The Psychology of Falsehood: A Human-Centric Survey of Misinformation Detection

Arghodeep Nandi[1], Megha Sundriyal[2], Euna Mehnaz Khan[3], Jikai Sun[3], Emily Vraga[3], Jaideep Srivastava[3], Tanmoy Chakraborty[1]
[1]Indian Institute of Technology Delhi, [2]Max Planck Institute for Security and Privacy,
[3]University of Minnesota - Twin Cities
{eez248395, tanchak}@iitd.ac.in, *megha.sundriyal@mpi-sp.org*,
{khan0586, sun00948, ekvraga, srivasta}@umn.edu

Abstract

Misinformation remains one of the most significant issues in the digital age. While automated fact-checking has emerged as a viable solution, most current systems are limited to evaluating factual accuracy. However, the detrimental effect of misinformation transcends simple falsehoods; it takes advantage of how individuals perceive, interpret, and emotionally react to information. This underscores the need to move beyond factuality and adopt more human-centered detection frameworks. In this survey, we explore the evolving interplay between traditional fact-checking approaches and psychological concepts such as cognitive biases, social dynamics, and emotional responses. By analyzing state-of-the-art misinformation detection systems through the lens of human psychology and behavior, we reveal critical limitations of current methods and identify opportunities for improvement. Additionally, we outline future research directions aimed at creating more robust and adaptive frameworks, such as neuro-behavioural models that integrate technological factors with the complexities of human cognition and social influence. These approaches offer promising pathways to more effectively detect and mitigate the societal harms of misinformation.

1 Introduction

The digital age has fundamentally transformed the way information is disseminated, leading to the rapid and widespread propagation of misinformation (Amoruso et al., 2020; Augenstein et al., 2023). Misinformation is more than just the existence of incorrect information; it also entails complex relationships between the information and the entities that consume it. As individuals navigate this complex network of information, their perceptions and behaviours are shaped by a number of psychological and social influences (Ecker et al., 2022).

Misinformation is primarily classified into various categories, including fake news, misleading information, fabricated content and similar other forms (Altay et al., 2023). However, these classifications often fail to account for the fact that even technically accurate information, when presented out of context or with bias, can also contribute to the spread of mass misperceptions. For example, the headline *"A doctor dies after receiving the second dose of a vaccine"* is factually correct. However, it contributed to widespread vaccine-related misperceptions by omitting essential contextual details such as the cause of death, timing, and medical background, which allowed the narrative to exploit emotional bias and reinforce existing confirmation biases (Grady and Mazzei, 2021). Similarly, satirical news, hyper-partisan reporting, and propaganda are frequently debated as sources of misinformation (Patwa et al., 2021; Alam et al., 2022a; Altay et al., 2023). To address these challenges, the focus needs to pivot from factual accuracy to comprehending how social audiences perceive and interpret information. In this study, we explore the role of perception, cognition, and social psychology in shaping the dynamics of misinformation.

Before the emergence of techniques addressing complex cases of misinformation, earlier studies primarily relied on relational operators to match claims with supporting evidences (Krishna et al., 2021; Thorne et al., 2018; Põldvere et al., 2023; Sundriyal et al., 2022b). These approaches focused on evidence retrieval, effectively using evidence to train models. Advancements in Large Language Models (LLMs) have significantly reshaped the misinformation detection landscape. Recent research focuses on nuanced aspects such as ambiguity, perception, and social energy to address contemporary challenges of misinformation (Song, 2021). Emerging literature highlights the application of LLMs to simulate user reactions based on different user profiles and generate interaction graphs (Wan et al., 2024). This underscores the growing importance of graph-based algorithms in

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8508–8525
November 4-9, 2025 ©2025 Association for Computational Linguistics

Survey Paper	Cognitive Framing	Human-Centric Analysis	Behavioral Insights	Relevance to HC
Guo et al. (2022)	✗ Not addressed	✗ No focus on human factors	✗ Absent	● Low
Hardalov et al. (2022)	✓ Discusses stance as a proxy for belief and agreement	✓ Explores how stance reflects user attitudes	▲ Mentions role in misinformation spread	● Moderate
Akhtar et al. (2023)	✓ Notes higher credibility of multimodal misinformation	▲ Highlights human susceptibility to images/videos	▲ Suggests need for user studies	● Moderate
Vladika and Matthes (2023)	✗ Focuses on technical aspects	✗ No discussion on user cognition	✗ Absent	● Low
Nakov et al. (2024)	▲ Touches on media bias perception	▲ Discusses impact on public trust	▲ Limited behavioral analysis	● Moderate
Panchendrarajan and Zubiaga (2024)	✗ Emphasizes linguistic challenges	✗ No cognitive aspects discussed	✗ Absent	● Low
Eldifrawi et al. (2024)	▲ Emphasizes the importance of explainability in fact-checking systems	▲ Discusses the need for user-understandable justifications	▲ Highlights the role of user trust in automated fact-checking	● Moderate

Table 1: Psychological and cognitive focus in misinformation survey papers. **Notation:** A cross (✗) indicates absence, a triangle (▲) signifies a tangential presence, and a tick (✓) denotes direct focus or presence. HC stands for Human Cognition.

combating misinformation. The future of misinformation prevention will likely leverage Graph Neural Networks (GNNs) in conjunction with LLMs to tackle misinformation's social and psychological complexities. With the objective of better addressing the complexities of misinformation in today's context, recent studies have also started to explore these issues by developing specialised datasets with ambiguous names (Chiang and Lee, 2024) and persuasive contents (Xu et al., 2024). Sundriyal et al. (2024) hypothesises that the user's reactions to misinformation often reveal its accuracy. They argue that collective judgment, as expressed through public reactions, can serve as a reliable signal of the truthfulness of a given piece of information. Recently, platforms have also adopted this idea, moving toward greater reliance on community-driven tools such as community notes for content moderation (Borenstein et al., 2025). Another essential factor in disseminating misinformation is the writing style. Whether deliberate or inadvertent, the style in which misinformation is presented can significantly impact its believability. Recent studies have introduced style-agnostic training methods to reduce the impact of writing styles on misinformation detection (Wu et al., 2024).

As the boundaries between AI models and psychological processes blur, there's a growing need to focus research on their intersection. This evolution signals the rise of stronger computational models that leverage data features and network structures to counter increasingly sophisticated misinformation campaigns. For instance, Loth et al. (2024) highlighted how Generative AI is transforming the landscape by automating the creation of misinformation (text and multimodal), effectively manipulating cognition, perception, and attitudes with minimal cost and unprecedented reach. Studies underscore the psychological impact of these advancements and the pressing need for countermeasures (Alam et al., 2022b; Kou et al., 2022).

As evident from Table 1, various surveys on misinformation lack the required focus on cognitive framing, human-centric analysis, and behavioural insights, and their overall relevance to human cognition remains moderate to low. In this paper, we argue for a shift toward analysing misinformation through the lens of its cognitive and psychological impacts. This transition is necessary because humans are not simply programmed machines, but conscious beings influenced by various biases and shaped by environmental and internal factors (Keller, 2010). Addressing these vulnerabilities requires innovative tools that integrate technology, social psychology, and insights from cognition.

2 Classic Simplicity, But Trust it Sparingly

With the rise of the Internet, the problem of misinformation has become increasingly significant. In

its early forms, misinformation often involved presenting incorrect facts in a manner that mimicked trustworthy sources. Automated Fact-Checking (AFC) modules were developed to combat this. A recent survey on AFC outlined four general stages involved in these systems (Eldifrawi et al., 2024). Interestingly, these stages can be compared to the multi-store model of memory (Atkinson and Shiffrin, 1968), which posits that memory consists of three key components: the sensory register (SR), short-term memory (STM), and long-term memory (LTM). Each AFC stage parallels memory store functions, as shown in Figure 1.

