PERSUASION WITH LARGE LANGUAGE MODELS: A SURVEY

A PREPRINT

Alexander Rogiers*, **Sander Noels, Maarten Buyl, and Tijl De Bie**

AIDA-IDLab, Electronics and Information Systems, Ghent University

ABSTRACT

The rapid rise of Large Language Models (LLMs) has created new disruptive possibilities for persuasive communication, by enabling fully-automated personalized and interactive content generation at an unprecedented scale. In this paper, we survey the research field of *LLM-based persuasion* that has emerged as a result. We begin by exploring the different modes in which *LLM Systems* are used to influence human attitudes and behaviors. In areas such as politics, marketing, public health, e-commerce, and charitable giving, such LLM Systems have already achieved human-level or even super-human persuasiveness. We identify key factors influencing their effectiveness, such as the manner of personalization and whether the content is labelled as AI-generated. We also summarize the experimental designs that have been used to evaluate progress. Our survey suggests that the current and future potential of LLM-based persuasion poses profound ethical and societal risks, including the spread of misinformation, the magnification of biases, and the invasion of privacy. These risks underscore the urgent need for ethical guidelines and updated regulatory frameworks to avoid the widespread deployment of irresponsible and harmful LLM Systems.

1 Introduction

Persuasive communication, defined as "any message that is intended to shape, reinforce, or change the responses of another, or others", has existed both as a practice and as an academic discipline for many decades [41]. Until recently, it has been the prerogative of humans to persuade others, initially through personal or small-group communication, and later through mass-communication channels and technologies such as the printed press, radio, and television.

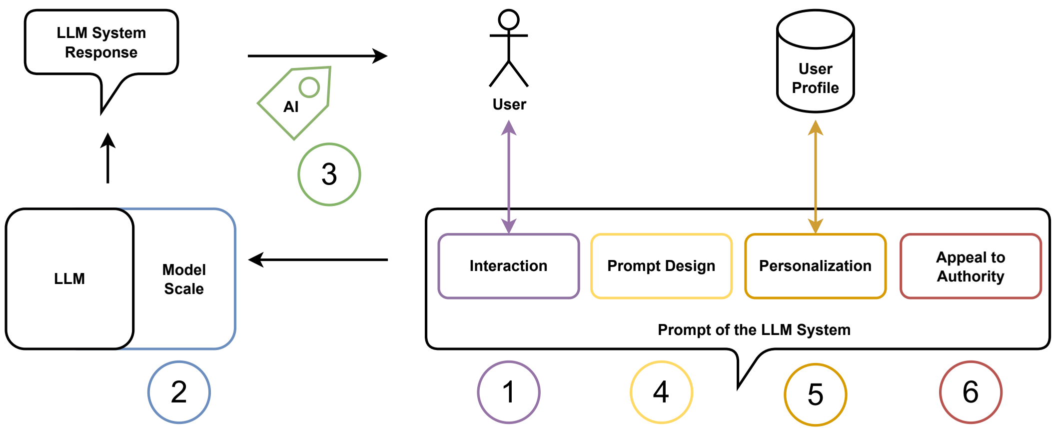

Figure 1: Overview of factors influencing the persuasiveness of an LLM System: (1) whether interactive dialogue is used, (2) the size of the LLM, (3) whether AI authorship is disclosed to users, (4) whether prompts are specifically engineered for persuasion, (5) the use of personal data for personalization, and (6) the use of authoritative language.

*Corresponding author. E-mail: alexander.rogiers@ugent.be

Since the rise of the internet, persuasion has become increasingly personalized, thus enhancing its reach and effectiveness [19, 27, 24]. Whereas earlier mass-communication via printed or audio-visual media could be targeted only to a limited extent to the demographics of a communication channel's audience, the internet allows far more fine-grained personalization, in some applications even down to the individual.

With the advent of Large Language Models (LLMs), marked by the release of ChatGPT in 2022, a new era in persuasive communication begins. In just a few years, LLMs have evolved from basic autoregressive text continuation models to models with remarkable emergent capabilities. We discuss an emergent capability that appears especially powerful and possibly concerning: their potential for intentionally generating human-like content that shapes attitudes, informs opinions, and drives behaviors across various domains. This evolution enables a significant change in the landscape of persuasion from selecting and presenting human-generated content, albeit in a personalized manner, to automatically generating tailored, context-aware, and hyper-personalized messages at scale.

Several key characteristics make *LLM Systems*, i.e. LLMs integrated with other technologies and deployed for specific applications, particularly effective for persuasive communication. First of all, they allow for unprecedented **personalization and adaptability** as they excel at tailoring messages to individual preferences and psychological profiles, creating targeted content that resonates with a targeted group or even a specific individual [3, 31, 5, 12, 13, 21]. Second and related to this, LLM Systems appear capable at **exploiting cognitive biases** that make humans vulnerable to persuasion and manipulation, thus allowing them to achieve persuasive impact beyond what was possible with classical scalable approaches [5, 9, 12, 13, 18]. Third, they allow for **interactive conversations**, enabling engaging experiences that can be more effective than traditional one-way messaging [2, 5, 12, 13, 29]. Fourth, despite known weaknesses of LLMs in terms of hallucinations, they are capable of a high degree of **message consistency** even across a prolonged personalized interaction [5, 14, 28, 16, 22]. And finally, in comparison to other persuasion methods that share the previously mentioned characteristics, LLM Systems are far more **scalable** because they can engage in hyper-personalized persuasive interactions at scale without proportional resource increases [2, 3, 9, 13, 14, 16].

During the past two years, these key characteristics of LLM Systems have led to the emergence of a new subfield of persuasion research, which we refer to as *LLM-based persuasion*.

The effectiveness of LLM Systems for persuasion depends on various configurable factors, as illustrated in Figure 1. These include whether interactive dialogue is enabled, the scale of the language model employed, whether AI authorship is disclosed to users, the use of persuasive prompt engineering, incorporation of personal data for personalization, and employment of authoritative language. Each of these factors can significantly impact the system's persuasive capabilities, as we will discuss in detail in Section 3.

Purpose of this survey. Through a systematic review of peer-reviewed articles, conference papers, and reputable industry reports published between 2022 and 2024, we provide a comprehensive overview of the state-of-the-art in LLM-based persuasion. We examine the applications across various domains of LLM Systems for persuasion, factors and design choices influencing their persuasiveness, the methodological techniques used to study their persuasiveness and thereby quantify progress, and the associated ethical and regulatory challenges. Our survey focuses on experimental studies that directly tested the persuasive capabilities of LLM Systems.

The rest of the paper is organized as follows:

- Section 2 reviews the application domains where LLM-based persuasion has been studied, including politics, marketing, public health, e-commerce, and efforts to combat misinformation.
- Section 3 analyzes the factors that influence the persuasive capabilities of LLM Systems, such as AI source labeling, model scale, and persuasion strategies.
- Section 4 examines the experimental design patterns employed in LLM persuasion studies, highlighting study methodologies, treatment condition structures, control strategies, and methods used to assess persuasiveness.
- Section 5 discusses the ethical considerations and regulatory landscape surrounding LLM-based persuasion.
- Section 6 concludes by summarizing key findings, identifying research gaps, and suggesting directions for future work.

By mapping the current state of research on LLM-based persuasion, this survey aims to provide a comprehensive overview of this rapidly evolving field. In doing so, we hope to raise awareness among researchers, practitioners, policy makers, and the general public, of the power of LLM Systems for persuasion today and in the near future, and to point out the resulting ethical and societal challenges that lie ahead of us.

