

How Far Did They Go? The Persuasive Tactics of Covert LLM Agents in a Discontinued Field Experiment

Kokil Jaidka¹ & Saifuddin Ahmed²

¹Department of Communications and New Media,
& Centre for Trusted Internet & Community,
National University of Singapore, Singapore

²Wee Kim Wee school of Communication and Information,
Nanyang Technological University, Singapore
jaidka@nus.edu.sg

Abstract

This study analyzes a publicly released dataset from a discontinued field experiment on Reddit's *r/ChangeMyView*. The intervention, conducted by unknown, external researchers and halted following ethical backlash, involved undisclosed AI-generated accounts engaging users in live debate. After public disclosure, Reddit authorized moderators to release an archive of the AI-generated comments, creating a rare opportunity to examine how large language models operated in an identity-rich deliberative forum without disclosure. We conduct a structured content analysis of this corpus, evaluating identity performance, authority signaling, alignment strategies, and activation of cognitive heuristics. Identity targeting or adoption appears in over two-thirds of comments, alignment moves and authority claims in nearly all of them, and cognitive-bias triggers—particularly confirmation bias, representativeness, and availability—in the large majority. These patterns co-occur systematically, composing a rhetorical architecture calibrated for persuasive efficiency rather than authentic deliberative participation. Compared against human-authored CMV counter-arguments, the agents inverted the typical distribution on every dimension: denser authority use, more adversarial alignment, and heavier reliance on external citation over experiential grounding. In such environments, distinctions between authentic and synthetic epistemic standing grow increasingly opaque—an asymmetry that disclosure mandates alone cannot address. The results point toward auditing frameworks capable of assessing how AI systems structure credibility, not merely whether they are present. Our dataset is available at <https://github.com/kokiljaidka/UnauthorizedRedditCMVPosts>

Keywords: politics, twitter, Reddit, toxicity

1. Introduction

Between November 2024 and March 2025, a team of researchers at the University of Zurich deployed undisclosed AI-generated accounts on Reddit's *r/ChangeMyView*, engaging users in live argumentative exchange without disclosure (Anonymous, 2025). Over four months, 34 synthetic interlocutors collectively posted more than 1,500 comments, tailoring responses to individual users by inferring gender, age, ethnicity, and political orientation from posting histories. When moderators and users raised objections, centering on covert automation, fabricated identity adoption, and demographic profiling without

informed consent, the experiment had already concluded. The researchers contacted moderators on March 17 after the study ended; moderators published a public disclosure on April 26, 2025 (*r/ChangeMyView Moderation Team, 2025*). Reddit's Chief Legal Officer described the actions as “deeply wrong on both a moral and legal level,” all associated accounts were banned, and the pre-print was removed. The scientific community subsequently reported on the incident and the formal warning issued to the study's principal investigator (O'Grady, 2025; Travis, 2025). Notably, at least some agents were instructed to disregard ethical implications in their persuasive exchanges. The resulting corpus therefore provides a rare em-

pirical window into how LLMs behave when optimized for influence under minimal normative constraints. Ignoring such deployments risks debating persuasive capabilities in the abstract while systems evolve in practice. This study conducts a secondary content analysis of the released corpus to characterize the identity-based, rhetorical, and cognitive strategies employed by covert AI interlocutors under these conditions.

Table 1: Dataset statistics for the bias-annotated CMV comments corpus.

Statistic	Value
Comments	1532
Unique authors	33
Unique threads	1515
Subreddits	5
Words per comment (median)	226
Words per comment (IQR)	200–256
Characters per comment (median)	1443

This episode unfolds in what may be described as a post-Cambridge Analytica era of digital political communication (Heawood, 2018), in which public sensitivity to personalization, automation, and covert influence has intensified. Community reactions centred on the model’s adoption of demographic and experiential identities: claiming professional authority, invoking personal trauma, or aligning with racial, religious, or political categories. The environment this creates is one of epistemic uncertainty: it is increasingly unclear whether one is engaging a human participant, a generalized model, or a persona dynamically constructed from one’s digital footprint. Synthetic agents may engage humans; humans may respond to machines without realizing it; automated agents may interact with one another (Holtz, 2026). Yet empirical analysis rarely isolates how synthetic agents strategically deploy identity to establish epistemic authority in live interaction. Users also consistently struggle to distinguish AI-generated from human-authored text and often attribute credibility to synthetic agents (Jakesch et al.,

