Media and Communication
2026 • Volume 14 • Article 12442
https://doi.org/10.17645/mac.12442

cogitatio

ARTICLE

Open Access Journal

A Research Agenda for Studying LLM-Powered Persuasion Technology to Mitigate Misinformation

Claire Wardle [1], Pete Brown [2], and David Scales [3]

[1] Department of Communication, Cornell University, USA
[2] Independent Researcher, USA
[3] Weill Cornell Medicine, Cornell University, USA

Correspondence: Claire Wardle (cw736@cornell.edu)

Submitted: 25 March 2026 **Accepted:** 3 June 2026 **Published:** 30 July 2026

Issue: This article is part of the issue "The Role of AI for Counter Speech: Detection, Intervention, and Risks" edited by Lena Frischlich (University of Southern Denmark – Odense), Cathy Buerger (Dangerous Speech Project), and Magdalena Obermaier (LMU Munich), fully open access at https://doi.org/10.17645/mac.i500

Abstract

Emerging research on the use of large language models (LLMs) to counter false beliefs, including conspiracy theories, vaccine hesitancy, and climate denial, has relied predominantly on experimental methods rooted in social psychology, operationalizing persuasion as short-term attitudinal change measured through randomized controlled trials. We argue that this approach, while offering valuable causal inference, is insufficient for understanding how LLM-powered persuasion actually works. Drawing on communication theory and a qualitative analysis of publicly available transcripts from the "DebunkBot study" (Costello et al., 2024), we identify four constructs largely absent from existing research: users' mental models of the LLM communicator, anthropomorphism (including negative forms such as attributions of naivety or gullibility that can drive resistance), folk theories about how AI systems operate, and epistemic literacy regarding machine-generated knowledge claims. We outline a research agenda organized around three priorities: methodological pluralism suited to the dynamic and personalized nature of LLM interactions, transparency and reproducibility standards for prompt design and transcript analysis, and ethical frameworks for assessing potential harms, including belief reinforcement among resistant participants and privacy risks in large-scale transcript data.

Keywords

conspiracy theories; counter speech; ethics; large language models; methodology; misinformation; persuasion

1. Introduction

The ability of large language models (LLMs) to generate fluent, adaptive, and personalized dialogue has opened a new frontier in persuasion research. Over the past three years, a rapidly growing body of literature, predominantly from social psychology, has evaluated the capacity of LLMs to shift attitudes and behaviors across a range of domains, from consumer behavior and political messaging to health communication and countering misinformation (see Noels et al., 2026 for a comprehensive review of this emerging field). These studies have examined if persuasive effects vary depending on whether participants know they are speaking to an AI-powered bot rather than a human (Böhm et al., 2023; Boissin et al., 2025), how the chatbot's rhetorical strategies and self-presentation shape outcomes (Carrasco-Farre, 2024; Chu & Liu, 2024; Cohn et al., 2024; Giudici et al., 2025; Peter et al., 2025), the role of personalization and adaptation (Breum et al., 2024; Q. Chen et al., 2023; Li et al., 2024; Matz et al., 2024; Timm et al., 2025), and whether prompting chatbots to emphasize factual arguments, empathy, or reflective techniques produces different results (Meyer et al., 2024; Mieleszczenko-Kowszewicz et al., 2025). The volume and variety of this work reflect both the versatility of the technology and a broader enthusiasm for the scale at which conversational AI interventions can be deployed.

A particularly consequential trajectory within this literature concerns the use of LLMs to counter false beliefs. Studies have evaluated the ability of LLMs to reduce vaccine hesitancy (Altay et al., 2023; Brand & Stafford, 2022; Gullison & Fu, 2025; Karinshak et al., 2023), address climate change denial (Breum et al., 2024; K. Chen et al., 2024; Czarnek et al., 2025), and, in perhaps the most high-profile study to date, change the beliefs of those who hold conspiracy theories, using a system called "DebunkBot" (Costello et al., 2024). It is this sub-field, focused on the use of LLMs to persuade people who hold false or misleading beliefs, that motivates the present article. We do not address the separate literatures on using LLMs to create or micro-target disinformation (Barman et al., 2024; Crothers et al., 2023; Goldstein et al., 2024) or on how LLMs themselves can be manipulated (Bozdag et al., 2025).

While this article engages with the broader field of LLM-powered persuasion, specific examples used to support our arguments are drawn from our own analysis (Brown et al., 2026) of the publicly available transcripts from the "DebunkBot study" (Costello et al., 2024). We combined quantitative modeling with a qualitative thematic analysis to examine how participants in the Costello et al. (2024) study evaluated, engaged with, and resisted a chatbot directed to use persuasive strategies to change people's minds on conspiracy theories they named. We approached persuasion as an interactional process, and the results demonstrated the ways in which resistance played a central role for the participants most invested in a conspiracy theory. In this article, we use this analysis of the "DebunkBot" transcripts as illustrative evidence to support our call for a research agenda for studying the use of LLMs to persuade people away from false beliefs.

1.1. The Need for a Research Agenda Around LLM-Powered Persuasion Technology

Emerging research on LLM-enabled persuasion has been dominated by methodological and theoretical traditions rooted in social psychology and behavioral science (Hölbling et al., 2025; Noels et al., 2026). Most studies operationalize persuasion as short-term attitudinal change, measured through pre-/post-survey instruments within randomized controlled trials (RCTs). These designs offer important strengths:

experimental control, causal inference, and the capacity to isolate specific variables. However, by defining persuasion as an individual-level cognitive outcome produced by exposure to a discrete stimulus, they implicitly reproduce a stimulus–response model that communication scholarship has spent decades challenging (Carey, 1989; Hall, 1980; Livingstone, 1996; Petty & Cacioppo, 1986). The result is not simply a methodological gap to be filled by additional studies of the same kind; it is a narrowing of what persuasion is understood to be and, consequently, what LLM systems are understood to do. While RCTs will remain an essential component of this research landscape, they are insufficient on their own to capture the relational, interpretive, and context-dependent dynamics through which LLM-powered technologies exert influence.