Table 1 provides an overview of the surveyed papers, showing for each paper the application domain it covered, the factors influencing persuasiveness it studied, and the experimental design it adopted.

Table 1: Summary of papers by application domains, influencing factors studied, and methodological characteristics. Filled circles (●) indicate the domain/factor was explicitly studied or the methodology was used in the paper.

	Application Domains					Influencing Factors						Methodology					Success Metrics							Persuasiveness			
Citation	Public Health	Politics	E-commerce	Misinfo. Red.	Charity	Interaction	Model Scale	AI Source Labeling	Prompt Design	Personalization	Authority	RCT Design	Between Subj.	Human Control	Pre-Post Measure	Long. Follow-up	Opinion Change	Agreement & Class.	Behavioral Intent	Engagement & Detect.	Perceived Effect.	Technical Metrics	Temporal	Superhuman	On Par	Inferior	No Comparison
[2]	●	○	○	○	○	●	○	○	○	○	●	●	●	○	●	●	●	○	●	●	○	○	●	○	●	○	○
[3]	○	●	○	○	○	○	○	○	●	○	●	●	●	●	○	●	○	○	●	●	○	○	○	○	○	○	○
[5]	○	●	○	●	○	●	○	○	●	○	○	○	○	○	○	●	○	○	●	●	○	○	○	●	○	○	○
[8]	●	○	○	○	○	○	○	●	○	○	●	●	●	●	○	○	●	○	○	●	●	○	○	○	●	○	○
[9]	○	○	○	●	○	●	○	○	●	○	○	○	●	●	○	○	○	●	○	○	●	○	○	○	●	○	○
[12]	○	○	●	○	○	●	○	○	○	●	○	○	○	○	○	●	○	●	○	●	○	○	○	○	○	○	●
[13]	○	○	○	●	○	●	○	○	○	●	○	●	●	○	●	●	●	○	●	○	○	○	●	○	○	○	●
[14]	○	●	○	○	○	○	●	○	●	○	○	○	●	●	○	○	●	●	○	○	●	○	○	○	●	○	○
[16]	●	○	○	●	●	●	○	○	●	●	○	○	○	○	○	○	●	○	●	●	●	●	○	○	●	○	○
[17]	○	●	○	○	○	○	○	○	●	○	○	●	●	●	○	○	○	●	○	○	●	○	○	○	●	○	○
[18]	○	●	○	○	○	○	○	○	●	●	○	○	○	●	○	●	○	●	○	●	○	○	○	○	●	○	○
[20]	○	●	○	○	○	○	○	○	○	●	○	●	●	○	○	●	○	○	●	○	○	○	○	○	●	○	○
[21]	○	●	○	○	○	○	○	○	●	○	●	●	●	●	○	○	●	○	○	●	○	○	○	●	○	○	○
[22]	○	●	○	○	○	○	●	○	○	○	○	●	●	●	○	○	●	○	○	●	○	○	○	○	●	○	○
[28]	●	○	○	○	○	○	○	●	●	○	●	○	●	●	○	○	●	○	○	○	●	○	○	●	○	○	○
[29]	○	○	○	●	○	●	○	○	○	○	●	○	○	○	○	○	●	○	○	○	○	●	●	○	○	○	●
[30]	●	○	○	○	○	○	○	●	○	○	○	●	●	●	○	○	○	○	●	●	○	○	○	●	○	○	○
[31]	●	●	●	○	○	○	○	●	●	●	○	●	●	●	○	○	●	●	●	●	●	○	○	○	○	○	●
[32]	○	○	●	○	○	○	○	○	●	●	○	○	○	●	○	○	○	○	●	●	●	○	○	○	●	○	○
[33]	○	○	○	●	○	●	○	○	○	○	●	○	●	○	○	○	●	○	○	○	●	○	○	○	○	○	●
[34]	○	●	○	●	○	○	○	●	○	○	○	●	●	●	○	○	○	●	○	●	●	○	○	○	●	○	○
[35]	○	○	○	○	●	○	○	○	●	○	○	○	○	●	○	●	○	○	●	●	●	○	○	○	○	○	●
[36]	○	●	○	○	○	●	●	●	●	●	○	●	●	●	○	○	●	●	●	●	●	○	○	●	○	○	○
[38]	○	○	●	○	○	○	○	○	○	○	○	●	●	○	●	○	○	○	●	●	●	○	○	●	○	○	○
[39]	○	●	○	○	○	○	○	○	●	●	○	○	●	○	○	●	○	○	○	●	●	○	○	●	○	○	○
[40]	●	○	○	●	○	○	○	○	●	○	○	○	○	●	○	○	○	●	○	●	●	●	○	○	●	○	○
[42]	●	●	○	○	○	●	○	●	○	○	●	○	●	●	○	○	●	○	○	○	●	○	○	○	○	●[2]	○
[44]	○	○	○	○	●	●	○	○	●	○	○	○	○	○	○	○	○	●	○	●	○	○	○	○	○	○	●
[45]	○	○	●	○	○	●	○	●	○	○	●	●	●	●	○	○	●	○	●	○	●	○	○	●	○	○	○
[46]	●	○	○	●	○	○	●	●	●	○	●	○	○	●	○	○	○	●	○	●	●	●	○	●	○	○	○

2 Application Domains

LLM Systems have been deployed across various domains to influence human behavior and attitudes. This section examines five key application areas, each presenting unique challenges and opportunities for persuasive AI.

2.1 Public Health

Public health applications uniquely combine the need for scientific accuracy with persuasive impact, where even small improvements in message effectiveness can have significant societal benefits.

Vaccination communication has been a key focus area. Altay et al. [2] developed a COVID-19 vaccine chatbot that effectively improved attitudes and intentions through interactive dialogues allowing users to choose questions about common vaccine hesitancies. Karinshak et al. [28] found that GPT-3-generated messages were perceived as more effective and evoked more positive attitudes than human-authored CDC messages, though participants preferred messages not labeled as AI-generated.

Beyond vaccines, Lim and Schmälzle [30] found that AI-generated health messages rated higher than human-generated ones when source was not disclosed, though disclosure reduced preference particularly among those with negative AI bias. Böhm et al. [8] investigated health advice, finding that while AI authors were perceived as less competent when source was transparent, there was no significant difference in content quality or intent to use the advice. Costello et al. [13] demonstrated success in using personalized LLM dialogues to address health-related conspiracy theories, achieving approximately 20% reduction in belief through their approach.

[2] Teigen et al. [42] study only the source labeling factor and use only AI generated messages. They conclude that Human source labels are more persuasive than AI source labels.

2.2 Politics

Political marketing, communication, and propaganda are obvious areas in which LLM Systems are bound to be deployed in practice. Several studies have specifically focused on this area.

Bai et al. [3] demonstrated that LLM-generated political messages could be as persuasive as human-written ones across different policy issues, with participants generally unable to distinguish between AI and human authorship. Palmer and Spirling [34] found that while LLM arguments were generally as convincing as human arguments, they had lower reading difficulty and a more positive tone, with awareness of AI authorship leading to slightly lower preference.

Although the evidence remains somewhat mixed and limited, microtargetting appears to enhance persuasiveness in politics. Hackenburg and Margetts [20] examined personalized political messages, finding that baseline LLM-generated messages were inherently persuasive even without extensive targeting. Simchon et al. [39] showed increased perceived persuasiveness when messages matched personality traits, though effect sizes were small. Hackenburg et al. [21] found that LLMs performing partisan role-play outperformed human experts particularly on right-leaning issues.