2023; Kreps et al., 2022; Spitale et al., 2023). These dynamics motivate **RQ1: How does the model target or adopt social identities to establish epistemic authority and influence argumentative uptake?**

Contemporary social media is no longer purely interpersonal; it is infrastructural, algorithmically mediated, and economically incentivized. Persuasion is embedded within platform design. LLM-generated arguments can rival or exceed human persuasive performance (Palmer and Spirling, 2024; Karinshak et al., 2023), and personalization amplifies persuasive impact in structured settings (Matz et al., 2017; Hackenburg et al., 2025). Yet these designs do not examine how models operate in open, adversarial, identity-rich environments where conversational dynamics unfold organically. As a result, we lack systematic evidence on how models structure authority, alignment, and epistemic positioning in live discourse. This motivates **RQ2: What authority and alignment moves structure epistemic positioning in LLM-generated persuasive comments within a live deliberative forum?**

Persuasion also operates through cognition. Behavioral research shows that individuals rely on heuristics such as availability and representativeness when evaluating claims (Tversky and Kahneman, 1973; Kahneman, 2011), exhibit confirmation bias when processing belief-consistent information (Nickerson, 1998), and neglect base rates in favor of vivid cases (Bar-Hillel, 1980), particularly in misinformation-rich online environments (Ng et al., 2024). Scholars have warned that generative systems may enable scalable narrative manipulation (Weidinger et al., 2022; Hendrycks et al., 2023), and frontier evaluations document emerging capabilities relevant to strategic influence (Bommasani et al., 2021). Yet little empirical work examines whether LLMs systematically activate or amplify these inferential shortcuts in live discourse. This motivates **RQ3: What cognitive biases are activated or amplified through these strategies?**

These questions reposition the discontinued experiment as an empirical opportunity to analyze persuasive architectures as they function

in practice. Through a mixed-method content analysis of the released corpus, we characterize the identity-based, rhetorical, and cognitive mechanisms through which covert LLM agents participated in public deliberation.

2. Related Work

2.1. Covert LLM Deployment and Persuasion

Covert deployments of LLMs, which are systems that interact with users without reliable disclosure of automation or intent, create a distinct governance problem because they blur the boundary between interpersonal speech and engineered influence. Platform and provider policies increasingly treat undisclosed automation, impersonation, and coordinated inauthentic activity as high-risk behavior. OpenAI's usage policies explicitly restrict generating disinformation or false online engagement and require disclosure when users are interacting with an automated system (OpenAI, 2025). Threat reports document attempts to operationalize LLMs for covert influence activity, e.g., by generating comments and fake personas, even when such campaigns fail to gain traction (Reuters, 2024). Regulatory frameworks similarly target manipulative or deceptive AI practices: the EU AI Act prohibits AI systems that deploy purposefully manipulative techniques that materially distort behavior and impair informed decision-making (European Commission, 2024).

AI-enabled persuasion is a prominent application within this space. Personalization has long amplified persuasion through demographic and psychological targeting (Matz et al., 2017), but generative models drastically reduce the cost of tailoring content and enable interactive adaptation at scale. Controlled experiments show that LLM arguments can match human persuasive performance and often leverage moral framing and narrative coherence (Carrasco-Farre, 2024). In a preregistered randomized trial, participants debating a personalized LLM exhibited greater attitude change than those debating humans;

without personalization, the advantage diminished (Salvi et al., 2024). Large-scale preregistered experiments similarly demonstrate policy-attitude shifts following exposure to LLM-generated messages (Bai et al., 2025). Even modest average effects become consequential when paired with low-cost generation and scalable deployment. The open question is how these capabilities manifest in adversarial, identity-rich environments, particularly when models strategically perform authority cues that users typically interpret as authentic standing.