Check-worthy Claims. We encounter numerous stimuli daily, but not all require our attention. Only claims that are significant, verifiable, and have the potential to cause harm or mislead should be addressed (Wright and Augenstein, 2020; Guo et al., 2022; Sundriyal et al., 2025). The concept of claim check-worthiness can be linked to the process of attention (Atkinson and Shiffrin, 1968), which facilitates the transfer of information from sensory memory to short-term memory in the multi-store memory model. A well-designed attention mechanism ensures that relevant claims are accurately identified and prioritized, enhancing the quality and reliability of the retrieved evidence or information.

Evidence Retrieval. When we receive new information, the brain compares it to what's stored in our long-term memory. It then pulls out the memories or facts that seem to best support the idea being considered. Human intelligence is often measured by how efficiently and effectively we can retrieve and apply this stored knowledge (Liesefeld et al., 2016). Similarly, the effectiveness of an AFC model is directly proportional to the performance of its evidence retrieval engine (Chen et al., 2024; Sundriyal et al., 2022b; Schlichtkrull et al., 2023). Various evidence selection approaches have been discussed as classification (Wadden et al., 2020) and regression problem; also, annotation distillation is used to mimic the annotator distribution (Glockner et al., 2024).

Veracity Prediction. Short-term memory holds the most critical information for processing, including sensory input and retrieved knowledge. Similarly, in AFC models, veracity prediction is performed using claims and the evidence that aligns with them. Both short-term memory and the AFC processing modules function as active centres where alignment, reasoning, comprehension, and classification occur in real-time (Baddeley and Hitch, 1974). Just as short-term memory processes and maintains information for immediate use, the AFC models rely on real-time interactions between claims and supporting evidence, dynamically updating their predictions as new data is processed. This parallel highlights the importance of adaptive, context-aware systems in both human cognition and misinformation detection models.

Justification Production. During veracity prediction, information undergoes processing and is organized to generate justifications. This process parallels the memory encoding process, where information is summarized and structured for effective storage (Panigrahy, 2019). Once encoded, the information can take the form of justifications when shared with users or become part of a knowledge base if stored in memory. Other types of models include relational models such as ProoFVer (Krishna et al., 2021) and MultiVers (Wadden et al., 2022). ProoFVer is based on the natural logic theory of compositional entailment, while MultiVers predicts rationale sentences from the evidence and classifies claim veracity using Longformer (Beltagy et al., 2020) as an encoder. These models are highly explainable; however, they face significant challenges when dealing with the context and pragmatics of claims. Additionally, these models do not effectively capture complex relationships between phrases or sentences, limiting their overall performance in certain scenarios.

In the current era of multimodal content, the scope and complexity of misinformation have expanded significantly (Segura-Bedmar and Alonso-Bartolome, 2022). Misinformation now includes advanced human engineering techniques to manipulate large audiences, making evidence-retrieval-based fact-checking and relational approaches increasingly insufficient for handling its complexities. Furthermore, while foundational, the memory model under study oversimplifies and overlooks phenomena related to emotion, bias, cognition and perception (Bennion et al., 2013). Both models also fail to incorporate human behaviour in response to a claim, instead focusing solely on its content or style. Misinformation is most harmful when it has the potential to influence the masses. Additionally, filtering claims based on their potential to disrupt human behaviour is cost-effective. This strategy would enhance the efficiency of real-time

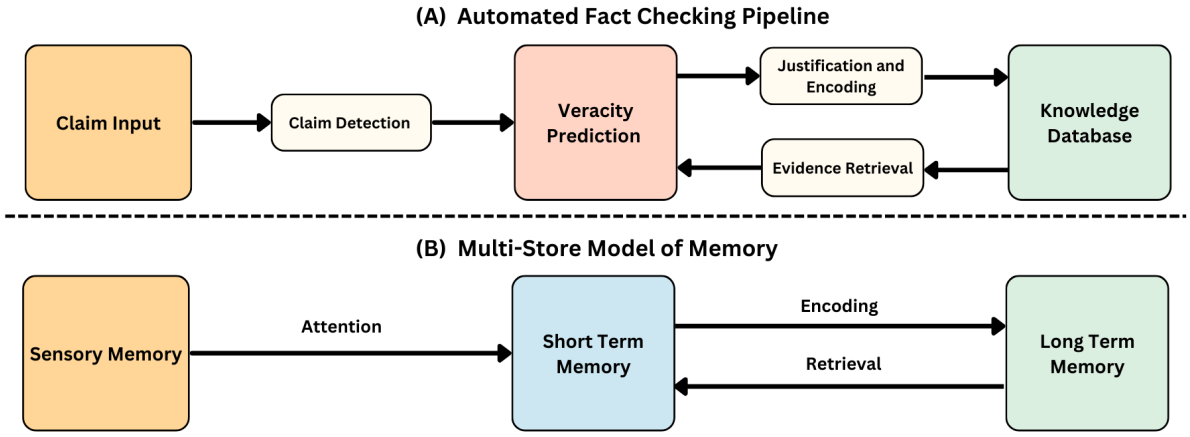

Figure 1: A comparison between (b) Automated Fact Checking Pipeline and (b) Multi-Store Model of Memory, illustrating how misinformation detection parallels human memory processes. The fact-checking pipeline retrieves and verifies claims against a knowledge database, similar to how human cognition processes, stores, and retrieves information.

fact-checking by giving importance to claims that pose large-scale risks. This highlights the need to align our research with recent advancements in psychological studies, providing a more robust framework for understanding and addressing the evolving challenges of misinformation.

In the next section, we will explore how adopting a different point of view can redefine the approach to misinformation, enabling the development of more generalizable, effective, and robust solutions.

3 Understanding the Gestalt of Lies

Wertheimer and Riezler (1984) introduced the concept of a *gestalt* to describe how an organised whole is perceived as greater than the sum of its parts. In the context of misinformation, this principle suggests that the persuasive power of falsehoods often does not lie in isolated claims but in how these claims are emotionally framed, repeated, and reinforced as a coherent narrative. Misinformation is not merely about verifying the veracity of individual claims, nor is it about assessing the truthfulness of their sum total (Starbird et al., 2019). This *gestalt framing* is a core mechanism behind the virality and resilience of misinformation. For instance, during the COVID-19 pandemic, individual claims (e.g., vaccines cause harm, coronavirus is a bioweapon) were either misleading or debunked when evaluated independently. Yet, when arranged into a larger, emotionally charged narrative of institutional betrayal, the gestalt effect overrides the individual debunkings. Viewers were persuaded not by any one fact, but by the cumulative structure that appears internally consistent and emotionally resonant. This further introduces the concepts of perception and *'qualia'*, the subjective, individual experience of interpreting information, into the traditional approach of verifying individual claims. Qualia, in this sense, refer to how each person internally experiences a claim or narrative, colored by prior beliefs, emotions, and contextual understanding (Chalmers, 1995). These subjective experiences can shape whether a piece of misinformation is accepted or rejected, regardless of its factual accuracy.

The gestalt nature of misinformation has been a major bottleneck in establishing a clear definition. Cognitive psychology provides models such as the multi-store model of memory (Atkinson and Shiffrin, 1968), which has been compared to traditional fact-checking models. However, these cognitive models fail to capture the broader gestalt of misinformation. Current models try to capture this gestalt in fragmented ways, addressing various aspects such as claim detection (Gupta et al., 2021; Sundriyal et al., 2021, 2022a), which focuses on identifying claims within a piece of content, yet often lacks the capacity to evaluate the broader context in which these claims are made. Claim simplifications aim to break down complex assertions into simpler components (Sundriyal et al., 2023b; Mittal et al., 2023), making it easier to assess their validity. Claim matching involves aligning detected claims with existing verified information (Kazemi et al., 2021a). Check-worthiness evaluates whether a claim is worth verifying (Sundriyal et al., 2023a),

considering factors like relevance and potential impact. Evidence retrieval and verification are central in sourcing relevant data to support or refute claims (Sundriyal et al., 2022b; Thorne et al., 2018). These elements, while essential to misinformation detection, often function in isolation, limiting their ability to work together in a cohesive framework.