The risks of political misuse were specifically examined by Goldstein et al. [17] and Durmus et al. [14]. Goldstein et al. [17] found GPT-3-generated propaganda was highly persuasive though slightly less compelling than human-created propaganda (43.5% vs 47.4% agreement), while Durmus et al. [14] showed that larger models like Claude 3 Opus could produce arguments comparable to human-crafted ones. Teigen et al. [42] demonstrated that AI-labeled arguments were consistently rated less persuasive than human-authored ones across all domains, with this effect particularly strong in expert domains.

2.3 E-commerce and Marketing

E-commerce applications highlight the intersection of persuasion with economic incentives, where message effectiveness directly translates to financial outcomes.

Chen et al. [12] examined how AI chatbots could enhance users' adoption intentions through personalized product recommendations, finding that both cognitive and emotional trust played important roles. Meguellati et al. [32] studied personality-based advertisement generation, finding that LLM-generated ads for openness-based traits performed comparably to human-written ads, though performance varied across different personality traits.

Shin and Kim [38] demonstrated that LLM-modified customer complaints significantly outperformed unedited ones in achieving higher perceived likelihood of compensation success, particularly when focusing on enhancing clarity and professionalism. Zhang and Gosline [45] found that AI-generated content was perceived as high quality when content origin was not disclosed, though human favoritism emerged when source was revealed.

2.4 Mitigating Misinformation

Misinformation mitigation represents a distinct challenge where LLM Systems aim to not only persuade but also maintain and enhance credibility.

Zhou et al. [46] found that AI-generated misinformation displayed distinct linguistic differences from human-created content, using more emotional and cognitive processing expressions, making detection more challenging. Spitale et al. [40] revealed that participants had difficulty distinguishing between GPT-3-generated and human-generated content, processing synthetic content more rapidly but with varied precision.

Interactive approaches showed particular effectiveness. Costello et al. [13] achieved significant reductions in conspiracy beliefs through personalized dialogues, while Metzger et al. [33] found that high authority communicative styles effectively increased trust more than capability disclaimers. Carrasco-Farre [9] showed that LLMs matched human persuasiveness by leveraging complex language and moral framing rather than emotional content.

2.5 Charity

Charitable giving applications involve unique persuasion challenges, requiring systems to motivate prosocial behavior without manipulation.

Yoon et al. [44] investigated how multiple coordinated chatbots could encourage charitable donations, finding that organizational alignment enhanced credibility though some approaches appeared overly insistent. Furumai et al. [16] demonstrated success with a multi-step dialogue approach that achieved superior scores in both persuasiveness and factual accuracy by grounding appeals in verifiable data and success stories.

3 Factors Influencing Persuasiveness

Although in general the performance of LLM Systems is impressive, some variability exists. Several authors have reported on LLM Systems achieving super-human performance in various persuasion-related tasks (e.g., Breum et al. [5], Hackenburg et al. [21], Karinshak et al. [28], Lim and Schmälzle [30], Salvi et al. [36], Simchon et al. [39], Zhang and Gosline [45], Zhou et al. [46]). Others have reported performance that is on par with human-level persuasion (e.g., Bai et al. [3], Goldstein et al. [17], Hackenburg and Margetts [20], Hackenburg et al. [22], Palmer and Spirling [34], Spitale et al. [40]), or in some conditions slightly inferior (e.g., Teigen et al. [42]).

The effectiveness of LLM Systems in persuasive communication is influenced by a complex interplay of factors. Understanding these factors is crucial for both researchers and practitioners in the field of AI-driven persuasion. This section examines six key elements that shape the persuasive capabilities of LLM Systems: interaction approach, model scale, prompt design, AI source labeling, personalization, and appeals to authority.

3.1 Interaction Approach

The manner in which an LLM System interacts with users can significantly affect its persuasive impact. Research has investigated interaction approaches ranging from *passive information delivery* over *targeted personalized messaging* to *interactive dialogues and debates*.

At one end of the spectrum, LLM Systems deliver content for users to consume with minimal interaction. Altay et al. [2] evaluated a chatbot delivering vaccine information through static text and interactive formats, finding that increased engagement time with static content could be equally effective. Bai et al. [3] demonstrated how LLM-generated political messages could influence public opinion through direct content delivery.

More sophisticated approaches incorporate personalization and targeted messaging. Matz et al. [31] developed systems aligning messages with users' personality traits, while Meguellati et al. [32] generated personalized advertisements for specific consumer segments. In political contexts, Hackenburg and Margetts [20] and Simchon et al. [39] demonstrated how personalized political messages could shape opinion dynamics. Their research showed that while personalization enhanced message relevance, its effectiveness varied depending on the political context and user characteristics.

The most advanced approaches enable interactive dialogues and debates. Breum et al. [5] demonstrated this through structured debates where LLMs acted as convincing agents, while Salvi et al. [36] revealed how conversation dynamics influence persuasiveness in debate settings. In addressing misinformation, Costello et al. [13] and Durmus et al. [14] showed how interactive dialogues could effectively counter false beliefs, with Costello et al. [13] reporting approximately 20% reduction in conspiracy beliefs through personalized dialogues.

Research suggests that active engagement typically achieves stronger persuasive effects than passive information delivery, particularly for complex topics requiring sustained attitude change. This advantage appears to stem from the ability to address specific user concerns in real-time and adapt persuasive strategies based on user responses.

3.2 Model Scale

The scale of the model, particularly its number of parameters and overall complexity, plays a significant role in determining the persuasive capabilities of LLM Systems. Research in this area has focused on how model size impacts language understanding, coherence, and ultimately, persuasive ability.

Several studies have demonstrated that larger LLMs exhibit enhanced persuasive capabilities due to improved language understanding and coherence. Breum et al. [5] conducted a comprehensive study using advanced systems like Llama-2-70B-chat and Claude 3 Opus, finding that increased model size correlated with more nuanced and contextually appropriate persuasive content. Durmus et al. [14] observed that larger models could produce arguments comparable in quality to those crafted by humans, particularly in tasks requiring sophisticated language understanding.

However, the relationship between model size and persuasive power is not linear. Hackenburg et al. [22] observed diminishing returns on persuasiveness as model size increases. Their study indicated that beyond a certain scale, additional parameters do not significantly enhance the LLM's persuasive capabilities. This suggests that while model scale is important, beyond a certain threshold, factors such as training data quality and fine-tuning strategies may play a larger role in enhancing persuasiveness.

An interesting aspect of larger models is their increased ability to perform zero-shot tasks, as highlighted by Furumai et al. [16]. They found that increased model size enables LLM Systems to leverage their generalization capabilities, making them more adaptable across various domains without specific training. This adaptability is particularly valuable in persuasive contexts where the ability to generate contextually appropriate and convincing arguments across diverse topics is essential.

3.3 AI Source Labeling

The impact of disclosing AI authorship on persuasiveness has been extensively studied, with findings suggesting complex effects on message reception and effectiveness.

Several studies have found that content perceived as human-authored tends to be viewed as more trustworthy and persuasive compared to content explicitly labeled as AI-generated. Böhm et al. [8] observed that AI-authored content is often associated with lower perceived competence when identity was known, although this did not affect content quality assessments or sharing intentions. Karinshak et al. [28] demonstrated that messages labeled as AI-generated were rated less favorably, showing a dispreference for AI-labeled content despite it being more persuasive objectively. Teigen et al. [42] found that AI-labeled arguments were perceived as less persuasive, particularly in expert domains, with expert human labels increasing persuasiveness significantly more than expert AI labels.

However, some studies suggest ways to mitigate these negative effects. Matz et al. [31] found that the impact of source disclosure was minimal on persuasiveness when messages were effectively personalized. Their research showed that knowledge of AI message generation did not significantly deter the effectiveness of persuasion, indicating robustness when combined with sophisticated personalization strategies.