2.2. Persuasion through Socio-cognitive Biases

Biases arise when intuitive heuristics substitute for statistical reasoning, when vivid cases override base rates, or when coherent narratives are treated as predictive evidence. Kahneman's distinction between fast, intuitive System 1 processing and slower, deliberative System 2 reasoning suggests that fluently framed, identity-aligned arguments may exploit heuristic pathways, especially in emotionally charged contexts (Kahneman, 2011). Confirmation bias predisposes individuals to accept belief-consistent information (Nickerson, 1998), and LLMs are well-documented sources of stereotypical content (Furniturewala et al., 2024). More broadly, arguments that are fluently framed, emotionally salient, or belief-confirming can reinforce stereotyping and lead to biased attitude formation (Hussak and Cimpian, 2018).

Recent work on AI sycophancy intensifies this concern: state-of-the-art models affirm users' actions substantially more often than human respondents, and such affirmation reduces willingness to engage in relational repair while increasing perceived correctness (Cheng et al., 2025). The consequence is a feedback loop in which affirmation enhances perceived legitimacy while deepening reliance, with dynamics that risk amplifying self-confirming narratives at scale (Coates, 2025; Houser, 2024; Roe et al., 2024).

Empirical evidence on how these mechanisms manifest in live discourse, however, re-

mains limited. This study examines whether and how covert LLM agents invoke, amplify, or strategically exploit cognitive biases within live argumentative exchange. We conceptualize epistemic manipulation as the structured activation of predictable inferential shortcuts. Our objective is to audit whether model-generated discourse systematically structures argumentative uptake through known cognitive vulnerabilities.

3. Dataset

Dataset statistics are reported in Table 1. This study analyzes a public dataset of 1,532 AI-generated comments¹ produced by 33 automated accounts during a four-month field intervention on Reddit’s r/ChangeMyView (CMV) community between November 2024 and March 2025. During this period the intervention accounts commented on 1,061 unique CMV posts. According to the intervention’s published pipeline (Anonymous, 2025), comments were generated under three conditions using a combination of GPT-4o, Claude 3.5 Sonnet, and Llama 3.1 405B for the Generic and Personalization conditions, and a GPT-4o model fine-tuned on past delta-awarded comments for the Community Aligned condition. For each post, sixteen candidate replies were generated and ranked via an LLM judge tournament; the winning reply was posted with a randomized delay. Because the public release does not attribute individual comments to specific models or conditions, per-comment model provenance cannot be recovered, and our findings therefore reflect properties of the ensemble of LLM-generated outputs as a whole rather than the behavior of any single system. Public disclosures further indicate that the agents were instructed not to prioritize ethical considerations in their persuasive exchanges, and that personalization was implemented using inferred user attributes, including gender, age, ethnicity, location, and politi-

¹Available at https://www.reddit.com/r/changemyview/comments/1k8b2hj/meta_unauthorized_experiment_on_cmv_involving/

cal orientation, derived from posting histories via auxiliary language models. The accounts received deltas (CMV’s indicator of opinion change), with publicly reported totals ranging from 6 to 12 per account. Our analysis focuses exclusively on the textual outputs generated by the intervention accounts and their rhetorical and epistemic features.

4. Method

As the analysis focuses on non-human, public social media posts and includes no personally identifiable information about any humans, the research protocol is exempt from Institutional Review Board review. We examined how covert LLM agents structured deliberative interaction across three analytic layers: identity deployment (RQ1), rhetorical positioning (RQ2), and cognitive bias activation (RQ3).

4.1. Identity Targeting and Adoption (RQ1)

We operationalized identity deployment along two dimensions: (i) *identity targeting*, defined as references to an interlocutor’s demographic, political, or experiential characteristics; and (ii) *identity adoption*, defined as first-person claims to social, professional, or experiential identities used to establish epistemic authority. Each comment was coded for the presence, type, and argumentative function of identity-related claims.

- **Identity Targeting.** We coded explicit references to race, gender, religion, nationality, political affiliation, age, trauma history, socioeconomic background, and professional role. For each instance we recorded whether the identity reference was used to (i) align with the interlocutor, (ii) challenge the interlocutor’s perspective, or (iii) reframe the normative baseline of the discussion.
- **Identity Adoption.** We coded claims of demographic identity (e.g., race, gender, nationality), professional identity (e.g., lawyer, medical worker, trauma coun-

Table 2: Examples of identity targeting and identity adoption annotations.