This gap emphasises the need for misinformation interventions that go beyond factual correctness. Detection systems must account for emotional framing, contextual cohesion, and the subjective experience (or *qualia*) of the information consumer. What matters is not just what is said, but how it is arranged, perceived, and interpreted collectively.

4 Forming and Revising Beliefs

In the battle against misinformation, understanding how humans *form* and *revise* their beliefs is crucial. Factual claims alone do not determine belief – how individuals interpret and internalise these claims plays a critical role. Thus, exploring both the structure of claims and the cognitive processes involved in belief formation provides a fuller picture of misinformation's impact.

Forming Beliefs. Two types of attitude change can be related to perceptions of the veracity of claims: incongruent change and congruent change (McGuire, 1969). Incongruent change occurs when a claim first believed to be 'True' is altered to 'False' or vice versa. Congruent change, on the other hand, involves increasing confidence in an existing label. Mathematically, these changes are typically modeled using a general loss function.

According to the psychological literature, incongruent changes are more challenging for humans than congruent changes (McGuire, 1969). Ranadive et al. (2023) highlighted that class-selective neurons tend to emerge within the initial few training epochs. This observation contrasts with the findings of human-based experiments, suggesting that achieving class label changes would require significantly more effort and extended training. A key limitation of this study is its focus on ResNet-50s trained on ImageNet instead of text-based data; extending the analysis to pretrained language models could offer deeper insights into class-selective neuron behavior across contexts. One possible explanation is the lack of human biases, stereotypes, and prewiring in these models, which may ease incongruent changes. Further research should include developing datasets to capture these cognitive attributes in language models.

Revising Beliefs. Both external and internal factors influence individuals' susceptibility to misinformation and motives to disseminate it (Sindermann et al., 2021). External factors relate to features of the information environment and social networks, while internal factors include personal traits and cognitive biases. Many meta-analyses have been conducted to identify the key psychological factors involved in the spread of misinformation (Munusamy et al., 2024; Nan et al., 2022; Sultan et al., 2024). Psychological concepts related to misinformation are detailed in Appendix A.1.

As implied by the mere-exposure effect (Zajonc, 1968), individuals may develop positive attitudes toward information they encounter repeatedly. This contributes to a true-news bias (Sultan et al., 2024). According to confirmation bias theory, people tend to place more trust in information that aligns with their existing beliefs (Klayman, 1995). Studies further show that people who already hold a misperception are more likely to accept misinformation that confirms it, intensifying polarization in public discourse (which aligns with congruent change discussed above) (Zhou and Shen, 2022). The heuristic-systematic model (HSM) posits that people may process information in a heuristic way, which is nonanalytic (Todorov et al., 2002) and makes people more easily to share misinformation (Sun and Xie, 2024b).

5 Cognitive and Social Complexities

Several fact-checking models have sought to break traditional barriers by incorporating human perspectives into their frameworks in recent years. Chen et al. (2024) highlights the challenges of evidence retrieval in real-world scenarios and emphasizes the need for a human-in-the-loop fact-checking system. This represents a larger trend in misinformation research, from rigorous, fact-based paradigms to more abstract, psychologically informed approaches. Next, we will look at significant psycho-social components that have been added into recent initiatives to counteract misinformation. A more temporal perspective can be found in Appendix A.2.

Social Energy and Perspectives. LLMs are limited in their ability to be used directly off-the-shelf for judging the veracity of news articles, where fac-

tual accuracy is essential. To alleviate this issue, Wan et al. (2024) proposed crucial steps in misinformation detection where LLMs may be introduced into the pipeline. Their model generates diverse reactions by leveraging varied user attributes and creates a user-news network using prompt-based techniques. To simulate a potential misinformation propagation process, three distinct strategies are implemented for LLMs: (i) generate a comment based on the news article, (ii) generate a comment in response to an existing comment, and (iii) select a comment for further engagement. The resulting network embodies an artificial social perspective (Song, 2021), which serves as the foundation for performing various tasks using GNNs. This approach integrates user interactions to enhance the model's ability to analyse and interpret complex patterns in misinformation detection. In AFaCTA (Ni et al., 2024), debates between LLM agents are used as a mechanism to improve factuality assessments. During these debates, the agents engage in back-and-forth discussions, constructing reasoned arguments that explore different perspectives on a claim. This collaborative process, when combined with step-by-step fact extraction, leads to more accurate and reliable performance of the fact-checking system.

Despite their innovation, these models face psycho-social limitations, notably the lack of interpretability in LLMs, which hinders verification of their real-world reliability. Such outputs may emerge from hallucinations (Guan et al., 2024) or biases (Kumar et al., 2024), affecting the model's robustness. Ensuring reliability requires metrics to assess the appropriateness and logic of these responses, improving overall stability and performance. Additionally, the scarcity of open-source LLMs raises concerns about behavioral variability, as training data differences can influence the generation of user perspectives. Prompt design also carries human biases, potentially limiting the range of outcomes. Addressing these issues requires extensive research using real-world datasets and focusing on the interpretability of models (Elhage et al., 2021). Overall, the research shows us a way to simulate a misinformation propagation framework, but the data used is questionable in terms of its reliability for combating real-world misinformation.

Heuristics, Bias, and Challenge of Entity Ambiguity in LLMs. Chiang and Lee (2024) highlighted a critical issue in LLMs, where information about multiple entities is merged within the same biography, misleading users who lack prior knowledge. While the proposed D-FActScore metric seeks to address this by evaluating factuality in the presence of entity ambiguity (Min et al., 2023; Chiang and Lee, 2024), it treats the problem primarily as a surface-level inconsistency rather than a deeper cognitive phenomenon. As Kahneman et al. (1982) noted, cognitive biases influence how individuals interpret and integrate information, suggesting that entity ambiguity is not merely a technical glitch but a reflection of heuristics and mental shortcuts users employ. For example, a politically biased reader might mix up 'George Soros' with unrelated people in a conspiracy article due to repeated exposure to partisan misinformation, showing how existing beliefs override facts that don't fit. In such cases, ambiguity fosters a gestalt of lies, where pieces from multiple biographies blend into a coherent but false narrative. The resulting plausibility is strengthened by the qualia of familiarity and coherence, which makes misinformation feel naturally true even when it is factually wrong. To address these deeper cognitive factors, future research should focus on improving evidence-retrieval models by developing systems that incorporate cognitive heuristics. These systems would go beyond simple search functionalities, actively mitigating biases and improving the reliability of information retrieval (Chaiken et al., 1989).

Future studies could also explore the role of false cognates, words that look similar across languages but have different meanings, in spreading misinformation and hate. This area of research holds significant potential for addressing a unique and impactful dimension of the misinformation problem.