3.4 Prompt Design

Prompt design significantly impacts the effectiveness of LLM Systems in persuasive communication. Several studies have demonstrated that prompt designs which establish a personal connection with users significantly enhance the persuasiveness of LLM-generated content.

Karinshak et al. [28] tested various prompting strategies, including zero-shot and few-shot prompts, finding that proper prompt design significantly affected message quality and relevance. Simchon et al. [39] discovered that prompts encouraging logical reasoning and evidence-based persuasion outperformed those relying on rhetorical or emotional language, resulting in more convincing arguments.

Furumai et al. [16] demonstrated that strategic prompting balancing persuasive communication with factual correctness ensures the reliability of LLM-generated content. Their research employed few-shot examples to guide zero-shot responses across multiple domains while maintaining strategic approaches. Pauli et al. [35] found that prompt variations in instructions to increase or decrease persuasive language significantly influenced text output, with reduction instructions leading to shorter, less persuasive text.

3.5 Personalization

Empirical evidence strongly supports the effectiveness of tailoring content to individual user characteristics. This personalization significantly enhances both engagement and persuasive impact across various domains.

Meguellati et al. [32] demonstrated that LLM personalization for specific personality traits like openness achieved user engagement similar to human-written content, though performance varied across different traits. Matz et al. [31] showed that personalized messages significantly influenced attitudes and intended behaviors more than generic messages, with effectiveness remaining robust even when AI authorship was disclosed.

In the political sphere, Simchon et al. [39] found that personalized political messages matching participant personality traits increased perceived persuasiveness, albeit with small effect sizes. Salvi et al. [36] demonstrated that including personalization in debate contexts significantly enhanced persuasive effectiveness, highlighting the importance of tailored messaging in dynamic interactions.

3.6 Appealing to Authority

The ability of LLM Systems to leverage authority and expertise represents another crucial aspect of their persuasive capabilities. Several studies have demonstrated the effectiveness of authority-based approaches in enhancing message credibility and impact.

Altay et al. [2] found that messages incorporating expert-verified information built greater credibility with users and were consequently more persuasive. Their study demonstrated that when LLM Systems present content as being endorsed by recognized authorities, users are more likely to trust and be influenced by the information.

Metzger et al. [33] investigated how conversational styles affect trust, finding that high authority prompts increased trustworthiness more than any disclaimers about limitations. Their research showed that when LLM Systems adopt authoritative tones and present information in a manner consistent with expert communication, trust levels are elevated, which in turn enhances persuasiveness.

The effectiveness of authority appeals appears particularly important in complex or specialized domains. Costello et al. [13] found that in contexts requiring high cognitive effort or involving complex arguments, such as addressing conspiracy theories, the perceived expertise of the LLM System significantly impacted its persuasive power. Similarly, Yoon et al. [44] demonstrated that authority conveyed through organizational alignment enhanced chatbot credibility, though the effectiveness varied based on the approach to authority presentation.

4 Experimental Design Patterns

Quantifying the effectiveness of automated approaches to persuasion, while ensuring high ecological validity, is a challenging task. This has been approached in widely differing ways across the literature, varying depending on the particular research questions addressed and methodological choices. This section examines the key characteristics of the experimental designs that have been developed and used, including study types, treatment condition structures, control strategies, and quantifications of success.

4.1 Study Types and Methodological Approaches

The reviewed studies demonstrate a clear preference for controlled experimental designs, with several key methodological approaches emerging. Randomized controlled trials (RCTs) form the backbone of many investigations, with studies like Altay et al. [2], Costello et al. [13], Hackenburg and Margetts [20], and Salvi et al. [36] employing this rigorous methodology to establish causal relationships. Pre-registered experiments, as used by Bai et al. [3], Böhm et al. [8], and Goldstein et al. [17], add another layer of methodological rigor by committing to analysis plans before data collection.

Between-subjects designs are particularly common, ranging from simple two-condition comparisons to complex factorial designs. Lim and Schmälzle [30] employed a between-subjects design with substantial sample sizes (151 and 183 participants across two studies) to examine source disclosure effects. Spitale et al. [40] utilized a between-subjects design with 697 participants to assess responses to AI-generated versus human-authored tweets. Zhang and Gosline [45] conducted a large-scale between-subjects study with 1203 participants comparing four different content generation paradigms. Hackenburg et al. [21] implemented a sophisticated nine-condition design examining various combinations of source attribution and message framing, while Teigen et al. [42] employed a 4x2 factorial design to investigate the interaction between source expertise and AI attribution.

Comparative studies were also prevalent, with Carrasco-Farre [9], Durmus et al. [14], and Palmer and Spirling [34] directly comparing different persuasive approaches. Zhou et al. [46] employed a mixed-methodology approach combining experimental user studies with algorithmic evaluations to assess misinformation detection mechanisms.

Some studies employed innovative methodological approaches, such as Breum et al. [5]'s synthetic persuasion dialogue and Li et al. [29]'s simulation of opinion network dynamics. Mixed-method approaches were also observed, as in Shin and Kim [38]'s combination of observational and experimental methodologies, while others like Chen et al. [12] and Metzger et al. [33], utilized survey-based designs to examine user perceptions and responses.

4.2 Treatment Condition Structures

Treatment conditions in these studies typically varied along four main dimensions, reflecting the specific research questions addressed. These research questions often relate to the factors influencing persuasiveness of LLM Systems discussed in Sec. 3.

1. **Source Attribution**: Many studies manipulated whether participants were aware of the AI authorship of content. For example, Palmer and Spirling [34] explicitly compared conditions where argument sources were revealed versus concealed, and Böhm et al. [8] examined transparent versus non-transparent author identity conditions. Karinshak et al. [28] further explored this by varying source attributions across different expertise levels. Zhang and Gosline [45] implemented baseline (no awareness), partially informed, and fully informed conditions regarding content origins. Lim and Schmälzle [30] specifically examined how source disclosure moderated the impact of negative attitudes toward AI.

2. **Message Characteristics**: Studies often varied the nature of the generated content. For instance, Carrasco-Farre [9] examined different persuasive strategies, and Simchon et al. [39] explored personality-congruent political ad variations. Pauli et al. [35] specifically focused on varying persuasive language characteristics in text pairs. Spitale et al. [40] implemented a novel four-way categorization of content (organic true, organic false, synthetic true, and synthetic false) to evaluate both truthfulness and origin perception.
3. **Interaction Patterns**: Several studies manipulated how participants engaged with the LLM Systems. Yoon et al. [44] tested different chatbot response patterns, while Costello et al. [13] implemented structured three-round conversations. Furumai et al. [16] explored various interaction strategies through their modular chatbot system.
4. **Targeting Approaches**: Multiple studies examined different targeting strategies. For example, Hackenburg and Margetts [20] compared accurate targeting, non-targeted messages, and false targeting conditions. Matz et al. [31] and Meguellati et al. [32] investigated personality-based targeting approaches.

4.3 Control Condition Strategies

Control conditions exhibited notable variation across studies, with five main approaches emerging:

1. **No-Message Controls**: Several studies (Goldstein et al. [17], Hackenburg et al. [22], Hackenburg et al. [21]) used no-message exposure as their baseline condition.
2. **Human-Authored Content**: Studies like Durmus et al. [14], Meguellati et al. [32], and Karinshak et al. [28] used human-written content as control conditions. Spitale et al. [40] and Zhou et al. [46] used human-created content as benchmarks for comparing AI-generated materials.
3. **Alternative Interaction Methods**: Some studies employed different interaction formats as controls. For example, Altay et al. [2] used brief text reading as a control for chatbot interaction, while Salvi et al. [36] used human-human debates as a control condition.
4. **Neutral Content**: Several studies used content-matched but neutral versions as controls, such as Griffin et al. [18]'s use of neutral article versions and Bai et al. [3]'s neutral message control. Lim and Schmälzle [30] employed unlabeled messages as a control condition to test source disclosure effects.
5. **Baseline Systems**: Technical studies often used existing systems as controls, exemplified by Furumai et al. [16]'s use of state-of-the-art chatbots and Breum et al. [5]'s baseline argument condition.