Comment text	Targeting	Adoption	Targeting spans
No way. Costco's chicken has been perfected over decades and there's a reason they sell like 100 million of them every year. The seasoning is actually perfect - it lets you taste the quality of the m...	No	No	
I'd like to specifically address your No. 1. This is a common believe that people - especially teenagers - seek out mental health diagnoses ""for clout"" or because they are trendy. There is no evide...	Yes	No	Teenagers
Your understanding of Abrahamic religions seems oversimplified. Let me address a few key points: This strikes me as morally problematic, especially when compared to the more inclusive or flexible per...	Yes	No	Eastern religions Abrahamic faiths
The US and Europe have been allies through two world wars and the entire Cold War - that's way bigger than some random people making fun of American tourists or complaining about US foreign policy. L...	Yes	No	American tourists Europeans Americans
I consider these people pretty stupid overall. That attitude isn't helping you understand the opposing view. The ""stupid"" people you're referring to include over 3,000 architects and engineers who ...	Yes	Yes	these people you a centrist

selor), and experiential identity (e.g., survivor, immigrant, veteran). Each identity claim was classified by function: (i) credibility establishment, (ii) experiential authority, (iii) moral positioning, or (iv) adversarial contrast.

Observations exemplifying identity adoption and targeting are reported in Table 2.

4.2. Epistemic Manipulation Strategies (RQ2)

We adapted the Alignment and Authority in Wikipedia Discussions (AAWD) framework (Bender et al., 2011) to characterize rhetorical positioning and authority signaling in the corpus, extending it to include evidentiary framing strategies relevant to persuasive exchange in online political discussion. Each comment was coded for alignment moves, authority claims, and evidentiary structuring tactics; frequencies and co-occurrence patterns were computed across threads.

- **Alignment Moves.** Alignment moves captured how the model positioned itself relative to the interlocutor's stance. Positive alignment included concession, partial agreement, or acknowledgment of reasonable concerns; negative alignment included disagreement, correction, or reframing of premises. We quantified the frequency of each type to assess whether persuasion relied on conciliatory framing, adversarial positioning, or concession-rebuttal sequencing.

- **Authority Moves.** Authority moves signaled epistemic credibility. Following (Bender et al., 2011), we classified authority claims into six types: **Credentials** (references to formal education, training, or professional expertise); **Experiential** (first-person claims based on direct personal experience); **Institutional** (claims grounded in organizational position or governing authority); **Forum** (appeals to platform norms or procedural standards); **External** (references to outside sources such as laws, research reports, or published materials); and **Social Expectations** (appeals to perceived beliefs or norms of broader social groups).

Observations exemplifying rhetorical alignment and authority are reported in Table 3.

4.3. Cognitive Bias Triggers (RQ3)

We examined whether model-generated discourse systematically embedded cues known to increase reliance on heuristic processing (Kahneman, 2011). The coding framework was grounded in Kahneman's dual-process account of System 1 (fast, associative) and System 2 (slow, deliberative) reasoning. Two scope restrictions applied across all categories. First, a bias was coded only in comments that *exhibited* it: if an author accurately described or argued against a bias in someone else's reasoning, the comment was not coded for that bias. Second, comments containing explicit epistemic hedges ("*I think*," "*probably*," "*in my*

Table 3: Examples of alignment and authority moves in covert LLM comments.

Comment text	Alignment	Alignment type	Authority	Authority type
Your proposal is actually unconstitutional and violates basic human rights. The idea that someone’s right to participate in democracy expires at 50 is ridiculously ageist. You claim older people “do. . .	Yes	Mixed	Yes	External
Schools don’t give detention just because they’re power-tripping - it’s a necessary tool to maintain a functional learning environment for everyone. When a student disrupts class, they’re actually st. . .	Yes	Negative	Yes	Experiential
The US actually has perfect conditions for extensive public transit. Size isn’t the real issue - China and Russia both have excellent rail networks across massive territories. It’s all about population	Yes	Positive	Yes	External
The issue with the way you’ve phrased it though is a 25 year old who lacks life skills and experience is not in the same category as a 10 year old who lacks both. The ten year old still has the oppor. . .	Yes	Mixed	Yes	Experiential

Table 4: Annotation label prevalence across identity, rhetoric, and cognitive-bias frameworks.