LLMs' Belief Towards Misinformation via Persuasive Conversation. (Xu et al., 2024) demonstrated that LLMs can shift correct beliefs when exposed to persuasive misinformation, a pattern that mirrors human susceptibility to persuasion. According to cognitive dissonance theory (Festinger, 1957), when individuals hold conflicting cognitions such as knowing a fact while simultaneously encountering persuasive counterclaims, the resulting psychological discomfort motivates them to reduce the conflict by adjusting their attitudes or beliefs. The behaviour of LLMs under persuasion resem-

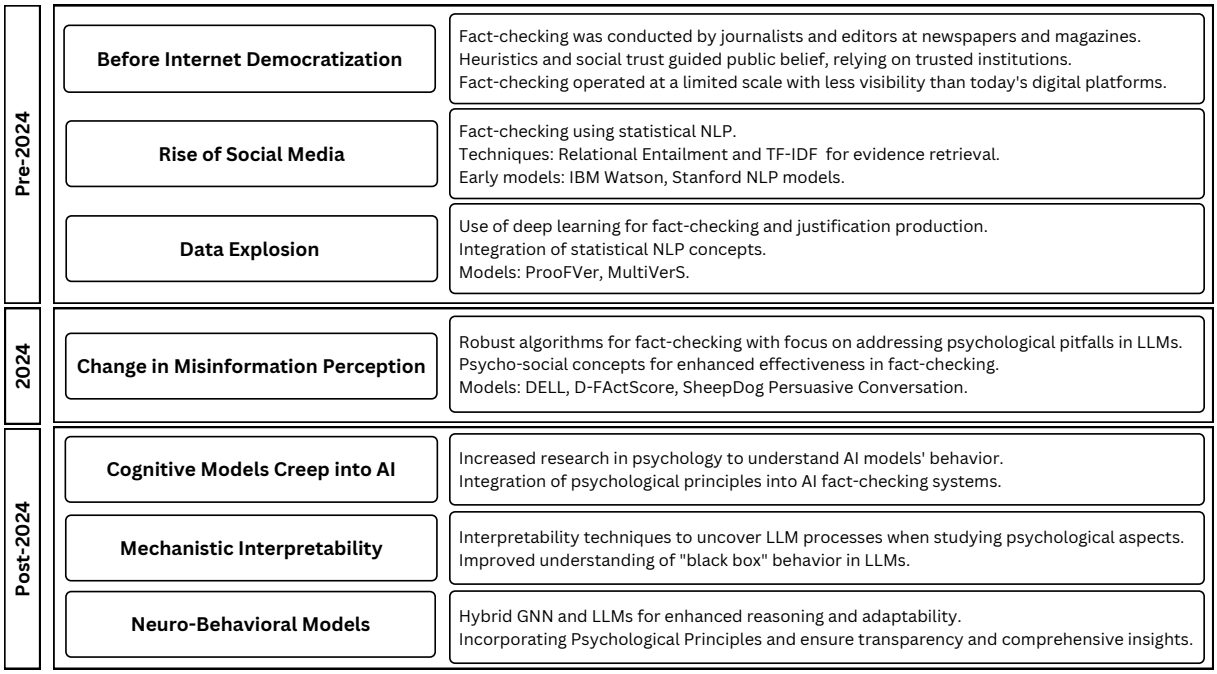

Figure 2: The timeline of the evolution of misinformation research, demonstrating significant advances in fact-checking approaches from earlier times (pre-2024), present (2024), and projected future (post-2024).

bles this dissonance-reduction process since rather than maintaining an initially accurate belief, the models align their outputs with persuasive inputs in a way similar to how humans align beliefs to restore internal coherence. Psychological studies further highlight how persuasion exploits cognitive dissonance by creating internal conflict and presenting belief change as a resolution (Crano and Prislin, 2006; Gass and Seiter, 2018). Thus, the observed susceptibility of LLMs under persuasive influence parallels well-established human tendencies, reinforcing the analogy between model behaviour and cognitive dissonance.

Another significant finding is that most LLMs are susceptible to persuasive misinformation, particularly when it aligns with their prior knowledge or training data. This susceptibility mirrors confirmation bias in humans (Nickerson, 1998). Additionally, the repetition strategy significantly increases the misinformation rate across most models. Repetition acts as a heuristic, making misinformation appear more credible without requiring systematic processing, aligning with the Heuristic-Systematic model theory (Chaiken et al., 1989). In persuasive settings, consecutive misleading statements can form a gestalt of lies, where each claim builds on the previous to create a coherent narrative. This cumulative flow makes the argument seem more convincing than any single claim alone (Xu et al., 2024). Furthermore, LLMs exhibit sycophancy, corresponding to the theory of Attitude-Behaviour Consistency (Wicker, 1969). While the literature uncovers valuable insights, it also highlights limitations. In humans, beliefs are shaped by more complex cognitive processes, including emotional investment, experiential memory, and subconscious biases. These deeper layers of human cognition are not addressed in this study.

Humans often resist persuasion through strategies such as counter-arguing, source skepticism, or reliance on prior knowledge (Brehm and Brehm, 2015). Incorporating these resistance mechanisms into LLMs could help develop more robust strategies to counter misinformation effectively.

Style-based Stereotypes. Wu et al. (2024) highlighted style manipulation using LLMs as a significant challenge to misinformation detection. It reveals that fake news camouflaged with LLM-generated styles substantially reduces state-of-the-art text-based detectors' effectiveness. Cognitive bias, a mental shortcut that aids quick situational analysis, is reflected in these models, exhibiting stereotypes toward certain styles when predicting veracity (Kahneman et al., 1982). Reframed news from trusted publishers leverages their credibility as a tool for deception, where credibility of the message source strongly influences compliance and belief (Pornpitakpan, 2004). Humans evaluate

persuasive content using either the central route, which relies on logical reasoning, or the peripheral route, focusing on stylistic cues (Petty and Cacioppo, 1986). Stylistic manipulation takes advantage of the peripheral route, bypassing critical analysis and enhancing the effectiveness of misinformation detection. Style cues can invoke the qualia of trust, similar to the feeling readers associate with reliable sources. This similarity makes misinformation seem credible by evoking the same subjective experience of trustworthiness.

However, several areas require further research. Emotional influence, combined with stylistic factors, could provide deeper insights. Investigating individual differences in perceiving styles as trustworthy could lead to developing robust and adaptable models. Additionally, the current evaluation framework focuses only on individual interactions with content, overlooking the social amplification of misinformation. Comparing and contrasting the effects of style and social influence could offer valuable insights into how these factors collectively and interactively shape belief in misinformation (Song, 2021).

6 Future Scope and Directions

The findings discussed in this work provide a foundational basis for future research in the field of misinformation, as illustrated in Figure 2. However, several areas still warrant further exploration. This section outlines potential future research directions, focusing on expanding current methodologies and exploring novel approaches.

Datasets Unlocking Psychological Biases. Recent datasets such as FARM (Xu et al., 2024) and AmbigBio (Chiang and Lee, 2024), fall short in addressing the complexity of multiple psychological biases simultaneously. Given that psychological biases, emotions, and perceptions are intricately linked and context-dependent, there is a clear need for datasets that better account for these intertwined factors, particularly in the context of misinformation. Creating such datasets will require extensive text annotations, necessitating collaboration between experts in linguistics and psychology. These datasets could enable the training of advanced models, including LLMs, to recognize (Lin et al., 2024) and address these biases effectively. However, the impact of training LLMs on such datasets on their downstream task performance remains uncertain. Investigating this aspect could provide valuable insights into the development of artificial general intelligence. This might also facilitate the downgrading of certain capabilities or offer moral and emotional reasoning in LLMs, ensuring their efficiency and making them safer to deploy.

Cognitive Modules for LLMs. Mechanistic interpretability (Elhage et al., 2021) and techniques like LoRA (Low Rank Adaptation) and Adapters offer a promising avenue for progress by modeling weight changes and creating specific role-based modules for LLMs. This could help integrate psychological modules into LLMs without the need for additional training or fine-tuning, thereby reducing the repetitive reliance on human labour and saving significant time and compute. At the same time, LLMs remain statistical models that are prone to misgeneralization and shortcut learning. The development of datasets that capture inherent psychological factors—particularly those involving human participants—would be crucial for enabling models to reflect nuanced cognitive phenomena more faithfully. Mechanistic interpretability techniques are still in their early stages, and the integration of cognitive modules into LLMs remains a speculative yet promising direction. These modules are not limited to fact-checking but also can be used for various other applications like reasoning, cognitive neuroscience, world models, and many more.

