consistency varies by content type. Employing context-aware inconsistency detection may help [91]. Moreover, anomalies that contradict common sense exhibit greater resilience than superficial flaws. At the semantic level, real-time and reliable knowledge update is challenging due to rapid large-scale misinformation dissemination. To mitigate this, an interactive user portal and real-time searches from verified sources can provide the latest information. Privacy concern is a key challenge at the human-level. The framework should auto-mask privacy-revealing content to ensure user privacy [67].

It is worth noting that AIGCs could also provide opportunities for countering misinformation. They can serve as real-time knowledge bases for evaluating information credibility. The expressive and visualization capabilities of AIGCs enable the generation of explainable and influential misinformation analyses for users. Moreover, AIGCs have the ability to analyze users' vulnerabilities and provide personalized recommendations for protection.

6 CONCLUSION AND DISCUSSION

This paper analyzes the challenges of misinformation in the AIGC era. We propose the conceptual framework for multi-modal misinformation detection, which addresses the future issues of large-scale generated misinformation and the impact of human psychology. Our design model operates on four levels: signal, perceptual, semantic, and human, covering the entire process of misinformation manipulation. It aims to provide real-time, comprehensive, and explainable detection measures to bridge the existing gap.

Lastly, the relationship between misinformation distributors and detectors is a constant cat-and-mouse game. Distributors evolve tactics while detectors enhance techniques to stay ahead. While technological solutions are essential, regulation, legislation, and education are vital in holding malicious actors accountable and empowering individuals with critical thinking and media literacy.

ACKNOWLEDGMENTS

This research is supported by the MOE (Ministry of Education Singapore) T3 Grant (MOE-MOET32022-0001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of MOE, Singapore.

REFERENCES

[1] 2022. IFCN Code of Principles. [Online]. Available: https://ifcncodeofprinciples.poynter.org/signatories. Accessed: June. 2. 2023.

[2] 2023. Midjourney creates fake events. https://interestingengineering.com/culture/ai-create-images-fake-events. Accessed: 2023-05-30.

[3] Sara Abdali. 2022. Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. *arXiv preprint arXiv:2203.13883* (2022).

[4] Hannah Abraham. 2023. Reports of Actor 'Saint Von Colucci' Dying of Cosmetic Surgeries to Resemble BTS Singer Jimin Appear to Be Elaborate Hoax That Used AI. [Online]. Available: https://variety.com/2023/film/asia/jimin-saint-von-colucci-hoax-ai-1235597897/. Accessed: Jun. 3. 2023.

[5] Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2022. HTLM: Hyper-Text Pre-Training and Prompting of Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=P-pPW1nxf1r

[6] Saadaldeen Rashid Ahmed, Emrullah Sonuç, Mohammed Rashid Ahmed, and Adil Deniz Duru. 2022. Analysis Survey on Deepfake detection and Recognition with Convolutional Neural Networks. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. 1–7. https://doi.org/10.1109/HORA55278.2022.9799858

[7] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* 35 (2022), 23716–23736.

[8] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Vzh1BFUCiIX

[9] Aleksey Averkin. 2022. Explainable Artificial Intelligence. In *Workshop on Intelligent Information Systems*. 4–6.

[10] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[11] Alice Baird, Shahin Amiriparian, and Björn Schuller. 2019. Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–5.

[12] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2022. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392* (2022).

[13] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*. Springer, 707–723.

[14] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542.

[15] Ali Borji. 2022. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586* (2022).

[16] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* 114, 28 (2017), 7313–7318.

[17] Philip J Calvert. 2001. Scholarly misconduct and misinformation on the World Wide Web. *The Electronic Library* 19, 4 (2001), 232–240.

[18] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226* (2023).

[19] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704* (2023).

[20] Xinran Chen. 2016. The influences of personality and motivation on the sharing of misinformation on social media. *IConference 2016 Proceedings* (2016).

[21] Xinran Chen and Sei-Ching Joanna Sin. 2013. 'Misinformation? What of it?'Motivations and individual differences in misinformation sharing on social media. *Proceedings of the American Society for Information Science and Technology* 50, 1 (2013), 1–4.

[22] Pride Chigwedere, George R Seage III, Sofia Gruskin, Tun-Hou Lee, and Max Essex. 2008. Estimating the lost benefits of antiretroviral drug use in South Africa. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 49, 4 (2008), 410–415.

[23] Jonas Colliander. 2019. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior* 97 (2019), 202–215.