4.4 Quantifications of success

The analysis of LLM persuasiveness employs diverse quantifications of success across studies, which can be categorized into seven main dimensions. Each dimension captures different aspects of persuasive impact and employs distinct methodological approaches, reflecting both traditional persuasion research methods and novel approaches specific to AI-generated content.

Opinion Change Opinion change measurements focus on shifts in participants' beliefs or attitudes following exposure to LLM-generated content. These studies employ various methodological approaches to capture both immediate and longer-term attitude shifts. Altay et al. [2] utilized pre-post surveys to measure changes in vaccine attitudes, while Costello et al. [13] implemented a 0-100 belief scale to track changes in conspiracy theory beliefs. In political contexts, Hackenburg et al. [22] measured shifts in policy stance agreement, and Griffin et al. [18] assessed attitude changes through article framing effects. More nuanced approaches included Carrasco-Farre [9]'s examination of agreement changes using linguistic metrics and Durmus et al. [14]'s use of 7-point Likert scales to measure stance changes.

Agreement with Messages and Classification Accuracy While related to opinion change, agreement measurements typically focus on immediate responses to persuasive content rather than longer-term attitude shifts. Bai et al. [3] employed 101-point policy support scales, providing fine-grained measurements of agreement levels. Spitale et al. [40] introduced a novel approach measuring participants' ability to correctly classify content origin and truthfulness, finding that participants struggled to differentiate between AI and human-authored content. Goldstein et al. [17] and Salvi et al. [36] examined agreement in political contexts, with the latter specifically focusing on debate scenarios. More complex approaches included Hackenburg et al. [21]'s use of role-playing variations and Hackenburg and Margetts [20]'s examination of targeting effects on agreement. Furumai et al. [16] contributed unique insights through worker evaluations of chatbot-generated persuasive content.

Behavioral Intent and Economic Metrics Moving beyond attitudinal measures, behavioral intent measurements provide insights into the practical impact of LLM persuasion. Böhm et al. [8] and Chen et al. [12] focused on participants' intentions to follow AI-generated advice, revealing an interesting disconnect between source perception and behavioral response. Zhang and Gosline [45] introduced economic metrics through willingness-to-pay measurements and satisfaction ratings using 7-point Likert scales. Shin and Kim [38] examined behavioral outcomes in customer service contexts, measuring the likelihood of concrete actions following LLM-edited complaint messages.

Engagement and Detection Metrics While behavioral intent focuses on future actions, engagement metrics assess how users interact with LLM Systems during persuasive exchanges. Meguellati et al. [32] examined engagement through product ratings and responses to personality-tailored advertisements, utilizing both behavioral and self-report measures. Zhou et al. [46] introduced detection-based metrics, measuring both human judgment accuracy and algorithmic detection rates for AI-generated content. Yoon et al. [44] provided insights into engagement patterns in multi-chatbot environments, measuring user preferences and interaction patterns across different chatbot configurations.

Perceived Effectiveness Perceived effectiveness measurements capture users' subjective evaluations of LLM persuasiveness through various methodological approaches. Lim and Schmälzle [30] assessed perceived message effectiveness through ratings and selection tasks, particularly examining how source disclosure affected evaluations. Karinshak et al. [28] utilized multi-dimensional effectiveness scales, including argument strength and trustworthiness measures. Metzger et al. [33] employed validated scales, including Jian et al. [26]'s trust scale and Bartneck et al. [4]'s human-likeness scale. Pauli et al. [35] and Simchon et al. [39] focused on comparative effectiveness ratings, while Teigen et al. [42] incorporated the GAAIS scale[37] for deeper analysis of AI-specific perceptions. Matz et al. [31] introduced economic measures through willingness-to-pay metrics, and Palmer and Spirling [34] examined perceived argument convincingness in political contexts. Breum et al. [5] used pairwise comparisons through crowdworkers to evaluate the relative persuasiveness of arguments.

Technical Metrics Complementing subjective measurements, a subset of studies employed computational and technical measures to assess persuasiveness. Carrasco-Farre [9] utilized a comprehensive suite of linguistic metrics, including readability scores, lexical complexity measures, and sentiment analysis tools. Li et al. [29] developed simulation metrics for analyzing opinion dynamics in network contexts. Shin and Kim [38] combined traditional persuasion measures with AI detection tools to assess message authenticity and effectiveness. Spitale et al. [40] analyzed processing speed and precision in tweet classification, providing insights into how participants engage with and evaluate AI-generated content.

Temporal Measurement Patterns The studies exhibit distinct patterns in their temporal approach to measurement, particularly regarding pre/post comparisons and longitudinal follow-up. A substantial number of studies employed pre/post measurements to assess immediate persuasive effects, with Altay et al. [2], Bai et al. [3], and Durmus et al. [14] using baseline and endline questionnaires to track attitude changes. Goldstein et al. [17] and Salvi et al. [36] measured agreement both before and after exposure to persuasive content, while Metzger et al. [33] assessed trust levels before and after conversations. Yoon et al. [44] implemented a comprehensive temporal structure with pre-surveys, interaction measurements, and post-surveys.

However, longitudinal follow-up measurements were relatively rare in the literature. Notable exceptions include Costello et al. [13], who conducted follow-up measurements at both 10 days and 2 months post-intervention to assess the durability of belief changes, and Altay et al. [2], who implemented a follow-up questionnaire 1–2 weeks after the initial experiment. Salvi et al. [36] included post-debate identity perception measurements, providing insights into lasting effects of the persuasive interaction. Most new studies including Spitale et al. [40], Lim and Schmälzle [30], and Zhang and Gosline [45] focused on immediate post-exposure measurements, further highlighting the scarcity of longitudinal studies in the field.

The diversity in approaches to quantify success in achieving persuasion reflects both the complexity of assessing persuasiveness and the field's methodological evolution. While some dimensions draw from established persuasion research traditions, others introduce novel metrics specifically designed for AI-generated content. The combination of subjective measures with objective metrics provides a more comprehensive understanding of LLM persuasiveness.

5 Ethical Considerations

The surveyed literature demonstrates that LLM Systems have been built that meet or exceed human capabilities in persuasion. Thus, their deployment in practice is bound to become more common as well. It should therefore be a matter of priority to consider the ethical implications associated with their deployment.

This section examines some of the potential dangers and pitfalls of using LLM Systems for persuasion, and briefly discusses the current state of regulation in this rapidly evolving field. We will summarize the concerns and associated suggestions voiced in the relevant literature, occasionally complemented with our own perspective.

5.1 Ethical and societal risks

The main ethical risks identified relate to information integrity and the risk of manipulation, deception, and coercion, the risk of societal polarization, the risk of biases being propagated by LLM Systems used for persuasion, the lack of transparency and accountability, and possible threats to privacy. We discuss them in turn below.