Recent studies have demonstrated that mechanistic interpretability can be applied effectively to investigate social biases. For example, studies (Yu and Ananiadou, 2025; Chandna et al., 2025; Vig et al., 2020) demonstrate that gender bias can be analysed in a controlled and experimentally tractable manner, such as by systematically swapping terms like “man” and “woman” in prompts. In contrast, the identification and modelling of cognitive biases remain less tractable, owing both to the scarcity of relevant datasets and to the inherent subjectivity of human cognition. While the current body of work highlights important progress, it also underscores that practical development of cognitive modules is still far from realization. A possible future direction is to train adapters on cognitive-bias datasets, once such resources become available, thereby creating reusable cognitive modules that extend beyond fact-checking to applications in reasoning, cognitive neuroscience, and world modelling. Such advancements would pave the way for more nuanced, ethical, and effective AI systems.

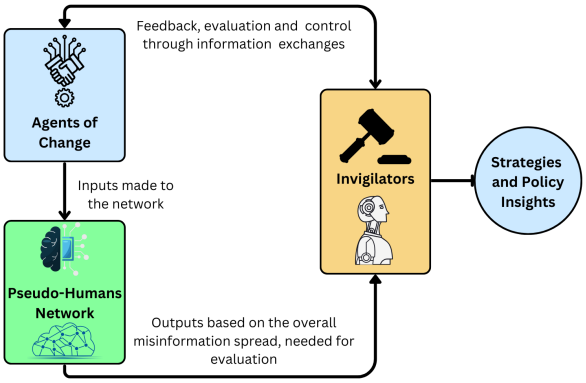

Figure 3: An illustration of the neuro-behavioural framework, showing the interaction between the three main components: agents of change, pseudo-humans network, and invigilators.

Neuro-Behavioural Models. World models have recently gained significant attention due to their ability to simulate the interaction between an agent and its environment (Ha and Schmidhuber, 2018). These models create a latent representation of the world, enabling them to predict environmental dynamics by integrating perception, memory, and decision-making. By simulating scenarios internally, world models allow agents to evaluate potential actions without direct interaction, making them effective for planning and problem-solving. Multi-agent systems are essential for addressing social problems, as they interact not only with the environment but also with one another (Guo et al., 2024). Neuro-Behavioural Models conceptualize social interactions within a simulated world. These models feature three primary types of agents, as illustrated in Figure 3: ▷ **Pseudo-humans**: Large models simulating human attributes, including perception, biases, and cognitive frameworks, with varying bias proportions for diversity. ▷ **Agents of Change**: Models interacting with pseudo-humans, providing inputs, analyzing outputs, and simulating scenarios. ▷ **Invigilators**: Models that continuously evaluate the network and provide feedback to the Agents of Change, enabling dynamic input adjustments. After analysis, they assist in developing new policies and strategies.

This high-level framework requires advanced techniques, such as GNN, Reinforcement Learning, and their integration with LLMs. Its design is further supported by the rapidly growing paradigm of LLMs as autonomous agents. Recent systems such as Auto-GPT (Significant Gravitas), BabyAGI (Nakajima), and Generative Agents (Park et al., 2023) demonstrate how LLM-based multi-agent systems can collaborate to perform diverse tasks. Similarly, ReAct (Yao et al., 2023) and HuggingGPT (Shen et al., 2023) illustrate the integration of reasoning and tool use within agentic workflows, while frameworks like ChatDev (Qian et al., 2024) and MetaGPT (Hong et al., 2024) show how specialised roles can be distributed among agents to collectively solve complex problems. These developments provide concrete demonstrations of feasibility and suggest that Neuro-Behavioural Models can emerge as a viable research direction. Advances in world models could further enable Neuro-Behavioural Models to better simulate and address these complex social challenges.

7 Conclusion

Through this survey, we aim to examine how misinformation research has evolved from a structural perspective, such as verifying multiple facts within a claim, to a holistic perspective, where augmented biases in datasets and psychological phenomena are integrated into misinformation detection frameworks. Misinformation has never been merely a matter of factual inaccuracies. Misinformation is instead a psychological and sociological issue that exploits human perception and reaction. This survey highlights the shift in misinformation research from factuality-centric approaches to cognitively grounded frameworks. We emphasize that combating misinformation requires more than detecting falsehoods; it also requires understanding belief. As misinformation narratives prioritise coherence, repetition, and emotional appeal over factual correctness, detection systems must progress beyond claim-level assessment. Future models should treat misinformation as a narrative, not just isolated claims. Neuro-behavioural simulations, psychological databases, and cognitive modules all present promising avenues. Integrating social cognition, and human-in-the-loop evaluations is no longer optional; it is essential for developing robust, adaptive, and trustworthy AI systems in the age of misinformation. Equally important is bridging mechanistic interpretability with behavioural insights to explain why models fail under persuasive or stylistic manipulation. Advancing in this direction could enable proactive interventions, where AI not only detects misinformation but anticipates and counters its psychological influence.

Limitations

The limitations in this survey can be summarized in the following points:

1. This survey focused exclusively on English fact-checking pipelines and associated cognitive phenomena. Exploring multilingual fact-checking and the variation of cognitive biases across languages and cultures remains a valuable direction for future research.

2. This survey does not include a taxonomy-based classification, as it aims to bridge two distinct domains—cognition and computation—in the context of fact-checking and claim veracity. Developing a meaningful taxonomy in this area requires further research and time, as continued experimentation across models is needed to establish proper classifications.

3. This survey primarily focuses on research published after 2021 to reflect recent advances in computational techniques for fact-checking.

4. While this study focuses on theoretical psychological concepts. Incorporating application-based findings, such as those from behavioral interventions and relating them to misinformation, presents a promising route for future research.

Acknowledgement

T. Chakraborty acknowledges the support of the Anusandhan National Research Foundation (DST/INT/USA/NSF-DST/Tanmoy/P-2/2024) and Rajiv Khemani Young Faculty Chair Professorship in Artificial Intelligence.

References

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022b. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643.

Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. 2023. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*.

Marco Amoruso, Daniele Anello, Vincenzo Auletta, Raffaele Cerulli, Diodato Ferraioli, and Andrea Raiconi. 2020. Contrasting the spread of misinformation in online social networks. *Journal of Artificial Intelligence Research*, 69:847–879.

RC Atkinson and RM Shiffrin. 1968. Human memory: A proposed system and its control processes (vol. 2). *The Psychology of Learning and Motivation: Advances in Research and Theory*.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, and 1 others. 2023. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*.

Alan D. Baddeley and Graham Hitch. 1974. Working memory. volume 8 of *Psychology of Learning and Motivation*, pages 47–89. Academic Press.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Kelly Bennion, Jaclyn Ford, Brendan Murray, and Elizabeth Kensinger. 2013. Oversimplification in the study of emotional memory. *JINS*, 19:1–9.

Nadav Borenstein, Greta Warren, Desmond Elliott, and Isabelle Augenstein. 2025. Can community notes replace professional fact-checkers? *CoRR*.

Jens B. H. Brehm and Jack W. Brehm. 2015. Strategies and motives for resistance to persuasion: An integrative framework. *Frontiers in Psychology*, 6:1201.

Marcus Butavicius, Kathryn Parsons, Malcolm Pattinson, and Agata McCormac. 2016. Breaching the human firewall: Social engineering in phishing and spear-phishing emails. *Preprint*, arXiv:1606.00887.

Shelly Chaiken, Akiva Liberman, and Alice H. Eagly. 1989. Heuristic and systematic information processing within and beyond the persuasion context. In James S. Uleman and John A. Bargh, editors, *Unintended Thought*, pages 212–252. Guilford Press.

David Chalmers. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–19.

Bhavik Chandna, Zubair Bashir, and Procheta Sen. 2025. Dissecting bias in llms: A mechanistic interpretability perspective. *Preprint*, arXiv:2506.05166.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proc of NAACL:HLT*, pages 3569–3587, Mexico City, Mexico. ACL.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proc of EMNLP*, pages 3495–3516, Abu Dhabi, United Arab Emirates. ACL.