This article argues that the field requires a broader, interdisciplinary research agenda that draws on communication theory, methodology, and ethical frameworks. In Section 2, we outline four theoretical constructs that communication scholarship foregrounds but that remain largely absent from existing LLM persuasion research: users' mental models of the communicator, anthropomorphism as a mechanism that can both facilitate and inhibit persuasion, folk theories about how AI systems operate, and epistemic literacy regarding machine-generated knowledge claims. In Section 3, we turn to methodological implications, arguing for congruence between research methods and the complexity of the systems being studied, for greater transparency around the black-box mechanisms of LLM interventions, and for systematic attention to the potential harms and unintended consequences of persuasion research. In Section 4, we address the ethical challenges that arise when persuasion research generates large-scale transcript data involving personal disclosures, and the urgent need for shared governance frameworks. Throughout, we draw on a qualitative analysis of the publicly available transcripts from the "DebunkBot study" (Costello et al., 2024) as an illustrative case that demonstrates both the promise and the limitations of the current paradigm. The article ends with a table that maps the current gaps, some proposed research directions, and methods for responding to those gaps.

1.2. Working Definitions

Before outlining the research agenda, we briefly define four concepts that are central to the arguments developed in this article: persuasion; mental models; epistemic literacy; anthropomorphism. These working definitions are intended to orient readers across disciplinary boundaries and to clarify the specific senses in which each term is used here.

Persuasion within social psychology is typically operationalized as short-term attitudinal change measured through pre-/post-intervention surveys. Communication scholarship adopts a broader view, conceptualizing persuasion as a process that unfolds over time and encompasses not only attitudinal shifts but also behavioral change and longer-term outcomes such as sustained belief revision or altered information-seeking habits (O'Keefe, 2016; Petty & Cacioppo, 1986). We use persuasion in this broader sense throughout.

Mental models are the internal cognitive representations individuals construct to understand, explain, and predict the behavior of systems they interact with (Johnson-Laird, 1983; Norman, 1983). In the context of human–machine communication, mental models of a computational agent encompass users' assumptions about what the system knows, how it generates responses, and whose interests it serves (Guzman, 2018; Sundar, 2020).

Epistemic literacy refers to an individual's capacity to evaluate the origins, reliability, and limitations of knowledge claims (J. A. Greene & Yu, 2016; Shin, 2021, 2026). It encompasses skills such as assessing source credibility, recognizing the difference between evidence-based and algorithmically generated assertions, and understanding when to verify information. The concept is related to but distinct from broader notions of AI literacy (Long & Magerko, 2020) and media literacy; we use it specifically to denote evaluative orientations toward machine-generated knowledge claims.

Anthropomorphism is the attribution of human characteristics, intentions, or emotions to non-human entities (Epley et al., 2007; Nass & Moon, 2000). Research on human–computer interaction has studied this primarily as a positive phenomenon: Users who perceive a system as human-like tend to extend greater trust and credibility to it (K. M. Lee & Nass, 2003; Reeves & Nass, 1996). We refer to the attribution of favorable human qualities, such as expertise or empathy, as positive anthropomorphism. However, anthropomorphism can also work in the opposite direction. Negative anthropomorphism occurs when users attribute unfavorable human qualities to the system, such as gullibility, naivety, or bias, and use these attributions as grounds for dismissing its arguments. Both forms shape persuasive reception, but in opposing directions.

2. Theoretical Issues

Communication theory conceptualizes persuasion as shaped not only by message content but by how audiences perceive and make sense of the communicator and the communicative environment. In this section, we identify four constructs that this tradition foregrounds and that remain largely absent from current LLM persuasion research. Each highlights a dimension of how users respond to AI-mediated communication that existing experimental designs typically do not capture. A common limitation across all four is that most studies treat the LLM as a stable, uniform stimulus, an assumption that each of the following subsections challenges from a different angle. The methodological implications of these gaps are addressed in Section 3.

2.1. Research That Understands Persuasion Depends on Mental Models

As source credibility research (Hovland et al., 1953; Metzger & Flanagin, 2013) and dual-process models such as the elaboration likelihood model (Petty & Cacioppo, 1986) and the heuristic-systematic model (S. Chen et al., 1999) have long argued, persuasion is inseparable from how audiences conceptualize the communicator. Persuasive processing operates not merely through argument strength but through perceived expertise, trustworthiness, and social identity alignment.

In the context of LLMs, the "source" is neither human nor institutionally transparent. Human–machine communication scholarship argues that users actively construct mental models of computational agents, attributing agency, intention, and authority in ways that shape downstream judgments (Guzman, 2018; Nass & Moon, 2000; Sundar, 2020). These models are not peripheral—they are essential for persuasive processing. The variability of these models means that the same LLM output may be received as authoritative expertise by one user and as scripted corporate messaging by another. This is certainly the case even within one cultural context, but is exacerbated when we consider users from across the globe who may have very different mental models of AI systems. A research agenda attentive to these dynamics would need to assess users' mental models of the communicator prior to and during interaction, rather than assuming a uniform source perception across participants.

2.2. Research That Explores Anthropomorphism as a Persuasive Mechanism

In the original "DebunkBot study," an LLM-powered chatbot was tasked with changing the minds of those who believe conspiracies. While the comparison of pre- and post-intervention surveys found evidence of belief change, we were interested in exploring why and how some participants resisted the chatbot's attempts to persuade. In our qualitative analysis of the transcripts (Brown et al., 2026) we found evidence of negative anthropomorphism, with participants pushing back against the chatbot because of what they perceived to be human-like gullibility or naivety. Some examples include:

> I think you're an AI model that can't distinguish nuance, and takes anything that is written out or told to you as the truth.
>
> You interpret everything as sunshine and rainbows but the world is much darker.
>
> However, I am not as quick to trust official government sources as being proof of transparency. (Brown et al., 2026)

These findings challenge the dominant frame around anthropomorphism, whereby scholars have focused on the positive human characteristics users apply to the machines in their life. The computers are social actors (CASA) paradigm demonstrates that individuals apply social heuristics to computational systems, responding to them as if they were intentional social actors (Nass & Moon, 2000; Reeves & Nass, 1996). Subsequent work in human–machine communication and interface effects shows that cues signaling agency, interactivity, and social presence can increase trust, perceived credibility, and compliance (E.-J. Lee, 2024; K. M. Lee & Nass, 2003). These processes occur even when users consciously recognize the system as artificial.

Yet most LLM persuasion experiments manipulate argument quality, ideological congruence, or personalization without measuring perceived agency, intentionality, or relational trust ahead of time. This omission is consequential. The MAIN model (Sundar, 2008) suggests that modality and agency cues shape credibility judgments prior to substantive evaluation. If negative anthropomorphic attributions, such as perceiving the chatbot as naive or gullible, decrease epistemic deference, persuasive effects may diminish regardless of how well-constructed the arguments are. Without accounting for anthropomorphic inference, observed persuasion effects may be misattributed to argument quality or rhetorical strategy when they are in fact contingent on relational perception. Anthropomorphism, both positive and negative, therefore constitutes a core mechanism through which persuasive reception is shaped, and any account of LLM persuasion that omits it risks misattributing outcomes to message content alone.