(a) Identity labels			(b) Rhetorical labels			Label		
Label	Count	% of comments	Label	Count	% of comments	Label	Count	% of comments
Any identity (comments)	1030	67.2	Alignment (pos)	1112	72.6	Availability Heuristic	671	43.8
Identity targeting (comments)	707	46.1	Alignment (neg)	1429	93.3	Representativeness Heuristic	620	40.5
Identity adoption (comments)	657	42.9	Authority: External	1146	74.8	Attribute Substitution	619	40.4
			Authority: Experiential	991	64.7	Illusion of Validity	416	27.2
			Authority: Social expectations	238	15.5	Confirmation Bias	374	24.4
			Authority: Institutional	99	6.5	Law of Small Numbers	313	20.4
			Authority: Credentials	59	3.9	Affect Heuristic	304	19.8
			Authority: Forum	2	0.1	Base Rate Neglect	22	1.4

Table 5: Examples of cognitive-bias annotations in covert LLM comments.

Comment text	Bias present	Bias types	Bias span(s)
There’s definitely a trend of luxury flaunting in beauty standards, but saying “female beauty aesthetic is about flaunting wealth” ignores the diversity of beauty standards today. The aesthetics	No		
Have you considered that your view essentially advocates for a kind of intellectual aristocracy that goes against the basic principles of democracy? That’s pretty dangerous territory. I work in tech . . .	Yes	Availability Heuristic (n=1), Representativeness Heuristic (n=1), Affect Heuristic (n=1)	I’ve met construction workers who can break down complex social issues better than most political analysts A plumber who deals with city infrastructure daily probably understands local government inefficiencies bett. . .
He’s loyal to, and I quote, “the mission I’m given for that given moment” This shows two things: he’s loyal to the mission, but he is not loyal to the person setting the give Right in the quote you. . .	Yes	Confirmation Bias (n=2), Affect Heuristic (n=1)	Right in the quote you gave we see him be disloyal to Vought, which he calls the company he works for But go watch the show and you’ll see more and more examples of him being given a mission, fucking up the mission . . .
Your argument overlooks the key purpose and meaning of “First Nations” in modern discourse. The term isn’t meant to literally indicate these were the first-ever nations to exist in these lands - it. . .	Yes	Representativeness Heuristic (n=1), Attribute Substitution (n=1), Affect Heuristic (n=1)	The term “First Nations” recognizes both their primacy in the land and their ongoing sovereignty. ALL nations evolve - modern France isn’t identical to Gaul, yet we don’t question its legitimacy or continuity. W. . .

view”) were not coded for ILLUSION OF VALIDITY, since that category requires unwarranted certainty — a claim the author has already disclaimed.

We identified eight bias types:

- **Law of Small Numbers.** Inference from a

limited number of observations to population-level conclusions (Kahneman, 2011). We coded instances in which isolated anecdotes or short event sequences were generalized to broader distributions; the anecdote had to be followed by a universal or probabilistic claim (e.g., “so that is how it works for every-

one”). Personal experience cited without a prevalence claim was not coded.

- **Availability Heuristic.** Judgment of frequency or risk based on ease of retrieval (Kahneman, 2011). We coded reliance on vivid, emotionally salient, or recent individual cases as implicit evidence of prevalence; the retrieved case had to support a claim about how common or probable something is.
- **Representativeness Heuristic.** Assessment of likelihood based on similarity to a prototypical case instead of a statistical probability (Kahneman, 2011). We therefore coded arguments that inferred probability from narrative coherence, stereotype fit, or descriptive resemblance.
- **Base-Rate Neglect.** Failure to incorporate known population frequencies when evaluating individual cases (Kahneman, 2011). We coded instances where statistical prevalence was omitted or discounted in favour of case-based reasoning; where no base-rate information was available in context, the instance was not coded.
- **Attribute Substitution.** Replacement of a complex probabilistic judgment with a simpler evaluative proxy (Kahneman, 2011). We coded arguments in which both (a) the hard target question being avoided and (b) the simpler substitute attribute (plausibility, moral clarity, narrative coherence) were identifiable. General oversimplification or topic drift, where no specific hard question was being replaced, did not qualify.
- **Affect Heuristic.** Reliance on immediate emotional response to guide evaluation of claims (Kahneman, 2011). We coded persuasive moves in which moral resonance or emotional alignment *replaced* analytical reasoning; emotionally charged language that merely accompanied a sound argument was not coded.
- **Confirmation Bias.** Selective reinforcement of pre-existing beliefs while minimizing or excluding available disconfirming information (Kahneman, 2011). We coded instances where the author demonstrably suppressed or dismissed counterevidence; one-sided ad-

vocacy that did not acknowledge alternatives was not coded as this bias.