Cheng-Han Chiang and Hung-yi Lee. 2024. Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations. *arXiv preprint arXiv:2402.05629*.

William Crano and Radmila Prislin. 2006. Attitudes and persuasion. *Annual review of psychology*, 57:345–74.

Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. 2021. Edited media understanding frames: Reasoning about the intent and implications of visual misinformation. In *Proc of ACL-IJCNLP*, pages 2026–2039, Online. ACL.

Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. 2024. Document-level claim extraction and decontextualisation for fact-checking. In *Proc of ACL*, pages 11943–11954, Bangkok, Thailand. ACL.

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):13–29.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact-checking: A survey on architectures and approaches. *arXiv preprint arXiv:2407.12853*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Martin Fajcik, Petr Motlicek, and Pavel Smrz. 2023. Claim-dissector: An interpretable fact-checking system with joint re-ranking and veracity prediction. In *Findings of ACL*, pages 10184–10205, Toronto, Canada. ACL.

Lisa K. Fazio, Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh. 2015. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5):993–1002.

Leon Festinger. 1957. *A Theory of Cognitive Dissonance*. Stanford University Press.

Robert Gass and John Seiter. 2018. *Persuasion: Social Influence and Compliance Gaining*.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proc of EMNLP*, pages 5916–5936, Abu Dhabi, United Arab Emirates. ACL.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *TACL*, 12:1–18.

Denise Grady and Patricia Mazzei. 2021. News: Doctor's death after covid vaccine is being investigated (the new york times) - behind the headlines - nlm.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111, Mexico City, Mexico. Association for Computational Linguistics.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *TACL*, 10:178–206.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proc of ACL*, pages 3785–3801, Dublin, Ireland. ACL.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Lesa: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188.

David Ha and Jürgen Schmidhuber. 2018. World models.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle,

United States. Association for Computational Linguistics.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. *Preprint*, arXiv:2308.00352.

Irving L. Janis. 1972. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin, Boston.

Shan Jiang and Christo Wilson. 2021. Structurizing misinformation stories via rationalizing fact-checks. In *Proc of ACL-IJCNLP*, pages 617–631, Online. ACL.

Marcia K. Johnson, Sheila Hashtroudi, and Daniel S. Lindsay. 1993. Source monitoring. *Psychological Bulletin*, 114(1):3–28.

D. Kahneman, P. Slovic, and A. Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021a. Claim matching beyond english to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021b. Claim matching beyond English to scale global fact-checking. In *Proc of ACL-IJCNLP*, pages 4504–4517, Online. ACL.

Evelyn Fox Keller. 2010. *The Mirage of a Space Between Nature and Nurture*. Duke University Press.

Joshua Klayman. 1995. Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418.

Ziyi Kou, Lanyu Shang, Yang Zhang, Zhenrui Yue, Huimin Zeng, and Dong Wang. 2022. Crowd, expert & ai: A human-ai interactive approach towards natural language explanation-based covid-19 misinformation detection. In *Proc of IJCAI*, pages 5087–5093.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2021. Proofver: Natural logic theorem proving for fact verification. *arXiv preprint arXiv:2108.11357*.

Divyanshu Kumar, Umang Jain, Sahil Agarwal, and Prashanth Harshangi. 2024. Investigating implicit bias in large language models: A large-scale study of over 50 llms. *Preprint*, arXiv:2410.12864.

H. R. Liesefeld, E. Hoffmann, and D. Wentura. 2016. Intelligence as the efficiency of cue-driven retrieval from secondary memory. *Memory*, 24(3):285–294.

Shuya Lin, Yuxiong Wang, Jonathan Dong, and Shiguang Ni. 2024. Detection and positive reconstruction of cognitive distortion sentences: Mandarin dataset and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6686–6701, Bangkok, Thailand. Association for Computational Linguistics.

Hui Liu, Wenya Wang, and Haoliang Li. 2023. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of ACL*, pages 9781–9796, Toronto, Canada. ACL.

Yanchen Liu, Mingyu Derek Ma, Wenna Qin, Azure Zhou, Jiaao Chen, Weiyan Shi, Wei Wang, and Diyi Yang. 2024. Decoding susceptibility: Modeling misbelief to misinformation through a computational approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15178–15194.

Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2024. Blessing or curse? a survey on the impact of generative ai on fake news. *arXiv e-prints*, pages arXiv–2404.

Chu Fei Luo, Radin Shayanfar, Rohan V Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2024. Misinformation with legal consequences (MisLC): A new task towards harnessing societal harm of misinformation. In *Findings of EMNLP*, pages 15749–15768, Miami, Florida, USA. ACL.

Cameron Martel, Gordon Pennycook, and David G. Rand. 2020. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1):47.

William J. McGuire. 1969. The nature of attitudes and attitude change. In Elliot Aronson and Gardner Lindzey, editors, *The Handbook of Social Psychology*, 2nd edition, volume 3, pages 136–314. Addison-Wesley, Massachusetts.

Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In *Proc of ACL*, pages 15817–15835, Toronto, Canada. ACL.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proc of EMNLP*, pages 12076–12100, Singapore. ACL.

Shubham Mittal, Megha Sundriyal, and Preslav Nakov. 2023. Lost in translation, found in spans: Identifying claims in multilingual social media. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3887–3902.

Shalini Munusamy, Kalaivanan Syasyila, Azahah Abu Hassan Shaari, Muhammad Adnan Pitchan, Mohammad Rahim Kamaluddin, and Ratna Jatnika.

2024. Psychological factors contributing to the creation and dissemination of fake news among social media users: a systematic review. *BMC psychology*, 12(1):673.

Yohei Nakajima. Babyagi. GitHub repository.

Preslav Nakov, Jisun An, Haewoon Kwak, Muhammad Arslan Manzoor, Zain Muhammad Mujahid, and Husrev Taha Sencar. 2024. A survey on predicting the factuality and the bias of news media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15947–15962, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoli Nan, Yuan Wang, and Kathryn Thier. 2022. Why do people believe health misinformation and who is at risk? a systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine*, 314:115398.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1890–1912, Bangkok, Thailand. Association for Computational Linguistics.

Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proc of EMNLP*, pages 2532–2544, Abu Dhabi, United Arab Emirates. ACL.

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 *Proc of ACL*, pages 6981–7004, Toronto, Canada. ACL.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Rina Panigrahy. 2019. How does the mind store information? *Preprint*, arXiv:1910.06718.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 21–29. Springer.

Richard Petty and John Cacioppo. 1986. The elaboration likelihood model of persuasion. *Advances in hydroscience*, 19:124–205.

Richard E. Petty and Jon A. Krosnick, editors. 1995. *Attitude Strength: Antecedents and Consequences*. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Nele Põldvere, Zia Uddin, and Aleena Thomas. 2023. The politifact-oslo corpus: A new dataset for fake news analysis and detection. *Information*.

Chanthika Pornpitakpan. 2004. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, 34:243 – 281.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Omkar Ranadive, Nikhil Thakurdesai, Ari S Morcos, Matthew Leavitt, and Stéphane Deny. 2023. On the special role of class-selective neurons in early training. *Preprint*, arXiv:2305.17409.

Sougata Saha and Rohini Srihari. 2024. Integrating argumentation and hate-speech-based techniques for countering misinformation. In *Proc of EMNLP*, pages 11109–11124, Miami, Florida, USA. ACL.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

Michael Sejr Schlichtkrull, Vladimir Karpukhin, Barlas Oguz, Mike Lewis, Wen-tau Yih, and Sebastian Riedel. 2021. Joint verification and reranking for open fact checking over tables. In *Proc of ACL-IJCNLP*, pages 6787–6799, Online. ACL.

Isabel Segura-Bedmar and Santiago Alonso-Bartolome. 2022. Multimodal fake news detection. *Information*, 13(6).