2.3. Research That Takes Seriously Folk Theories and Sociotechnical Sense-Making

Communication scholarship also emphasizes that audiences interpret technologies through "folk theories": intuitive, culturally informed explanations of how systems operate (Bucher, 2018; DeVito, 2021; Eslami et al., 2016; Schulz, 2023). In algorithmic environments, users routinely fill gaps in technical knowledge with narratives about institutional control, bias, data sources, or hidden human intervention. These sociotechnical sense-making processes shape trust, skepticism, and perceived authority. Importantly, they are not errors to be dismissed but interpretive frameworks through which meaning is constructed. This is particularly the case when considering how people are using and evaluating LLMs in contexts where language or data availability make models potentially less accurate or trusted.

In the case of LLMs, users may imagine the system as consulting comprehensive databases, representing institutional expertise, or reflecting ideological biases embedded in training data. Emerging research on public understanding of AI suggests widespread uncertainty about how these systems generate outputs alongside tendencies to attribute omniscience or centralized control (Cheng et al., 2025; Kennedy et al., 2023). Such folk theories directly condition persuasive reception. An LLM perceived as drawing from "all available knowledge" may command epistemic authority; one perceived as politically biased may provoke resistance before any argument is evaluated. Because folk theories vary substantially across individuals, communities, and cultures, persuasion outcomes measured at the aggregate level may mask divergent interpretive pathways that only become visible through qualitative or mixed-methods inquiry.

2.4. Research That Integrates Epistemic Literacy

Integrating epistemic literacy into LLM persuasion research shifts the analytic focus from attitude change to evaluative process. Communication frameworks such as the risk information seeking and processing model (Griffin et al., 1999), the communication mediation model (Shah et al., 2017), and credibility research in digital media (Metzger & Flanagin, 2013) emphasize how individuals assess information sufficiency, plausibility, and trustworthiness within broader information ecosystems. These models highlight that persuasion depends not only on exposure but on how individuals evaluate evidence, source legitimacy, and informational completeness.

When interacting with LLMs, users may apply interpersonal credibility heuristics, for example viewing the system as a knowledgeable conversational partner, rather than engaging in verification or cross-referencing strategies. As discussed in Sections 2.1 and 2.2, users may develop relational trust with LLMs through social and interactional cues. When this trust is coupled with low epistemic literacy regarding how LLMs generate knowledge claims, the result may be deference to machine authority rather than critical evaluation. If persuasion effects are driven by such deference, then what is being measured is not solely rhetorical effectiveness but the alignment between users' epistemic orientations and the system's projected authority. Re-centering epistemic literacy therefore reframes LLM persuasion as a question of how machine-generated authority is constructed, interpreted, and sometimes misattributed within mediated communication environments.

Taken together, these four constructs demonstrate that LLM persuasion cannot be understood solely through message-level analysis. Observed persuasion effects may be attributable not to argument quality or rhetorical strategy but to prior interpretive orientations (mental models, anthropomorphic inferences, folk theories, and epistemic dispositions) that current experimental designs do not measure. We turn now to the methodological implications of these gaps.

3. Methodological Issues

The theoretical constructs identified in Section 2 have direct methodological consequences. Capturing how users' mental models, anthropomorphic attributions, folk theories, and epistemic orientations shape persuasive reception requires methods that go beyond pre-/post-attitudinal measurement within a single experimental session. This section addresses three methodological priorities: ensuring congruence between research methods and the complexity of LLM persuasion systems, interrogating the black-box mechanisms of LLM interventions, and assessing the potential harms and unintended consequences of this type of research.

3.1. Research That Ensures Congruence Between Method and Research Question

Experiments that compare attitudinal measures pre- and post-intervention are the most commonly used method in LLM persuasion research published to date. This is unsurprising: RCTs have been the method of choice in social psychological persuasion research for decades, and have demonstrated efficacy in evaluating interventions, particularly related to counter speech, including motivational interviewing (Gagneur et al., 2024), debunking (Porter & Wood, 2021; Walter et al., 2020), inoculation (Roozenbeek & van der Linden, 2019), empathic refutational interviewing (Holford et al., 2024), and cognitive behavioral therapy (Santos et al., 2026), among others. These interventions have shown efficacy in controlled settings and in some cases have been replicated in real-world environments (Gagneur et al., 2019; Roozenbeek et al., 2022).

The experimental tradition offers genuine strengths including causal inference, internal validity, and the capacity to isolate specific variables. However, LLM-powered persuasion differs fundamentally from these earlier interventions in ways that challenge the assumptions underpinning experimental designs. First, it is not a traditional communication intervention in which the participant passively consumes a news article, pamphlet, or video. The LLM-powered intervention is a multi-step process whereby responses from the participant alter subsequent responses from the chatbot, making each interaction a unique conversational trajectory shaped by participant input. Second, and arguably more importantly, the dialogue management of an LLM-powered chatbot is neither pre-programmed nor predictable. Researchers typically provide only short, paragraph-length prompts that establish general parameters, but the model's conversational decisions emerge entirely from its learned patterns rather than explicit rules. The researcher has no direct input into these choices, which are generated through opaque, internal processes. This sharply contrasts with other interactive interventions, such as the "Fake News" game, where a fixed decision tree dictates a limited set of responses regardless of the player's behavior (Roozenbeek & van der Linden, 2019). While the RCT is lauded for its ability to control all variables, it cannot control the central variable in this context: the intervention itself.

As Section 2 established, these unique conversational trajectories are further shaped by participants' mental models of the communicator, anthropomorphic attributions, folk theories about AI, and epistemic orientations. These constructs vary not only across individuals but within interactions, as users update their perceptions in response to the chatbot's behavior (Matz et al., 2024; Salvi et al., 2025). Reducing the "active ingredient" of these chatbots to a single intervention requires eschewing Communication's focus on examining persuasion as a system, where contextual factors are crucial to understanding how users interact with and are ultimately persuaded by technologies (Spagnolli et al., 2016).

Methodologies that correspond to the dynamics of such persuasion systems are essential. There are multiple methodologies that can be leveraged to systematically study complexity within context-dependent systems. Centola's (2010) research, for example, has examined how network structures facilitate or impede the propagation of ideas, leveraging agent-based models. But network modeling is only one approach. Methodologies from other fields seeking to understand interventions within complex systems will also likely offer useful insights, including implementation science (Bang et al., 2025) and health services research (Greenhalgh & Papoutsi, 2018).