5.1.1 Information integrity

Mis-, dis-, and malinformation: a controversial subject The nature and impact of *mis- and disinformation* is receiving increasing amounts of attention, both in academic research and in the public sphere. It is a controversial field, not least because there is lack of consensus about both the definition and the actual risks posed by mis- and disinformation today. While *misinformation* is often used as an umbrella term for all false information, it is sometimes specified as falsehood *without* intention to cause harm or deceive, in contrast to *disinformation* that denotes false information *with* malicious intent [25]. Controversy in the field is illustrated in a June 2024 special issue of Nature on ‘fake news’, which featured two articles that appear to contradict each other. Ecker et al. [15] argues that misinformation is even more harmful for democracy than we might think. Budak et al. [6], on the other hand, analyzed the reach and impact of misinformation, and found that the number of people exposed to misinformation remains limited today, that those who do get exposed are relatively unaffected by it, and that those who are affected by it were looking for it anyway. A similar conclusion had already been reached earlier by Altay et al. [1].

Both views may be valid, as they use very different definitions of ’misinformation’. While Budak et al. [6] adopt a narrow definition as demonstrably false and inflammatory content, Ecker et al. [15]’s definition is more expansive, and includes *true* information that may be harmful, sometimes referred to as *malinformation*. The problematically loose manner in which terms such as mis- and disinformation are being used, not only in popular parlance but also in scientific discourse, has been discussed in depth by Uscinski et al. [43].

Can (and should) LLM Systems bring order in the information chaos? Many authors argue that a narrow definition is insufficient, as much harm can be done by spreading information that is factually true. They argue that true facts can also be misleading, inflammatory, or otherwise harmful, if they are presented in an emotionally charged manner; if other truths are omitted or concealed to present a biased picture of reality; or when the truthful claims are easy to misunderstand by people who lack the required background.

This view has compelled some researchers to advocate or investigate the use of LLM Systems in combatting misinformation, sometimes in its expansive definition. For example, Altay et al. [2] explored the use of an LLM-based chatbot to counter misinformation about COVID-19 vaccines. Just minutes of interaction with the chatbot significantly increased people’s vaccination intentions, and made their attitudes toward COVID-19 vaccines more positive. Chen and Shu [11] study more generally how to use LLM Systems to combat misinformation, including LLM-generated misinformation. Costello et al. [13] find that conversational LLM Systems can be effective in durably countering belief in unsubstantiated or false conspiracy theories.

Such use of persuasive LLM Systems sounds attractive (and to some, even inevitable), particularly for authorities, policy makers, and the media, who are expected to inform the public. Even so, the question remains under which circumstances, if ever, the persuasive power of LLM Systems can be rightly and safely employed. Delegating the power to influence the public’s views on what is false and what is true, let alone on what is just and what is unjust, even to well-intentioned people in a position of authority, carries obvious risks in complex debates where truth may be elusive or where multiple perspectives on the same facts are possible [6, 1]. The additional persuasive power offered by LLM Systems greatly adds to these risks.

In their article on catastrophic AI risks, Hendrycks et al. [23] warn against such centralized control and monopolization of trusted information. They describe two ways in which this may happen. First, advanced AI systems could “centralize the creation and dissemination of trusted information”, allowing only the technically most powerful and capable actors to shape popular narratives. Second, AI-enabled censorship may arise, starting with good intentions in the form of fact-checking. However, such efforts may fail to solve the misinformation problem, while they may inspire certain governments to deploy 'fact-checking' AI systems to suppress true information they consider undesirable (e.g., by alleging it is malinformation).

Risks of LLM Systems for spreading mis-, dis-, and malinformation Many studies primarily focus on the risk that LLM Systems can be used to generate mis-, dis-, and malinformation, rather than on their potential to counter it. Carrasco-Farre [9] and Goldstein et al. [17] highlight the broader implications of LLM-generated misinformation on society and the integrity of democratic processes, pointing out "the dual potential of LLMs to both enhance and undermine information integrity" [9]. Breum et al. [5] also investigate the persuasive dynamics *between* LLM Systems, concluding that they "have the potential of playing an important role in collective processes of opinion formation in online social media". And Spitale et al. [40] find that GPT-3 is not only better than humans at producing accurate information that is easy to understand, but also at creating compelling disinformation.

Besides the risk of monopolization of truth, Hendrycks et al. [23] also warn against the risks of AI systems that "pollute the information ecosystem with motivated lies" due to their ability to generate highly personalized disinformation at an unprecedented scale. They point out that this problem is exacerbated by the emotional connections and trust relationships people forge with companion chatbots. To make matters worse, such chatbots have access to very private information, which they can exploit for enhanced persuasion, controlled by large technology corporations and other powerful actors.

The political sphere is particularly vulnerable to the effects of AI-generated misinformation. Bai et al. [3] discuss the risks for democratic discourse, raising concerns about the integrity of political processes in an era of advanced AI-driven communication. Böhm et al. [8] and Palmer and Spirling [34] explore how AI-generated content can shape public opinion and potentially manipulate societal narratives. Hackenburg and Margetts [20] examine the implications of microtargeting political messages using LLM Systems, which can subtly manipulate individuals based on their psychological profiles.

Finally, Carroll et al. [10] focus specifically on *manipulation*, which can be defined as covert persuasion, or as persuasion involving coercion or deception. They also highlight the risk of AI systems manipulating without the intent of the system designers.

5.1.2 Societal polarization

A possible effect of the information chaos is societal polarization. Matz et al. [31] point out that LLM-generated personalized persuasive messages can exacerbate echo chambers and social polarization. By tailoring content to individual preferences, these systems risk creating information bubbles where users are predominantly exposed to content that aligns with their existing views, potentially reinforcing biases and limiting exposure to diverse perspectives. Hendrycks et al. [23] contend that pervasive persuasive AI systems might erode a common sense of reality, causing people to "retreat even further into ideological enclaves".

5.1.3 Bias and Unfairness

The persuasiveness of LLM Systems also poses concerns for the reinforcement of biases in the LLM. Indeed, recent research has shown that LLMs tend to reflect the ideological worldview of their creators, significant normative differences in the content they generate across different models and languages [7]. Li et al. [29] suggest that, in the modelling of collective opinion dynamics, adding LLM agents with opposed or random opinions may reduce the impact of any particular LLM's bias on people's opinions. As stressed by Karinshak et al. [28], failure to address such biases may impact have an unforeseen and dangerous impact on important areas like public health messaging.

5.1.4 Privacy Concerns

The use of LLM Systems for personalized persuasion introduces substantial privacy concerns, particularly due to the value that personal data can be exploited to personalize persuasive messaging [31]. Specifically in the political domain, Hackenburg and Margetts [20] raise concerns regarding microtargeting, emphasizing the privacy implications of using personal data to create personalized political messages.

5.2 Regulation and Ethical Guidelines

5.2.1 Legal Frameworks

The rapid advancement of LLM Systems in persuasive communication has outpaced existing legal frameworks, creating a need for updated laws and regulations. For example, Bai et al. [3] and Matz et al. [31] both emphasize the necessity for regulation to prevent misuse of LLM Systems in generating misinformation or manipulative content. Their work highlights the gap between current legal structures and the capabilities of modern AI systems. Simchon et al. [39] call for policy interventions to address the regulatory gaps in AI-mediated political messaging, particularly in the context of microtargeting. Their research underscores the need for legal frameworks that can effectively govern the use of AI in sensitive areas like political campaigning.