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in

hugging face. In *Advances in Neural Information Processing Systems*, volume 36, pages 38154–38180. Curran Associates, Inc.

Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. 2021. Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims. In *Proc of ACL-IJCNLP*, pages 5468–5481, Online. ACL.

Jiasheng Si, Yibo Zhao, Yingjie Zhu, Haiyang Zhu, Wenpeng Lu, and Deyu Zhou. 2024. CHECKWHY: Causal fact verification via argument structure. In *Proc of ACL*, pages 15636–15659.

Significant Gravitas. AutoGPT.

C. Sindermann, H. S. Schmitt, D. Rozgonjuk, J. D. Elhai, and C. Montag. 2021. The evaluation of fake and true news: on the role of intelligence, personality, interpersonal trust, ideological attitudes, and news consumption. *Heliyon*, 7(3):e06503.

Y. Song. 2021. Social energy and trust. *International Journal of Social Science Studies*, 9(1):1–10.

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proc of HCI*, 3:1–26.

Mubashir Sultan, Alan N Tump, Nina Ehmann, Philipp Lorenz-Spreen, Ralph Hertwig, Anton Gollwitzer, and Ralf HJM Kurvers. 2024. Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proc of National Academy of Sciences*, 121(47):e2409329121.

Y. Sun and J. Xie. 2024a. Do heuristic cues affect misinformation sharing? evidence from a meta-analysis. *Journalism & Mass Communication Quarterly*, 0(0).

Yanqing Sun and Juan Xie. 2024b. Do heuristic cues affect misinformation sharing? evidence from a meta-analysis. *Journalism & Mass Communication Quarterly*, page 10776990241284597.

Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2023a. Leveraging social discourse to measure check-worthiness of claims for fact-checking. *arXiv preprint arXiv:2309.09274*.

Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2025. Leveraging rationality labels for explainable claim check-worthiness. *IEEE Transactions on Artificial Intelligence*.

Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. 2023b. From chaos to clarity: Claim normalization to empower fact-checking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6594–6609.

Megha Sundriyal, Harshit Choudhary, Tanmoy Chakraborty, and Md Shad Akhtar. 2024. Crowd intelligence for early misinformation prediction on social media. *arXiv preprint arXiv:2408.04463*.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022a. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proc of EMNLP*, pages 7701–7715, Abu Dhabi, United Arab Emirates. ACL.

Megha Sundriyal, Ganeshan Malhotra, Md Shad Akhtar, Shubhashis Sengupta, Andrew Fano, and Tanmoy Chakraborty. 2022b. Document retrieval and claim verification to mitigate covid-19 misinformation. In *Proc of CONSTRAINT*, pages 66–74.

Megha Sundriyal, Parantak Singh, Md Shad Akhtar, Shubhashis Sengupta, and Tanmoy Chakraborty. 2021. Desyr: definition and syntactic representation based claim detection on the web. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1764–1773.

C. Sunstein. 2002. The law of group polarization. *Journal of Political Philosophy*, 10:175 – 195.

Charles Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science - AMER J POLIT SCI*, 50:755–769.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. Relic: Retrieving evidence for literary claims. *Preprint*, arXiv:2203.10053.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proc of NAACL:HLT*.

Alexander Todorov, Shelly Chaiken, and Marlone D Henderson. 2002. The heuristic-systematic model of social information processing. *The persuasion handbook: Developments in theory and practice*, 23:195–211.

John C. Turner, Michael A. Hogg, Penelope J. Oakes, Stephen D. Reicher, and Margaret S. Wetherell. 1987. *Rediscovering the Social Group: A Self-Categorization Theory*. Basil Blackwell.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Kateryna Tymoshenko and Alessandro Moschitti. 2021. Strong and light baseline models for fact-checking joint inference. In *Findings of ACL-IJCNLP*, pages 4824–4830, Online. ACL.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proc of EMNLP*, pages 7534–7550, Online. ACL.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of NAACL*, pages 61–76, Seattle, United States. ACL.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. DELL: Generating reactions and explanations for LLM-based misinformation detection. In *Findings of ACL*, pages 2637–2667, Bangkok, Thailand. ACL.

Max Wertheimer and Kurt Riezler. 1984. Gestalt theory. *Social Research*, 51(1/2):305–327.

Allan W. Wicker. 1969. Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, 25(4):41–78.

Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proc of ACL*, pages 2448–2460, Dublin, Ireland. ACL.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proc of KDD*.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proc of ACL*, pages 16259–16303, Bangkok, Thailand. ACL.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Zeping Yu and Sophia Ananiadou. 2025. Understanding and mitigating gender bias in llms via interpretable neuron editing. *Preprint*, arXiv:2501.14457.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. *Preprint*, arXiv:2305.12692.

Robert B Zajonc. 1968. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1.

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

A Appendix

A.1 Psychological Foundations and Misinformation

Understanding the psychological undercurrents that shape belief formation and information processing is essential to the study of misinformation. A list of additional psychological concepts central to the theme of this paper is provided in Table 2. The aim of this table is twofold: to familiarise the reader with abstract yet foundational constructs from cognitive science, and to bridge two traditionally distinct domains – cognitive psychology and computational social science. By grounding computational models in well-established psychological theory, we aim to enhance both the interpretability and effectiveness of misinformation detection systems.

Table 2 expands several such constructs that have proven instrumental in explaining why individuals find misinformation persuasive, why they struggle to abandon false beliefs, and how socio-cognitive mechanisms influence the spread of inaccurate content. One of the most influential of these is Cognitive Dissonance (Festinger, 1957), which refers to the psychological discomfort caused by holding conflicting beliefs. Individuals try to resolve this discomfort by adjusting their attitudes or by filtering information selectively (Taber and Lodge, 2006). This process is closely linked to Confirmation Bias (Nickerson, 1998) – the inclination to favour, interpret, and recall information that supports pre-existing beliefs. Collectively, these cognitive tendencies help explain why misinformation

often remains compelling and is accepted as truth, even when confronted with opposing facts.

The Heuristic-Systematic Model (Chaiken et al., 1989) further expands our understanding of how people process information. It proposes two parallel routes of evaluation: a heuristic mode based on cognitive shortcuts and peripheral cues, and a systematic mode grounded in deliberate, analytical reasoning. Misinformation often thrives in heuristic conditions, exploiting superficial cues such as authority (Butavicius et al., 2016), emotion (Sun and Xie, 2024a; Martel et al., 2020), or repetition (Fazio et al., 2015; Nyhan and Reifler, 2010). Relatedly, Anchoring Bias (Tversky and Kahneman, 1974) describes how initial exposure to a specific claim—regardless of its truth value—can anchor subsequent beliefs, skewing judgment.

The Availability Heuristic (Tversky and Kahneman, 1973) provides insight into how repeated exposure or vivid examples (Johnson et al., 1993) can distort our perception of truth, as people tend to judge the likelihood or credibility of events based on how easily instances come to mind. Similarly, the Framing Effect (Tversky and Kahneman, 1981) demonstrates how the same information, when presented differently (e.g., as a gain or a loss), can significantly alter decision-making and belief acceptance. Misinformation often leverages emotional framing to manipulate these biases.

At the group level, Groupthink (Janis, 1972) highlights the risks of conformity and suppressed dissent in tightly knit or ideologically homogeneous communities. It explains how group cohesion can impair critical evaluation and accelerate the unchecked dissemination of misinformation (Turner et al., 1987; Sunstein, 2002). Additionally, Persuasion Theory (McGuire, 1969) and the principle of Attitude-Behavior Consistency (Petty and Krosnick, 1995) emphasize the role of effective communication and the intensity, accessibility, and situational relevance of attitudes in predicting behavioural responses to misinformation.