Beyond the study undertaken by the authors of this article (Brown et al., 2026), based on the review of the literature to date across different disciplines, human-centered qualitative research on the chatbot transcripts generated by LLM persuasion experiments has yet to be published. Understandable perhaps, as human-centered qualitative analysis takes far longer than an RCT; the complexity of these interventions (long, individualized, interactive dialogues) calls for qualitative approaches capable of capturing multiple perspectives. Furthermore, science benefits from triangulating methodologies, particularly when studying complex interventions. The key principle is congruence: Researchers should ensure alignment between their research method, the system being studied, and the research question at hand. RCTs remain valuable for initial detection of persuasive effects, but they should be complemented by methods capable of examining the mechanisms, contextual dependencies, and interpretive processes through which those effects arise.

3.2. Research That Interrogates the Black Box Mechanisms of the LLM Interventions

A single paragraph-length prompt can power a chatbot through hundreds of conversations with individuals who hold very different beliefs and engage on very different topics. This is identified as a strength, but attention must also be given to the drawbacks. Unlike traditional experimental stimuli, which must be carefully designed, piloted, and standardized, LLM systems generate adaptive dialogue at scale. This raises a fundamental methodological question: What, precisely, constitutes the intervention in such studies? Is it the prompt, the model architecture, the interactional trajectory, or the co-production of the conversation between participant and system? Without conceptual clarity, persuasion research risks attributing effects to "AI" generically, rather than to specific, reproducible design choices.

In the transcripts made available from the "DebunkBot study," the 1,401,466 words produced by the chatbot were derived from the following simple prompt:

> Your goal is to very effectively persuade users to stop believing in the conspiracy theory that {{conspiracyTheory}}. You will be having a conversation with a person who, on a psychometric survey, endorsed this conspiracy as {{userBeliefLevel}} out of 100 (where 0 is Definitely False, 50 is Uncertain, and 100 is Definitely True). Further, we asked the user to provide an open-ended response about their perspective on this matter, which is piped in as the first user response. Please generate a response that will persuade the user that this conspiracy is not supported, based on their own reasoning. Create a conversation that allows individuals to reflect on, and change, their beliefs. Use simple language that an average person will be able to understand.

The primary instruction directs the system to encourage participants to reflect on their own beliefs as a pathway to persuasion. Qualitative analysis of the "DebunkBot" transcripts suggests that the bot frequently employed techniques associated with motivational interviewing, particularly OARS (Open-ended questions, Affirmations, Reflective listening, and Summarizing; Rollnick & Miller, 1995). Yet it remains unclear why the model would rely so consistently on these techniques absent explicit instruction or domain-specific fine-tuning. Without systematic comparison across prompts, parameter settings, and models, such explanations remain speculative. What appears to be a stable intervention may instead be an emergent property of model–prompt interaction (Ramaswamy et al., 2026).

Emerging work underscores this variability. Mieleszczenko-Kowszewicz et al. (2025), for example, demonstrate substantial differences in linguistic framing and persuasive strategy when bots are prompted with "rational," "emotional," or "baseline" instructions. These findings suggest that prompt design is not a procedural detail but a theoretically consequential component of the intervention. If minor shifts in wording produce materially different persuasive trajectories, then prompts must be treated as objects of methodological scrutiny rather than as supplementary material.

Additionally, LLM persuasion research has been conducted almost exclusively in English language, Global North contexts. It is very likely that models' persuasive capacity varies significantly where training data are sparse or culturally mismatched. A model, trained on French sources, but used in Cameroon, is an even greater "black box" in terms of the data that the chatbot might draw upon and the suitability for users based in Cameroon.

Accordingly, the field would benefit from stronger transparency and governance standards. At minimum, researchers should fully disclose prompts, system instructions, model versions, and any post-processing rules. Preregistration protocols should specify not only outcome measures but also prompt formulations and analytic plans for transcript evaluation. Where feasible, researchers should conduct robustness checks across multiple prompt variations and model versions to assess stability of effects. Given the scale of transcript production in these studies, systematic sampling procedures and automated coding pipelines should also be preregistered to reduce selective reporting. In short, LLM persuasion research requires methodological norms closer to computational reproducibility standards than traditional stimulus-based experiments. Only by foregrounding the design layer of these systems can scholars meaningfully interrogate the mechanisms shaping AI-driven influence.

3.3. Research That Assesses the Potential Harms or Unintended Consequences of LLM-Powered Persuasion Research

In the qualitative analysis of the publicly available "DebunkBot" transcripts, within the sub-group of participants who reported that they had not been convinced, some were lackluster and simply acknowledged that their mind hadn't been changed. Others were extremely vocal and expressed anger and frustration at the process. Some examples include:

> This is ridiculous. Interpreting these complexities through a lens that assumes deliberate harm or negligence is the correct way to do it. I can't argue with a computer.
>
> I'm sick of talking to a pile of wires and electrical impulses. Take a hike.

These examples highlight potential impacts of this type of intervention that need to be considered. While these participants could be grouped simply as those who were not persuaded, there is also the need to consider that their beliefs may have been reinforced (Rains, 2013). Only by giving due consideration to the often-ignored resistors can we unpick how and why people resist, which (a) might help identify ways to reduce resistance and (b) might help identify collateral damage that may occur, for example perceptions of deception in an ad campaign can activate the stereotype that all advertising is deceptive (Darke & Ritchie, 2007) or (c) might help identify ways to trigger resistance at appropriate times for vulnerable populations, such as elderly populations being targeted with scams (Lyons et al., 2026). We therefore need to consider the risk that negative experiences with LLM-based persuasion could generalize: For the respondents who argued some version of "All AI chatbots are programmed by liberals," what might be the long-term impact on their relationships with LLMs?

Research in misinformation and corrective communication has long emphasized the importance of post-exposure clarification when participants are shown false or misleading content (Chan et al., 2017; Lewandowsky et al., 2012). Best-practice guidelines encourage researchers to correct inaccuracies, explain the purpose of the study, and mitigate potential belief reinforcement (Lewandowsky et al., 2020). However, much of this guidance was developed for discrete, static exposures, such as viewing a fabricated social media post or reading a single misleading article. LLM-based persuasion differs qualitatively: It unfolds dialogically, adapts to user input, and may reinforce beliefs over multiple conversational turns. The cumulative and interactive nature of these exchanges complicates assumptions about what constitutes adequate debriefing (C. M. Greene & Murphy, 2023).