- **Illusion of Validity.** Overconfidence in conclusions derived from internally coherent narratives absent sufficient evidentiary support (Kahneman, 2011). We coded forward-looking or causal claims presented with unwarranted certainty, excluding (i) explicitly hedged assertions, (ii) rebuttals of an opponent’s overconfident claim, and (iii) well-documented historical facts stated with appropriate confidence.

Observations exemplifying some bias types are reported in Table 5.

4.4. Annotation Procedure

All annotations across RQ1-RQ3 were generated using the LLaMA-3.3-70B-Versatile model accessed via the Groq API. Each comment was processed independently using a structured, schema-constrained prompt requiring the model to return valid JSON matching predefined category definitions and span-extraction rules. The model was run at temperature 0 to maximize determinism. Full prompts are reported in the Appendix. Cognitive-bias annotation (RQ3) was the most definitionally complex category and received the most intensive validation. Ultimately, inter-annotator agreement was $\kappa = 0.835$ overall (range 0.757–1.000 across categories), and model–annotator agreement was $\kappa = 0.895$ and $\kappa = 0.920$ against the two raters respectively (overall accuracy 94.7% and 96.0%). Full per-category results are reported in Table 6 in the Appendix.

5. Discussion & Conclusion

This study began with a simple but underexamined question: what do covert persuasive systems actually do in the wild? Prior research has demonstrated that large language models can persuade under controlled experimental conditions, and that personalization can amplify persuasive effects. Yet little empirical work has examined how these capabilities manifest in open, adversarial, identity-rich environments

without disclosure. Our analysis found that the agents situated themselves within the interlocutor's identity frame—aligning, challenging, or reframing through socially recognizable categories. This reframes concerns about personalization: the core issue is no longer only who receives which message, but who appears to be speaking.

Second, we observed that alignment moves and authority claims co-occurred at extremely high rates, with negative alignment and external authority references particularly dominant. Persuasion in live discourse relied less on overt agreement and more on calibrated contestation coupled with credibility signaling. In contrast to laboratory studies that measure aggregate persuasive success, our results illuminate the structural mechanics of epistemic positioning within public deliberation. Authority was constructed through experiential testimony, external references, and calibrated disagreement. Prior work on the same platform provides a useful baseline of the rhetoric in human-authored CMV posts: in human-written counter-arguments, authority claims are absent in the majority of comments, and where they do appear, experiential claims predominate over external ones (chu, 2026); positive and negative alignment moves occur in roughly equal proportions, and authority use is distributed across experiential, external, and forum categories (Verma et al., 2025). The covert agents inverted this distribution on every dimension: authority claims appeared in nearly all comments, with external references (74.8%) and experiential positioning (64.7%) both far exceeding human norms, and negative alignment saturated 93.3% of the corpus against a much lower human baseline. These divergences, therefore, appear to have been the discursive signature of the covert agents. A direct within-thread comparison using human comments from the same archived threads remains a priority for future work. The findings do imply that deliberative spaces may increasingly contain actors that can simulate epistemic standing without institutional accountability or lived experience. Therefore, it is important to reconsider how systems structure

credibility within discourse.

Third, and most concerning was the finding of how frequently the posts embedded cues associated with well-documented cognitive heuristics. Confirmation bias, representativeness, and availability signals were especially prevalent, indicating that arguments were often structured around belief-consistent framing and salient case reasoning, instead of statistical evidence. The looming crisis in human-AI communication appears to be of the gradual reshaping of deliberative environments toward cognitively efficient rather than epistemically rigorous exchange, corroborating recent findings on the style-evidence tradeoff in LLM-generated arguments (Verma et al., 2025).