Together, these psychological constructs form a theoretical backbone for understanding the psychological vulnerabilities exploited by misinformation. Their inclusion in computational frameworks not only improves model performance but also strengthens the interpretability and societal relevance of misinformation detection systems. This integration captures the core aim of this paper, to harmonize algorithmic detection methods with human psychological patterns, fostering interventions against misinformation that are both psychologically insightful and ethically responsible.

A.2 Overview of Recent Literature

Table 3 presents a comprehensive list of recent studies on misinformation and fact-checking, examined from a psychological lens. This compilation not only identifies the psychological phenomena – either explicitly studied or implicitly embedded – in these works but also offers a temporal perspective, highlighting how psychological framing has gained prominence in more recent studies compared to earlier efforts. As such, the table serves as a valuable compass for researchers aiming to explore the evolving intersection of psychology and computational misinformation research.

A clear pattern emerges from this landscape, that while nearly all of the surveyed works concentrate on core tasks such as misinformation detection, fact-checking, and claim structuring, only a subset actively or inactively incorporates psychological theory to enhance their methodologies or explain user susceptibility. In particular, some of these stand out for their deep integration of foundational psychological constructs, including the Framing Effect, Confirmation Bias, and Cognitive Dissonance (Wu et al., 2024; Si et al., 2024; Xu et al., 2024; Liu et al., 2024). These works draw upon classical frameworks such as the Heuristic-Systematic Model, bridging decision-making psychological phenomena with computational fact-checking models to enrich both understanding and performance.

Other studies venture into less traditionally studied but equally impactful psychological or sociological constructs. For example, (Wan et al., 2024) and (Ni et al., 2024) introduce concepts such as Social Energy and Groupthink, highlighting the cognitive dynamics at the group level that influence belief propagation and acceptance of collective misinformation. This growing focus on social cognition marks a shift from isolated user modelling to more context-aware interactional paradigms.

The application of psychological theory extends further into the detection of multimodal misinformation. (Gupta et al., 2022) and (Da et al., 2021) incorporate the priming effect and the attribute error, respectively, to unravel how different modalities, textual, visual, or combined, shape perception and credibility judgments. Similarly, (Chiang and Lee, 2024) and (Saha and Srihari, 2024) examine availability heuristics and group thinking to account for how cognitive shortcuts and peer influence con-

Psychological Concept	Definition
Cognitive Dissonance (Festinger, 1957)	A psychological discomfort experienced when holding two or more conflicting cognitions, leading individuals to adjust their attitudes or behaviors to reduce inconsistency.
Confirmation Bias (Nickerson, 1998)	The tendency to search for, interpret, and recall information in a way that confirms one's preexisting beliefs or hypotheses.
Persuasion (McGuire, 1969)	The process by which a person's attitudes or behavior are influenced by communication from others, often via reciprocity, authority, or social proof.
Heuristic-Systematic Model (Chaiken et al., 1989)	A model proposing two modes of information processing: heuristic (using mental shortcuts) and systematic (in-depth and analytical), affecting how persuasive messages are judged.
Anchoring Bias (Tversky and Kahneman, 1974)	The tendency to rely too heavily on the first piece of information encountered (the "anchor") when making decisions.
Availability Heuristic (Tversky and Kahneman, 1973)	A mental shortcut where individuals estimate the probability of events based on how easily examples come to mind.
Groupthink (Janis, 1972)	A mode of thinking where desire for consensus in cohesive groups leads to suppression of dissent and poor decision-making.
Framing Effect (Tversky and Kahneman, 1981)	A cognitive bias where individuals' decisions are influenced by the way information is presented, such as emphasizing potential gains or losses.
Attitude-Behavior Consistency (Petty and Krosnick, 1995)	The degree to which a person's attitudes predict their behavior, influenced by attitude strength, accessibility, and context.

Table 2: Key psychological concepts relevant to misinformation and their definitions.

tribute to the spread of misinformation.

However, the survey also reveals an evident disparity: several technically sophisticated studies such as (Pan et al., 2023), (Fajcik et al., 2023), and (Wright et al., 2022)—do not explicitly consider psychological constructs, suggesting a persistent gap between computational efficacy and cognitive realism. This observation underscores the significance of this review, as it highlights the need for more integrative and interdisciplinary approaches that not only optimize detection accuracy but also deepen our understanding of why and how users engage with misinformation.

Taken together, the growing incorporation of psychological theories into misinformation research signals a paradigm shift. In the future, there is strong potential for future studies to integrate cognitive and behavioural principles more deliberately into the development and evaluation of misinformation detection systems, resulting in tools that are not only intelligent but also psychologically informed.

Previous Work	Research Focus	Psy. Study	Psychological Phenomenon
Wu et al. (2024)	Misinformation Detection, Fact-Checking Models	✓	Framing Effect, Cognitive Bias
Xu et al. (2024)	Misinformation Impact, Domain-Specific Techniques	✓	Persuasion, Heuristic-Systematic Model, Confirmation Bias, Cognitive Dissonance, Attitude-Behaviour Consistency
Wan et al. (2024)	Fact-Checking Models, Misinformation Detection	✓	Social Energy
Ni et al. (2024)	Claim Structuring, Misinformation Detection	✓	Social Energy, Groupthink
Chiang and Lee (2024)	Claim Structuring, Misinformation Detection	✓	Cognitive Bias, Availability Heuristics
Saha and Srihari (2024)	Misinformation Detection, Multimodal Techniques	✓	Groupthink, Availability Heuristic
Si et al. (2024)	Fact-Checking Models, Argument Structure Reasoning	✓	Heuristic-Systematic Model, Cognitive Dissonance
Liu et al. (2024)	Misinformation Detection, Motivated Reasoning	✓	Heuristic-Systematic Model, Cognitive Dissonance
Deng et al. (2024)	Claim Structuring, Domain-Specific Techniques	✗	N/A
Luo et al. (2024)	Claim Structuring, Domain-Specific Techniques	✓	Negativity Bias
Pan et al. (2023)	Fact-Checking Systems, Claim Structuring	✗	N/A
Fajcik et al. (2023)	Fact-Checking Systems, Claim Structuring	✗	N/A
Mendes et al. (2023)	Misinformation Detection, Domain-Specific Techniques	✓	Anchoring Bias
Yue et al. (2023)	Domain-Specific Techniques, Fact-Checking Models	✗	N/A
Liu et al. (2023)	Multimodal Techniques, Claim Structuring	✓	Modality Bias
Gupta et al. (2022)	Multimodal Techniques, Fact-Checking Models	✓	Priming Effect
Thai et al. (2022)	Claim Structuring, Multimodal Techniques	✗	N/A
Wright et al. (2022)	Fact-Checking Models, Domain-Specific Techniques	✗	N/A
Ousidhoum et al. (2022)	Claim Structuring, Fact-Checking Models	✗	N/A
Chen et al. (2022)	Claim Structuring, Misinformation Detection	✗	N/A
Glockner et al. (2022)	Misinformation Impact, Multimodal Techniques	✓	Confirmation Bias
Sundriyal et al. (2022a)	Claim Structuring, Fact-Checking Models	✗	N/A
Jiang and Wilson (2021)	Claim Structuring, Misinformation Detection	✗	N/A
Kazemi et al. (2021b)	Domain-Specific Techniques, Claim Structuring	✓	Cultural Bias
Sheng et al. (2021)	Claim Structuring, Fact-Checking Models	✗	N/A
Tymoshenko and Moschitti (2021)	Fact-Checking Systems, Misinformation Detection	✗	N/A
Da et al. (2021)	Multimodal Techniques, Misinformation Impact	✓	Visual Bias, Attribution Error
Schlichtkrull et al. (2021)	Fact-Checking Models, Claim Structuring	✗	N/A

Table 3: Overview of misinformation-related papers categorized by their research focus, with indicators of psychological phenomena studied. Psy. Study indicates whether (✓) or not (✗) the work involves the study of any psychological phenomena.