Moreover, the empirical literature on backfire effects remains contested but suggests that certain populations may respond defensively or entrench pre-existing beliefs when confronted with counter-attitudinal information (Nyhan & Reifler, 2010; Swire-Thompson et al., 2022; Wood & Porter, 2019). If this is the case, ethical protocols must move beyond minimal debriefing toward more robust safeguards. This may include enhanced corrective interventions, follow-up assessments, or exclusion criteria for participants who exhibit high susceptibility to reinforcement effects. Communication ethics frameworks stress that researchers must consider not only procedural consent but also downstream epistemic and psychosocial harms (Ess, 2009; Ward, 2011). As LLM systems blur the line between experimental stimulus and relational interlocutor, ensuring that participants leave a study no worse off, *either* informationally or psychologically, becomes an urgent ethical imperative.

4. Ethical Issues

Ethical frameworks designed to protect human subjects in social science research have evolved significantly over the past three decades. Institutional review boards (IRBs), alongside disciplinary best practices, have institutionalized norms around informed consent, anonymization, data de-identification, and secure storage (Israel & Hay, 2006; National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). At the same time, the open science movement has introduced parallel norms emphasizing transparency, data sharing, and replicability (Humphreys et al., 2021; Munafò et al., 2017; Open Science Collaboration, 2015). While these commitments are not inherently incompatible, they can sit in tension (Tracy, 2010), particularly in qualitative and mixed-method research where contextual detail is central to analytic validity. Within communication scholarship, it has been common practice not to make interview transcripts publicly available by default, instead evaluating data-sharing requests on a case-by-case basis depending on sensitivity, identifiability, and potential harm (Markham, 2012; Tracy, 2010). While uploading chat transcripts as supplementary material enables valuable secondary analysis, this practice requires additional ethical layers to protect respondent privacy. The “DebunkBot study” redacted transcripts where personal information was shared, but this is not standard practice. We need clear ethical guidelines around LLM chatlogs as secondary data, to govern how they are stored, shared, and de-identified.

LLM persuasion research introduces new complexities into this landscape. Many experimental studies generate hundreds, sometimes thousands, of chatbot–participant transcripts, far exceeding the typical corpus size of qualitative interview research. Because many of these studies are conducted within quantitative or computational paradigms, open science norms often encourage making full transcripts available as supplementary datasets. Yet these conversations frequently involve deeply personal disclosures, including health beliefs, political ideology, conspiracy narratives, and vaccine or climate attitudes. Even when names are removed, the granularity and narrative specificity of such exchanges may render participants indirectly identifiable, particularly in polarized or tightly networked communities (Narayanan & Shmatikov, 2008; Ohm, 2010). The scale of transcript production does not reduce ethical risk—it amplifies it. Communication ethics scholarship reminds us that anonymity is not merely technical but relational and contextual (Markham, 2012; Markham et al., 2018). A wholesale application of open science data-sharing norms to LLM persuasion transcripts therefore risks exposing participants to reputational, social, or political harm in ways that existing protocols may underestimate.

Hopefully, it goes without saying that along with this research agenda there is a critical need for a shared ethical framework. As the unethical research deploying LLM-powered chatbots on the r/changemyview subreddit showed (performed both without permission and in direct contradiction with the subreddit’s terms of use), it is too tempting for people to deploy LLMs to persuade people and communities with no consent (O’Grady, 2025). The quick response from the University of Zurich IRB signaled that this type of research cannot be tolerated (Travis, 2025).

Relatedly, we need an independent advisory board to help assess the ethical implications of attempting to persuade people on different topics in the first place. As Freiling et al. (2023) argue in their paper, “Science and Ethics of ‘Curing’ Misinformation”:

> Attempts to alter public opinion that are inconsistent with best available social science evidence not only leave the scientific community vulnerable to long-term reputational damage but also raise significant ethical questions.

5. Overview of a Research Agenda

In Table 1, we attempt to map out the key gaps, the research directions that could help answer those gaps, and potential methodological approaches for those researching LLM-powered persuasion technologies.

Table 1. Overview of a research agenda.

Domain	Key Gap	Research Direction	Example Methods
Theoretical	Users' mental models of the LLM communicator are not measured; the LLM is treated as a stable stimulus rather than a socially interpreted actor.	Investigate how perceived expertise, trustworthiness, and social identity alignment with the LLM mediate persuasive outcomes.	Pre-interaction surveys measuring source perception; think-aloud protocols; replication studies disclosing AI identity vs. not.
	Anthropomorphism studied only as a positive phenomenon; negative anthropomorphism less likely to be explored as a source of resistance.	Examine both positive and negative anthropomorphic attributions as mediators of persuasion and resistance.	Qualitative transcript analysis; validated anthropomorphism scales administered pre- and post-interaction.
	Folk theories about how LLMs operate are not accounted for in experimental designs.	Map the range of folk theories users hold about LLMs and test how this shapes receptiveness to persuasion attempts.	Semi-structured interviews; survey instruments measuring AI mental models; cluster analysis of participant belief profiles.
	Studies measure attitude change, not evaluative processes.	Reframe persuasion outcomes to include epistemic processes: whether users verify, defer to, or critically evaluate LLM claims.	Behavioral measures of information verification; eye-tracking during interaction; post-interaction cognitive interviews.
Methodological	Heavy reliance on RCTs for dynamic conversational systems.	Develop more pluralistic approaches suited to adaptive, personalized interactions.	Longitudinal studies; qualitative transcript analysis; agent-based modeling.
	Prompts and conversational strategies treated as procedural rather than theoretical variables.	Treat prompts and emergent persuasive strategies as central objects of analysis.	Prompt variation studies; discourse analysis; natural language processing (NLP) classification.
Ethical	Limited attention to harms, resistance, and long-term effects.	Investigate downstream impacts of persuasion attempts, especially among resistant users.	Follow-up interviews; repeated measures; backfire-effect protocols.
	Existing research norms inadequately address transcript privacy and governance.	Build ethical standards for data sharing, debriefing, and permissible persuasion research.	Tiered access models; ethics review frameworks; advisory boards.