Various regulatory frameworks worldwide regulate some aspect of LLM Systems for persuasion. Data protection laws such as the European General Data Protection Regulation (GDPR)[3], the California Privacy Rights Act (CPRA)[4], and the Chinese Personal Information Protection Law (PIPL)[5], offer certain privacy protections that may limit the extent to which LLM Systems can exploit personal data for personalization. Regulations specifically targetting AI, such as the European AI Act[6] and the Chinese "Provisions on the Administration of Deep Synthesis Internet Information Services"[7] in particular require transparency regarding the AI origin of AI-generated content. Under its risk-based framework, the EU AI Act also bans certain uses of AI, such as for 'harmful subliminal manipulation'. Additionally, "AI systems intended to be used for influencing the outcome of an election or referendum" is a considered a high-risk application, which is subject to certain requirements and conformity assessment, thereby somewhat mitigating possible abuse of persuasive LLM Systems for political gain. The EU AI Act also demands that providers of so-called "General-Purpose AI Models with Systemic Risk" assess and mitigate them for systemic risks—a rather malleable concept.

These limited regulatory guardrails provide some protection depending on the application domain. Yet, it is fair to say that the most powerful generic LLM Systems, such as social media bots, chatbots, and social AI companions, which may engage in persuasion across many domains, remain largely unregulated.

5.2.2 Ethical Guidelines and Best Practices

In response to the ethical challenges posed by LLM Systems in persuasion, several researchers have proposed guidelines and best practices. For example, Carrasco-Farre [9] suggests promoting literacy in AI-generated content discernment and implementing ethical standards to ensure responsible use of persuasive AI technologies. Goldstein et al. [17] advocate for ethical reviews to manage risks associated with AI-generated propaganda, emphasizing the need for proactive measures to prevent misuse of these powerful tools. And for public health communication, Karinshak et al. [28] focus on developing best practices for AI-generated health messaging, emphasizing the importance of factuality and the need for careful human oversight.

Also providers of LLM Systems have adopted ethical guidelines in their terms and conditions. For example, Anthropic's *Acceptable Use Policy* prohibits the use "for activities and applications where persuasive content could be particularly harmful. [Anthropic does] not allow Claude to be used for abusive and fraudulent applications [...], deceptive and misleading content [...], and use cases such as political campaigning and lobbying" [14]. How strictly such policies are interpreted in practice, and how they can be enforced, is less clear.

5.2.3 Accountability Mechanisms

The need for clear accountability mechanisms in the design, deployment, and consequences of LLM-driven persuasion is a recurring theme in the literature. Hendrycks et al. [23] advocate for comprehensive impact assessments as part of a responsible deployment strategy for AI technologies. Their work emphasizes the importance of evaluating potential risks and establishing clear lines of responsibility in AI-driven persuasion systems.

The development of robust accountability frameworks remains an ongoing challenge, requiring collaboration between technologists, ethicists, policymakers, and other stakeholders to ensure responsible innovation in this rapidly evolving field.

[3] https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/
[4] https://thecpra.org/
[5] https://personalinformationprotectionlaw.com/
[6] https://artificialintelligenceact.eu/the-act/
[7] https://www.loc.gov/item/global-legal-monitor/2023-04-25/china-provisions-on-deep-synthesis-technology-enter-into-effect/

6 Conclusion and Future Directions

This survey has explored the multifaceted landscape of LLM Systems in persuasive communication, revealing both their significant potential and the complex challenges they present. LLM Systems have emerged as powerful tools for persuasion, offering unique advantages in scalability, efficiency, and personalization. The versatility and potential impact of LLM Systems is evident from their application across diverse domains, including public health, politics, e-commerce, efforts to combat misinformation, and promoting charitable giving.

The effectiveness of LLM-driven persuasion is influenced by multiple factors, including AI source labeling, model scale, and specific persuasion strategies. The field is very young, and yet, many surveyed papers already managed to achieve persuasiveness capabilities on par with or exceeding human performance. It is reasonable to assume that further research will lead to substantial additional gains in persuasive power of LLM Systems—if not in general, then at least in specific application contexts. Such research could focus on technical enhancements to LLM Systems for persuasion, or on the development of comprehensive benchmarks and metrics to enable reproducible and nuanced evaluations of such systems.

By highlighting the current state-of-the-art, this survey may serve both as a structured inventory of recent research on LLM Systems for persuasion, and as a warning to the academic community, policy makers, and the general public about the risks harbored by this technology. Indeed, the deployment of LLM Systems for persuasion raises significant ethical concerns, such as the risk pervasive personalized misinformation and the monopolization of trusted information, of societal polarization, of amplification of biases, of lack of accountability, and of incentives to invade user privacy. Other important ethical questions remain unanswered, such as about the impact of the language on persuasive capabilities of LLM Systems, and what the long-term effects are of exposure to AI-generated persuasive content on individuals and society, on public discourse and actual decision-making.

These ethical concerns underscore the need for robust ethical guidelines and regulatory frameworks. Unfortunately, existing ethical guidelines are little known and used, inconsistent, and unenforceable, and current regulatory initiatives appear insufficient to address these challenges. The present survey may be read as an encouragement, or perhaps a wake-up call, to address these gaps.

Acknowledgements

This research was funded by the BOF of Ghent University (BOF20/IBF/117), the Flemish Government (AI Research Program), the BOF of Ghent University (BOF20/IBF/117), and the FWO (11J2322N, G0F9816N, 3G042220, G073924N). This work is also supported by an ERC grant (VIGILIA, 101142229) funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Sacha Altay, Manon Berriche, and Alberto Acerbi. Misinformation on misinformation: Conceptual and methodological challenges. *Social media+ society*, 9(1):20563051221150412, 2023.

[2] Sacha Altay, Anne-Sophie Hacquin, Coralie Chevallier, and Hugo Mercier. Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Journal of Experimental Psychology. Applied*, 29(1):52–62, March 2023. ISSN 1939-2192. doi: 10.1037/xap0000400.

[3] Hui Bai, Jan Voelkel, Johannes Eichstaedt, and Robb Willer. Artificial Intelligence Can Persuade Humans on Political Issues, September 2023. URL `https://www.researchsquare.com/article/rs-3238396/v1`. ISSN: 2693-5015.

[4] Christoph Bartneck, Dana Kulic, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1:71–81, 01 2008. doi: 10.1007/s12369-008-0001-3.

[5] Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. The Persuasive Power of Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:152–163, May 2024. ISSN 2334-0770. doi: 10.1609/icwsm.v18i1.31304. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/31304`.

[6] Ceren Budak, Brendan Nyhan, David M Rothschild, Emily Thorson, and Duncan J Watts. Misunderstanding the harms of online misinformation. *Nature*, 630(8015):45–53, 2024.

[7] Maarten Buyl, Alexander Rogiers, Sander Noels, Iris Dominguez-Catena, Edith Heiter, Raphael Romero, Iman Johary, Alexandru-Cristian Mara, Jefrey Lijffijt, and Tijl De Bie. Large language models reflect the ideology of their creators, 2024. URL `https://arxiv.org/abs/2410.18417`.

[8] Robert Böhm, Moritz Jörling, Leonhard Reiter, and Christoph Fuchs. People devalue generative AI's competence but not its advice in addressing societal and personal challenges. *Communications Psychology*, 1(1):1–10, November 2023. ISSN 2731-9121. doi: 10.1038/s44271-023-00032-x. URL `https://www.nature.com/articles/s44271-023-00032-x`. Publisher: Nature Publishing Group.

[9] Carlos Carrasco-Farre. Large Language Models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments, April 2024. URL `http://arxiv.org/abs/2404.09329`. arXiv:2404.09329 [cs].

[10] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing Manipulation from AI Systems, October 2023. URL `http://arxiv.org/abs/2303.09387`. arXiv:2303.09387 [cs].

[11] Canyu Chen and Kai Shu. Combating Misinformation in the Age of LLMs: Opportunities and Challenges, November 2023. URL `http://arxiv.org/abs/2311.05656`. arXiv:2311.05656 [cs].