[24] Nick Couldry and Jose Van Dijck. 2015. Researching social media as if the social mattered. *Social Media+ Society* 1, 2 (2015), 2056305115604174.

[25] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 492–502.

[26] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the national academy of Sciences* 113, 3 (2016), 554–559.

[27] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71 (2023), 102642.

[28] Ullrich KH Ecker. 2017. Why rebuttals may not work: the psychology of misinformation. *Media Asia* 44, 2 (2017), 79–87.

[29] Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* 1, 1 (2022), 13–29.

[30] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011* (2023).

[31] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint arXiv:2301.04246* (2023).

[32] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. arXiv:2305.04790 [cs.CV]

[33] Ankur Gupta, Neeraj Kumar, Purnendu Prabhat, Rajesh Gupta, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. 2022. Combating fake news: Stakeholder interventions and potential solutions. *Ieee Access* 10 (2022), 78268–78289.

[34] GM Harshvardhan, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. 2020. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review* 38 (2020), 100285.

[35] Christy Galletta Horner, Dennis Galletta, Jennifer Crawford, and Abhijeet Shirsat. 2021. Emotions: The unexplored fuel of fake news on social media. *Journal of Management Information Systems* 38, 4 (2021), 1039–1066.

[36] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. 2023. Implicit Identity Driven Deepfake Face Swapping Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4490–4499.

[37] Ramji Jaiswal, Upendra Pratap Singh, and Krishna Pratap Singh. 2021. Fake News Detection Using BERT-VGG19 Multimodal Variational Autoencoder. In *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE, 1–5.

[38] Hollyn M Johnson and Colleen M Seifert. 1994. Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of experimental psychology: Learning, memory, and cognition* 20, 6 (1994), 1420.

[39] Maurits Kaptein, Panos Markopoulos, Boris De Ruyter, and Emile Aarts. 2015. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies* 77 (2015), 38–51.

[40] Ravi Kashyap and ChatGPT OpenAI. 2023. A First Chat with ChatGPT: The First Step in the Road-Map for AI (Artificial Intelligence)... *Available at SSRN* (2023).

[41] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274.

[42] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823* (2023).

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[44] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.

[45] Zheng Li, Zijian Wang, Ming Tan, Ramesh Nallapati, Parminder Bhatia, Andrew Arnold, Bing Xiang, and Dan Roth. 2022. DQ-BART: Efficient Sequence-to-Sequence Model via Joint Distillation and Quantization. *arXiv preprint arXiv:2203.11239* (2022).

[46] Tianyun Liu, Diqun Yan, Rangding Wang, Nan Yan, and Gang Chen. 2021. Identification of fake stereo audio using SVM and CNN. *Information* 12, 7 (2021), 263.

[47] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. 2006. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security* 1, 2 (2006), 205–214.

[48] Theresa M Marteau, Amanda J Sowden, and David Armstrong. 1998. Implementing research findings into practice: beyond the information deficit model. *Getting research findings into practice* 2 (1998), 36–42.

[49] Ken Mishima and Hayato Yamana. 2022. A survey on explainable fake news detection. *IEICE TRANSACTIONS on Information and Systems* 105, 7 (2022), 1249–1257.

[50] Muhammad Naeem. 2021. The role of social media to generate social proof as engaged society for stockpiling behaviour of customers during Covid-19 pandemic. *Qualitative Market Research: An International Journal* 24, 3 (2021), 281–301.

[51] Kuldeep Nagi. 2018. New social media and impact of fake news on society. *ICSSM Proceedings, July* (2018), 77–96.

[52] OpenAI. 2023. GPT-4 Technical report. *arXiv preprint arXiv:2303.08774* (2023).

[53] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. 2022. DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=ygoNPRiLxw

[54] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[55] Paulo Max Gil Innocencio Reis, João Paulo Carvalho Lustosa da Costa, Ricardo Kehrle Miranda, and Giovanni Del Galdo. 2016. ESPRIT-Hilbert-based audio tampering detection with SVM classifier for forensic analysis via electrical network frequency. *IEEE Transactions on Information Forensics and Security* 12, 4 (2016), 853–864.

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[57] Victoria Rubin and Tatiana Lukoianova. 2013. Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online* 24, 1 (2013), 4.

[58] Victoria L Rubin. 2022. Credibility Assessment Models and Trust Indicators in Social Sciences. In *Misinformation and Disinformation: Detecting Fakes with the Eye and AI*. Springer, 61–94.