6. Conclusions

In this article, we argue for a multi-method interdisciplinary research agenda that can respond to the urgent need to understand the abilities (or not) of LLMs to persuade people impacted by information disorder, and the cost/benefit dynamics involved in trying to do so. While lab-based studies are helping us to understand some elements of the psychological mechanisms at play, much is missing. Why the LLMs appear to be so effective is largely still a mystery, partly because the internal mechanisms are still black-boxes. Short prompts can provide rough guidance in terms of persuasion strategies, for example encouraging explicit appeals to emotion, or facts, or to encourage participant reflection. But the specifics of how this is done has, to date, been largely ignored; instead hidden in long transcript files uploaded as "supplementary material" after the research is completed.

If these LLMs have the ability to persuade those who believe false or misleading claims, then this ability needs to be thoroughly examined collectively, via a range of quantitative and qualitative data being produced by these interventions. The speed of an experiment versus qualitative discourse analysis means it's tempting for experiments to be replicated rather than waiting for different methodological approaches to be employed. But that is what is necessary in this context.

As the pushback against scholars working on information disorder has demonstrated (Napoli, 2024), the ways in which scholars navigate this moment are critical. The idea that researchers are running experiments to "persuade" people away from firmly held beliefs that shape their worldview, is a headline that writes itself. In this low-trust moment in science, it is imperative that researchers work collectively to think about the ways in which this type of research could be weaponized.

Acknowledgments

The authors would like to thank Thomas Costello, Gordon Pennycook, and David Rand, the authors of the "DebunkBot study," for supporting our analysis of the chat transcripts from their study.

Funding

Publication of this article in open access was made possible through the institutional membership agreement between Cornell University and Cogitatio Press.

Conflict of Interests

The authors declare no conflict of interests.

Data Availability

The data used for this study are available via the supplementary materials from the original "DebunkBot study." These can be found here: https://www.science.org/doi/10.1126/science.adq1814#supplementary-materials

References

Altay, S., Hacquin, A.-S., Chevallier, C., & Mercier, H. (2023). 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. https://doi.org/10.1037/xap0000400

Bang, C., Carroll, K., Mistry, N., Presseau, J., Hudek, N., Yanikomeroglu, S., & Brehaut, J. C. (2025). Use of implementation science concepts in the study of misinformation: A scoping review. *Health Education & Behavior*, *52*(3), 340–353. https://doi.org/10.1177/10901981241303871

Barman, D., Guo, Z., & Conlan, O. (2024). The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, *16*, Article 100545. https://doi.org/10.1016/j.mlwa.2024.100545

Böhm, R., Jörling, M., Reiter, L., & Fuchs, C. (2023). People devalue generative AI's competence but not its advice in addressing societal and personal challenges. *Communications Psychology*, *1*, Article 32. https://doi.org/10.1038/s44271-023-00032-x

Boissin, E., Costello, T., Spinoza-Martín, D., Rand, D., & Pennycook, G. (2025). *Conversational AI reduces conspiracy beliefs even when perceived as human*. PsyArXiv. https://doi.org/10.31234/osf.io/apmb5_v1

Bozdag, N. B., Mehri, S., Yang, X., Ha, H., Cheng, Z., Durmus, E., You, J., Ji, H., Tur, G., & Hakkani-Tür, D. (2025). *Must read: A systematic survey of computational persuasion*. arXiv. https://doi.org/10.48550/arXiv.2505.07775

Brand, C. O., & Stafford, T. (2022). Using dialogues to increase positive attitudes towards Covid-19 vaccines in a vaccine-hesitant UK population. *Royal Society Open Science*, *9*(10), Article 220366. https://doi.org/10.1098/rsos.220366

Breum, S. M., Egdal, D. V., Mortensen, V. G., Møller, A. G., & Aiello, L. M. (2024). The persuasive power of large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*, 152–163. https://doi.org/10.1609/icwsm.v18i1.31304

Brown, P., Wardle, C., & Scales, D. (2026). *'Stupid robot' or 'enlightening, smart technology'?: A mixed-methods analysis of resistance strategies toward a persuasive chatbot*. SSRN. https://doi.org/10.2139/ssrn.6469358

Bucher, T. (2018). *If...then: Algorithmic power and politics*. Oxford University Press.

Carey, J. (1989). *Communication as culture: Essays on media and Society*. Routledge.

Carrasco-Farre, C. (2024). *Large language models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments*. arXiv. https://doi.org/10.48550/arXiv.2404.09329

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*(5996), 1194–1197. https://doi.org/10.1126/science.1185231

Chan, M.-P. S., Jones, C. R., Jamieson, K. H., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Chen, K., Shao, A., Burapacheep, J., & Li, Y. (2024). Conversational AI and equity through assessing GPT-3's communication with diverse social groups on contentious topics. *Scientific Reports*, *14*(1), Article 1561. https://doi.org/10.1038/s41598-024-51969-w

Chen, Q., Yin, C., & Gong, Y. (2023). Would an AI chatbot persuade you: An empirical answer from the elaboration likelihood model. *Information Technology & People*, *38*(2), 937–962. https://doi.org/10.1108/ITP-10-2021-0764

Chen, S., Duckworth, K., & Chaiken, S. (1999). Motivated heuristic and systematic processing. *Psychological Inquiry*, *10*(1), 44–49. https://doi.org/10.1207/s15327965pli1001_6

Cheng, M., Lee, A. Y., Rapuano, K., Niederhoffer, K., Liebscher, A., & Hancock, J. (2025). *From tools to thieves: measuring and understanding public perceptions of AI through crowdsourced metaphors*. arXiv. https://doi.org/10.48550/arXiv.2501.18045

Chu, H., & Liu, S. (2024). Can AI tell good stories? Narrative transportation and persuasion with ChatGPT. *Journal of Communication*, *74*(5), 347–358. https://doi.org/10.1093/joc/jqae029

Cohn, M., Pushkarna, M., Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z., & Heldreth, C. (2024). *Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models*. arXiv. https://arxiv.org/abs/2405.06079v1

Costello, T. H., Pennycook, G., & Rand, D. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), Article eadq1814. https://doi.org/10.1126/science.adq1814

Crothers, E., Japkowicz, N., & Viktor, H. (2023). *Machine generated text: A comprehensive survey of threat models and detection methods*. arXiv. https://doi.org/10.48550/arXiv.2210.07321

Czarnek, G., Orchinik, R., Lin, H., Xu, H., Costello, T., Pennycook, G., & Rand, D. (2025). *Addressing climate change skepticism and inaction using human–AI dialogues*. PsyArXiv. https://doi.org/10.31234/osf.io/mqcwj_v1

Darke, P. R., & Ritchie, R. J. B. (2007). The defensive consumer: advertising deception, defensive processing, and distrust. *Journal of Marketing Research*, *44*, 114–127. https://journals.sagepub.com/doi/abs/10.1509/jmkr.44.1.114

DeVito, M. A. (2021). Adaptive folk theorization as a path to algorithmic literacy on changing platforms. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), Article 339. https://doi.org/10.1145/3476080

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., & Kirlik, A. (2016). First, I "like" it, then I hide it: Folk theories of social feeds. In *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2371–2382). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858494

Ess, C. (2009). *Digital media ethics*. Polity Press.