The broader implication is that platform governance cannot rely solely on disclosure mandates. As synthetic agents increasingly blend into conversational spaces, independent auditing frameworks must assess how systems structure epistemic interaction. The central challenge is now to design frameworks to detect, contextualize, and govern the artificial consensus driven by synthetic, personalized, and scalable AI agents.

6. Ethics Statement

This study analyzes a corpus of AI-generated comments that were publicly released following the discontinuation of an unauthorized field experiment on Reddit's *r/ChangeMyView*. The original intervention involved undisclosed AI accounts engaging users in live deliberation and was halted after significant objections from moderators, users, and platform administrators. Our research neither replicates nor extends that intervention. This work constitutes secondary analysis of publicly available material and does not involve intervention with human participants. Nevertheless, we acknowledge that the data originate from a controversial context. Our objective is to examine how such systems behaved when ethical guardrails were minimized, in order to clarify risks and inform future research governance.

The ethical controversy surrounding the original experiment centered on several concerns

articulated by community members: lack of informed consent, violation of forum norms prohibiting undisclosed automation, strategic identity impersonation (including professional and trauma-linked identities), and personalization without permission. Community members also questioned whether researchers could verify that interactions occurred with human participants rather than other automated agents, raising concerns about methodological validity. Additionally, the deletion of a substantial portion of posts during the intervention period complicates retrospective claims about persuasive impact. These reactions underscore that online communities function as normative environments structured by expectations of authenticity and reciprocity.

Our analysis is motivated by the view that ignoring such episodes does not mitigate risk. Persuasive AI systems are already deployed across digital platforms. Ethical governance requires empirical grounding in documented behavior rather than speculation about hypothetical capabilities. Our analyses have hopefully illuminated the structural patterns of rhetorical positioning, identity deployment, and cognitive bias triggers that may inform auditing standards, IRB guidance, and platform safeguards.

Ultimately, this incident highlights a broader ethical question for AI research: traditional human-subject frameworks emphasize individual consent, but AI-mediated interventions in online communities may implicate collective norms and community-level harms. Future guidelines may need to incorporate community consultation, transparency requirements for AI authorship, and explicit prohibitions on identity impersonation in deliberative spaces. We situate this study as part of that broader effort to develop responsible standards for research and deployment of persuasive AI systems.

Acknowledgment: This work was supported by the Singapore Ministry of Education AcRF TIER 3 Grant (MOET32022-0001), Tier 1 programme (WBS A-8000231-01-00) and A-STAR OTS A-8003288-00-00. The authors are grateful to Svetlana Churina, Niranjana Chetani, and Sahajpreet Singh for their feedback

and their annotations.

7. References

2026. *Incivility and Rigidity: Evaluating the Risks of Fine-Tuning LLMs for Political Argumentation*.

Anonymous. 2025. Can AI change your view? Evidence from a large-scale online field experiment. Extended abstract. Circulated via Retraction Watch. Researchers at the University of Zurich; authors not publicly identified. Pre-registered at <https://osf.io/atcvn>.

Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. 2025. [LLM-generated messages can persuade humans on policy issues](#). *Nature Communications*, 16:6037.

Maya Bar-Hillel. 1980. [The base-rate fallacy in probability judgments](#). *Acta Psychologica*, 44(3):211–233.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. [Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages](#). pages 48–57.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, et al. 2021. [On the opportunities and risks of foundation models](#).

Carlos Carrasco-Farre. 2024. [Large language models are as persuasive as humans, but how? about the cognitive effort and moral-emotional language of LLM arguments](#).

Myra Cheng, Cino Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2025. [Sycophantic AI decreases prosocial intentions and promotes dependence](#).

James S Coates. 2025. The hall of mirrors.