[59] Tanmay Sachan, Nikhil Pinnaparaju, Manish Gupta, and Vasudeva Varma. 2021. SCATE: shared cross attention transformer encoders for multimodal fake news detection. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 399–406.

[60] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[61] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. 2023. Raising the Cost of Malicious AI-Powered Image Editing. *arXiv preprint arXiv:2302.06588* (2023).

[62] Jasmyne A Sanderson, Simon Farrell, and Ullrich KH Ecker. 2022. Examining the role of information integration in the continued influence effect using an event segmentation approach. *PloS one* 17, 7 (2022), e0271566.

[63] Jia-Wen Seow, Mei-Kuan Lim, Raphaël C-W Phan, and Joseph K Liu. 2022. A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing* (2022).

[64] Shweta Sharma. 2023-03-24. Trump shares deepfake photo of himself praying as AI images of arrest spread online. https://www.independent.co.uk/news/world/americas/us-politics/donald-trump-ai-praying-photo-b2307178.html. (2023-03-24).

[65] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. *arXiv preprint arXiv:2303.17580* (2023).

[66] Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management* 58, 5 (2021), 102618.

[67] Anshu Singh, Shaojing Fan, and Mohan Kankanhalli. 2021. Human attributes prediction under privacy-preserving conditions. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4698–4706.

[68] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 39–47.

[69] Shahab Saquib Sohail, Faiza Farhat, Yassine Himeur, Mohammad Nadeem, Dag Øivind Madsen, Yashbir Singh, Shadi Atalla, and Wathiq Mansoor. 2023. The Future of GPT: A Taxonomy of Existing ChatGPT Research, Current Challenges, and Possible Future Directions. *Current Challenges, and Possible Future Directions* (2023).

[70] Chenguang Song, Kai Shu, and Bin Wu. 2021. Temporally evolving graph neural network for fake news detection. *Information Processing & Management* 58, 6 (2021), 102712.

[71] Nishant Subramani and Delip Rao. 2020. Learning efficient representations for fake speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5859–5866.

[72] Melanie B Tannenbaum, Justin Hepler, Rick S Zimmerman, Lindsey Saul, Samantha Jacobs, Kristina Wilson, and Dolores Albarracín. 2015. Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological bulletin* 141, 6 (2015), 1178.

[73] Tonyia J Tidline. 1999. The mythology of information overload. (1999).

[74] Aniruddha Tiwari, Rushit Dave, and Mounika Vanamala. 2023. Leveraging Deep Learning Approaches for Deepfake Detection: A Review. *arXiv preprint arXiv:2304.01908* (2023).

[75] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[76] Sander Van der Linden, Costas Panagopoulos, and Jon Roozenbeek. 2020. You are fake news: Political bias in perceptions of fake news. *Media, Culture & Society* 42, 3 (2020), 460–470.

[77] 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*. 4425–4434.

[78] Nathan Walter and Riva Tukachinsky. 2020. A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication research* 47, 2 (2020), 155–177.

[79] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv:2301.02111 [cs.CL]

[80] Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. 2022. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*.

[81] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).

[82] Anita Whiting and David Williams. 2013. Why people use social media: a uses and gratifications approach. *Qualitative market research: an international journal* (2013).

[83] Manli Wu and Yiming Pei. 2022. Linking social media overload to health misinformation dissemination: An investigation of the underlying mechanisms. *Telematics and Informatics Reports* 8 (2022), 100020.

[84] Yihan Wu, Xu Tan, Bohan Li, Lei He, Sheng Zhao, Ruihua Song, Tao Qin, and Tie-Yan Liu. 2022. AdaSpeech 4: Adaptive Text to Speech in Zero-Shot Scenarios. arXiv:2204.00436 [eess.AS]

[85] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management* 58, 5 (2021), 102610.

[86] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2733–2743.

[87] Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling. arXiv:2303.03926 [cs.CL]

[88] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-Aware Multi-modal Fake News Detection. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*. Springer, 354–367.

[89] Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. 2019. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 836–837.

[90] Yanmengqian Zhou and Lijiang Shen. 2022. Confirmation bias and the persistence of misinformation on climate change. *Communication Research* 49, 4 (2022), 500–523.

[91] Jianming Zhu, Peikun Ni, and Guoqing Wang. 2020. Activity minimization of misinformation influence in online social networks. *IEEE Transactions on Computational Social Systems* 7, 4 (2020), 897–906.

[92] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).