[12] Qian Chen, Changqin Yin, and Yeming Gong. Would an AI chatbot persuade you: an empirical answer from the elaboration likelihood model. *Information Technology & People*, ahead-of-print(ahead-of-print), January 2023. ISSN 0959-3845. doi: 10.1108/ITP-10-2021-0764. URL `https://doi.org/10.1108/ITP-10-2021-0764`.

[13] Thomas H. Costello, Gordon Pennycook, and David Rand. Durably reducing conspiracy beliefs through dialogues with AI, April 2024. URL `https://osf.io/xcwdn`.

[14] Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, April 2024. URL `https://www.anthropic.com/news/measuring-model-persuasiveness`.

[15] Ullrich Ecker, Jon Roozenbeek, Sander van der Linden, Li Qian Tay, John Cook, Naomi Oreskes, and Stephan Lewandowsky. Misinformation poses a bigger threat to democracy than you might think. *Nature*, 630(8015): 29–32, 2024.

[16] Kazuaki Furumai, Roberto Legaspi, Julio Vizcarra, Yudai Yamazaki, Yasutaka Nishimura, Sina J. Semnani, Kazushi Ikeda, Weiyan Shi, and Monica S. Lam. Zero-shot Persuasive Chatbots with LLM-Generated Strategies and Information Retrieval, July 2024. URL `http://arxiv.org/abs/2407.03585`. arXiv:2407.03585 [cs].

[17] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is AI-generated propaganda? *PNAS Nexus*, 3(2):pgae034, February 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae034. URL `https://doi.org/10.1093/pnasnexus/pgae034`.

[18] Lewis D. Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly T. Mai, Maria Vau, Matthew Caldwell, and Augustine Marvor-Parker. Susceptibility to Influence of Large Language Models, March 2023. URL `http://arxiv.org/abs/2303.06074`. arXiv:2303.06074 [cs].

[19] Rosanna E Guadagno and Robert B Cialdini. Online persuasion and compliance: Social influence on the internet and beyond. *The social net: The social psychology of the Internet*, pages 91–113, 2005.

[20] Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121, June 2024. doi: 10.1073/pnas.2403116121. URL `https://www.pnas.org/doi/10.1073/pnas.2403116121`. Publisher: Proceedings of the National Academy of Sciences.

[21] Kobi Hackenburg, Lujain Ibrahim, Ben M. Tappin, and Manos Tsakiris. Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues, December 2023. URL `https://osf.io/ey8db`.

[22] Kobi Hackenburg, Ben M. Tappin, Paul Röttger, Scott Hale, Jonathan Bright, and Helen Margetts. Evidence of a log scaling law for political persuasion with large language models, June 2024. URL `http://arxiv.org/abs/2406.14508`. arXiv:2406.14508 [cs].

[23] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An Overview of Catastrophic AI Risks, October 2023. URL `http://arxiv.org/abs/2306.12001`. arXiv:2306.12001 [cs].

[24] Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen. Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological science*, 23(6):578–581, 2012.

[25] Cherilyn Ireton. *Journalism,'Fake News'& Disinformation: Handbook for Journalism Education and Training*. UNESCO Publishing, 2018.

[26] Jiun-Yin Jian, Ann Bisantz, and Colin Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4:53–71, 03 2000. doi: 10.1207/S15327566IJCE0401_04.

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

[28] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–29, April 2023. ISSN 2573-0142. doi: 10.1145/3579592. URL https://dl.acm.org/doi/10.1145/3579592.

[29] Chao Li, Xing Su, Haoying Han, Cong Xue, Chunmo Zheng, and Chao Fan. Quantifying the Impact of Large Language Models on Collective Opinion Dynamics, January 2024. URL https://papers.ssrn.com/abstract=4688547.

[30] Sue Lim and Ralf Schmälzle. The effect of source disclosure on evaluation of AI-generated messages: A two-part study. *Computers in Human Behavior: Artificial Humans*, 2(1):100058, January 2024. ISSN 29498821. doi: 10.1016/j.chbah.2024.100058. URL http://arxiv.org/abs/2311.15544. arXiv:2311.15544 [cs].

[31] S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1):4692, February 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-53755-0. URL https://www.nature.com/articles/s41598-024-53755-0. Publisher: Nature Publishing Group.

[32] Elyas Meguellati, Lei Han, Abraham Bernstein, Shazia Sadiq, and Gianluca Demartini. How Good are LLMs in Generating Personalized Advertisements? In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, pages 826–829, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 9798400701726. doi: 10.1145/3589335.3651520. URL https://doi.org/10.1145/3589335.3651520.

[33] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–19, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642122. URL https://doi.org/10.1145/3613904.3642122.

[34] Alexis Palmer and Arthur Spirling. Large Language Models Can Argue in Convincing Ways About Politics, But Humans Dislike AI Authors: implications for Governance. *Political Science*, 75(3):281–291, September 2023. ISSN 0032-3187, 2041-0611. doi: 10.1080/00323187.2024.2335471. URL https://www.tandfonline.com/doi/full/10.1080/00323187.2024.2335471.

[35] Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. Measuring and Benchmarking Large Language Models' Capabilities to Generate Persuasive Language, June 2024. URL http://arxiv.org/abs/2406.17753. arXiv:2406.17753 [cs].

[36] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial, March 2024. URL http://arxiv.org/abs/2403.14380. arXiv:2403.14380 [cs].

[37] Astrid Schepman and Paul Rodway. The general attitudes towards artificial intelligence scale (gaais): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human–Computer Interaction*, 39(13):2724–2741, 2023. doi: 10.1080/10447318.2022.2085400. URL https://doi.org/10.1080/10447318.2022.2085400.

[38] Minkyu Shin and Jin Kim. Large Language Models Can Enhance Persuasion Through Linguistic Feature Alignment, February 2024. URL https://papers.ssrn.com/abstract=4725351.

[39] Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus*, 3(2):pgae035, February 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae035. URL https://doi.org/10.1093/pnasnexus/pgae035.

[40] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850, June 2023. doi: 10.1126/sciadv.adh1850. URL https://www.science.org/doi/10.1126/sciadv.adh1850. Publisher: American Association for the Advancement of Science.

[41] James B Stiff and Paul A Mongeau. *Persuasive communication*. Guilford Publications, 2016.

[42] Cassandra Teigen, Jens Koed Madsen, Nicole Lauren George, and Sayeh Yousefi. Persuasiveness of arguments with AI-source labels. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0), 2024. URL https://escholarship.org/uc/item/6t82g70v.

[43] Joseph Uscinski, Shane Littrell, and Casey Klofstad. The importance of epistemology for the study of misinformation. *Current opinion in psychology*, page 101789, 2024.

[44] Sion Yoon, Tae Eun Kim, and Yoo Jung Oh. Designing and Evaluating Multi-Chatbot Interface for Human-AI Communication: Preliminary Findings from a Persuasion Task, June 2024. URL http://arxiv.org/abs/2406.19648. arXiv:2406.19648 [cs].

[45] Yunhao Zhang and Renée Gosline. Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making*, 18:e41, January 2023. ISSN 1930-2975. doi: 10.1017/jdm.2023.37. URL https://www.cambridge.org/core/journals/judgment-and-decision-making/article/human-favoritism-not-ai-aversion-peoples-perceptions-and-bias-toward-generative-ai-human-experts 419C4BD9CE82673EAF1D8F6C350C4FA8.

[46] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–20, New York, NY, USA, April 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581318. URL https://dl.acm.org/doi/10.1145/3544548.3581318.