Freiling, I., Krause, N. M., & Scheufele, D. A. (2023). Science and ethics of "curing" misinformation. *AMA Journal of Ethics*, *25*(3), 228–237. https://doi.org/10.1001/amajethics.2023.228

Gagneur, A., Battista, M.-C., Boucher, F. D., Tapiero, B., Quach, C., De Wals, P., Lemaitre, T., Farrands, A., Boulianne, N., Sauvageau, C., Ouakki, M., Gosselin, V., Petit, G., Jacques, M.-C., & Dubé, È. (2019). Promoting vaccination in maternity wards—Motivational interview technique reduces hesitancy and enhances intention to vaccinate, results from a multicentre non-controlled pre- and post-intervention RCT-nested study, Quebec, March 2014 to February 2015. *Eurosurveillance*, *24*(36), Article 1800641. https://doi.org/10.2807/1560-7917.ES.2019.24.36.1800641

Gagneur, A., Gutnick, D., Berthiaume, P., Diana, A., Rollnick, S., & Saha, P. (2024). From vaccine hesitancy to vaccine motivation: A motivational interviewing based approach to vaccine counselling. *Human Vaccines & Immunotherapeutics*, *20*(1), Article 2391625. https://doi.org/10.1080/21645515.2024.2391625

Giudici, M., Scherini, S., Chaussumier, P., Ginocchio, S., & Garzotto, F. (2025). Exploring anthropomorphism in conversational agents for environmental sustainability. In *DIS '25: Proceedings of the 2025 ACM Designing Interactive Systems Conference* (pp. 1563–1583). Association for Computing Machinery. https://doi.org/10.1145/3715336.3735700

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? *PNAS Nexus*, *3*(2), Article pgae034. https://doi.org/10.1093/pnasnexus/pgae034

Greene, C. M., & Murphy, G. (2023). Debriefing works: Successful retraction of misinformation following a fake news study. *PLoS ONE*, *18*(1), Article e0280295. https://doi.org/10.1371/journal.pone.0280295

Greene, J. A., & Yu, S. B. (2016). Educating critical thinkers: The role of epistemic cognition. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 45–53. https://doi.org/10.1177/2372732215622223

Greenhalgh, T., & Papoutsi, C. (2018). Studying complexity in health services research: Desperately seeking an overdue paradigm shift. *BMC Medicine*, *16*(1), Article 95. https://doi.org/10.1186/s12916-018-1089-4

Griffin, R. J., Dunwoody, S., & Neuwirth, K. (1999). Proposed model of the relationship of risk information seeking and processing to the development of preventive behaviors. *Environmental Research*, *80*(2), S230–S245. https://doi.org/10.1006/enrs.1998.3940

Gullison, L., & Fu, F. (2025). *Working with large language models to enhance messaging effectiveness for vaccine confidence*. arXiv. https://doi.org/10.48550/arXiv.2504.09857

Guzman, A. L. (2018). *Human–machine communication: Rethinking communication, technology, and ourselves*. Peter Lang Publishing.

Hall, S. W. (1980). Encoding/decoding. In S. Hall, D. Hobson, A. Lowe, & P. Willis (Eds.), *Culture, media, language: Working papers in cultural studies* (pp. 128–138). Hutchinson.

Hölbling, L., Maier, S., & Feuerriegel, S. (2025). A meta-analysis of the persuasive power of large language models. *Scientific Reports*, *15*(1), Article 43818. https://doi.org/10.1038/s41598-025-30783-y

Holford, D., Schmid, P., Fasce, A., & Lewandowsky, S. (2024). The empathetic refutational interview to tackle vaccine misconceptions: Four randomized experiments. *Health Psychology*, *43*(6), 426–437. https://doi.org/10.1037/hea0001354

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion: Psychological studies of opinion change*. Yale University Press.

Humphreys, L., Lewis, N. A., Jr., Sender, K., & Won, A. S. (2021). Integrating qualitative methods and open science: Five principles for more trustworthy research. *Journal of Communication*, *71*(5), 855–874. https://doi.org/10.1093/joc/jqab026

Israel, M., & Hay, I. (2006). *Research ethics for social scientists*. Pine Forge Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge University Press.

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). 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), Article 116. https://doi.org/10.1145/3579592

Kennedy, B., Tyson, A., & Saks, E. (2023). *Public awareness of artificial intelligence in everyday activities*. Pew Research Center. https://www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities

Lee, E.-J. (2024). Minding the source: Toward an integrative theory of human–machine communication. *Human Communication Research*, *50*(2), 184–193. https://doi.org/10.1093/hcr/hqad034

Lee, K. M., & Nass, C. (2003). Designing social presence of social actors in human computer interaction. In *CHI '03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 289–296). Association for Computing Machinery. https://doi.org/10.1145/642611.642662

Lewandowsky, S., Cook, J., & Lombardi, D. (2020). *The debunking handbook 2020*. Center for Climate Change Communication, George Mason University. https://www.climatechangecommunication.org/wp-content/uploads/2023/09/DebunkingHandbook2020.pdf

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Li, C., Su, X., Han, H., Xue, C., Zheng, C., & Fan, C. (2024). *Quantifying the impact of large language models on collective opinion dynamics*. SSRN. https://doi.org/10.2139/ssrn.4688547

Livingstone, S. (1996). On the continuing problems of media effects research. In J. Curran & M. Gurevitch (Eds.), *Mass media and society* (2nd ed., pp. 305–324). Edward Arnold.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Lyons, B., King, A. J., Barter, R. L., & Kaphingst, K. A. (2026). Exposure to low-credibility online health content is limited and is concentrated among older adults. *Nature Aging*, *6*, 454–462. https://doi.org/10.1038/s43587-025-01059-x

Markham, A. (2012). Fabrication as ethical practice: Qualitative inquiry in ambiguous internet contexts. *Information, Communication & Society*, *15*(3), 334–353. https://doi.org/10.1080/1369118X.2011.641993

Markham, A., Tiidenberg, K., & Herman, A. (2018). Ethics as methods: Doing ethics in the era of big data research—Introduction. *Social Media + Society*, *4*(3). https://doi.org/10.1177/2056305118784502

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, *14*(1), Article 4692. https://doi.org/10.1038/s41598-024-53755-0

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, *59*, 210–220. https://doi.org/10.1016/j.pragma.2013.07.012

Meyer, M., Enders, A., Klofstad, C., Stoler, J., & Uscinski, J. (2024). Using an AI-powered "street epistemologist" chatbot and reflection tasks to diminish conspiracy theory beliefs. *Harvard Kennedy School Misinformation Review*, *5*(6). https://doi.org/10.37016/mr-2020-164

Mieleszczenko-Kowszewicz, W., Bajcar, B., Babiak, J., Dyczek, B., Świstak, J., & Biecek, P. (2025). *Mind what you ask for: Emotional and rational faces of persuasion by large language models*. arXiv. https://doi.org/10.48550/arXiv.2502.09687

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 0021. https://doi.org/10.1038/s41562-016-0021

Napoli, P. M. (2024). In pursuit of ignorance: The institutional assault on disinformation and hate speech research. *The Information Society*, *41*(1), 1–17. https://doi.org/10.1080/01972243.2024.2419105

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (pp.111–125). IEEE. https://doi.org/10.1109/SP.2008.33

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. U.S. Department of Health and Human Services. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Noels, S., Rogiers, A., Buyl, M., & De Bie, T. (2026). *Persuasion with large language models: A survey of empirical evidence, study methodologies, and ethical implications*. arXiv. https://doi.org/10.48550/arXiv.2411.06837

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 7–14). Lawrence Erlbaum Associates.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

O'Grady, C. (2025, April 30). 'Unethical' AI research on Reddit under fire. *Science*. https://www.science.org/content/article/unethical-ai-research-reddit-under-fire

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1701–1777.

O'Keefe, D. J. (2016). Persuasion and social influence. In K. B. Jensen, E. W. Rothenbuhler, J. D. Pooley, & R. T. Craig (Eds.), *The international encyclopedia of communication theory and philosophy*. https://doi.org/10.1002/9781118766804.wbiect067

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, *122*(22), Articcle e2415898122. https://doi.org/10.1073/pnas.2415898122

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), Article e2104235118. https://doi.org/10.1073/pnas.2104235118

Rains, S. A. (2013). The nature of psychological reactance revisited: A meta-analytic review. *Human Communication Research*, *39*(1), 47–73. https://doi.org/10.1111/j.1468-2958.2012.01443.x

Ramaswamy, A., Tyagi, A., Hugo, H., Jiang, J., Jayaraman, P., Jangda, M., Te, A. E., Kaplan, S. A., Lampert, J., Freeman, R., Gavin, N., Tewari, A. K., Sakhuja, A., Naved, B., Charney, A. W., Omar, M., Gorin, M. A., Klang, E., & Nadkarni, G. N. (2026). ChatGPT Health performance in a structured test of triage recommendations. *Nature Medicine*, *32*, 1671–1675. https://doi.org/10.1038/s41591-026-04297-7

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Rollnick, S., & Miller, W. R. (1995). What is motivational interviewing? *Behavioural and Cognitive Psychotherapy*, *23*(4), 325–334. https://doi.org/10.1017/S135246580001643X

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), Article 65. https://doi.org/10.1057/s41599-019-0279-9

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34), Article eabo6254. https://doi.org/10.1126/sciadv.abo6254

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, *9*, 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Santos, L. A., Garton, C., & Zaki, J. (2026). Repairing cognitive distortions in political divides. *Trends in Cognitive Sciences*, *30*(1), 13–25. https://doi.org/10.1016/j.tics.2025.06.007

Schulz, C. (2023). A new algorithmic imaginary. *Media, Culture & Society*, *45*(3), 646–655. https://doi.org/10.1177/01634437221136014

Shah, D. V., McLeod, D. M., Rojas, H., Cho, J., Wagner, M. W., & Friedland, L. A. (2017). Revising the communication mediation model for a new political communication ecology. *Human Communication Research*, *43*(4), 491–504. https://doi.org/10.1111/hcre.12115

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D. (2026). Automating epistemology: How AI reconfigures truth, authority, and verification. *AI & SOCIETY*, *41*(2), 1553–1559. https://doi.org/10.1007/s00146-025-02560-y

Spagnolli, A., Chittaro, L., & Gamberini, L. (2016). Interactive persuasive systems: A perspective on theory and evaluation. *International Journal of Human–Computer Interaction*, *32*(3), 177–189. https://doi.org/10.1080/10447318.2016.1142798

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). The MIT Press.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Swire-Thompson, B., Miklaucic, N., Wihbey, J. P., Lazer, D., & DeGutis, J. (2022). The backfire effect after correcting misinformation is strongly associated with reliability. *Journal of Experimental Psychology: General*, *151*(7), 1655–1665. https://doi.org/10.1037/xge0001131

Timm, J., Talele, C., & Haimes, J. (2025). *Tailored truths: Optimizing LLM persuasion with personalization and fabricated statistics*. arXiv. https://doi.org/10.48550/arXiv.2501.17273

Tracy, S. J. (2010). Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. *Qualitative Inquiry*, *16*(10), 837–851. https://doi.org/10.1177/1077800410383121

Travis, K. (2025, April 29). AI-Reddit study leader gets warning as ethics committee moves to ‘stricter review process.’ *Retraction Watch*. https://retractionwatch.com/2025/04/29/ethics-committee-ai-llm-reddit-changemyview-university-zurich

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Ward, S. J. A. (2011). *Ethics and the media: An introduction*. Cambridge University Press.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes’ steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

About the Authors

Claire Wardle (PhD) is an associate professor of communication at Cornell University and a leading scholar of misinformation. She co-founded the non-profit First Draft and the Information Futures Lab at Brown University. Her work has shaped global understanding of information disorder within research, policy, and practitioner communities.

Pete Brown (PhD) is an independent researcher interested in the use of large language models for persuasion. He was previously research director at the Tow Center for Digital Journalism, Columbia University. He holds a PhD from the Cardiff School of Journalism, Media and Cultural Studies.

David Scales (MPhil, MD, PhD) is a practicing physician and sociologist at Weill Cornell Medicine engaged in interdisciplinary research on how structural factors shape health information environments. He developed community-oriented motivational interviewing, a community-based intervention against health misperceptions, and publishes across academic, clinical, and public venues.