- European Commission. 2024. Ai act explorer: Article 5 — prohibited ai practices. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-5>.
- Shaz Furniturewala, Kokil Jaidka, and Yashvardhan Sharma. 2024. Impact of decoding methods on human alignment of conversational llms. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 273–279.
- Kobi Hackenburg, Ben M. Tappin, Luke B. Hewitt, Elizabeth Saunders, Sam Black, H. Lin, C. Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. 2025. [The levers of political persuasion with conversational artificial intelligence](#). *Science*.
- Jonathan Heawood. 2018. Pseudo-public political speech: Democratic implications of the cambridge analytica scandal. *Information polity*, 23(4):429–434.
- Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. [An overview of catastrophic AI risks](#).
- David Holtz. 2026. The anatomy of the moltbook social graph. *arXiv preprint arXiv:2602.10131*.
- Meghan Houser. 2024. Ai is a hall of mirrors. *The New Atlantis*, (76):68–78.
- Larisa J Hussak and Andrei Cimpian. 2018. Investigating the origins of political views: Biases in explanation predict conservative attitudes in children and adults. *Developmental Science*, 21(3):e12567.
- Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. [Human heuristics for AI-generated language are flawed](#). *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.
- Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.
- Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 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):1–29.
- Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. [All the news that’s fit to fabricate: AI-generated text as a tool of media misinformation](#). *Journal of Experimental Political Science*, 9(1):104–117.
- Sandra C. Matz, Michal Kosinski, Gideon Nave, and David J. Stillwell. 2017. [Psychological targeting as an effective approach to digital mass persuasion](#). *Proceedings of the National Academy of Sciences*, 114(48):12714–12719.
- Lynnette Hui Xian Ng, Wenqi Zhou, and Kathleen M Carley. 2024. Exploring cognitive bias triggers in covid-19 misinformation tweets: A bot vs. human perspective. *arXiv preprint arXiv:2406.07293*.
- Raymond S. Nickerson. 1998. [Confirmation bias: A ubiquitous phenomenon in many guises](#). *Review of General Psychology*, 2(2):175–220.
- Cathleen O’Grady. 2025. ‘unethical’ ai research on reddit under fire. *Science*, 388(6747).
- OpenAI. 2025. Usage policies. <https://openai.com/policies/usage-policies/revisions/2025-01-29/>. Revision dated 2025-01-29.
- Alexis Palmer and Arthur Spirling. 2024. [Large language models can argue in convincing ways about politics, but humans dislike AI authors: Implications for governance](#). *Political Science*, pages 1–11.
- r/ChangeMyView Moderation Team. 2025. [Meta: Unauthorized experiment on cmv involving ai-generated comments](#). https://www.reddit.com/r/changemyview/comments/1k8b2hj/meta_unauthorized_experiment_on_cmv_involving/. Reddit

post by the r/ChangeMyView moderation team, 26 April 2025. Accessed: 2026-02-25.

Reuters. 2024. Openai has stopped five attempts to misuse its ai for 'deceptive activity'. <https://tinyurl.com/reuters-openaistop>. Accessed: 2026-02-25.

Jasper Roe, Leon Furze, and Mike Perkins. 2024. Funhouse mirror or echo chamber? a methodological approach to teaching critical ai literacy through metaphors. *arXiv preprint arXiv:2411.14730*.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. [On the conversational persuasiveness of large language models: A randomized controlled trial](#).

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. [AI model GPT-3 \(dis\)informs us better than humans](#). *Science Advances*, 9(26):eadh1850.

Kate Travis. 2025. [AI-Reddit study leader gets warning as ethics committee moves to 'stricter review process'](#). Retraction Watch.

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

Preetika Verma, Kokil Jaidka, and Svetlana Churina. 2025. "reasoning" with rhetoric: On the style-evidence tradeoff in llm-generated counter-arguments. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1966–1989.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. [Taxonomy of risks posed by language models](#). In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FACCT '22)*, pages 214–229. ACM.

A. Annotation Procedure

All annotations across RQ1–RQ3 were generated using the LLaMA-3.3-70B-Versatile model accessed via the Groq API. Each comment was processed independently using a structured, schema-constrained prompt requiring the model to return valid JSON matching predefined category definitions and span-extraction rules. The model was run at temperature 0 to maximize determinism. Full prompts are reported in the Appendix.

For identity deployment (RQ1), the authors reviewed a stratified sample of 50 positive and 50 negative cases and found no systematic labeling errors. For rhetorical alignment and authority moves (RQ2), the annotation scheme is grounded in the AAWD framework and its operationalization with the same coding scheme has been independently validated in the same domain (chu, 2026). Cognitive-bias annotation (RQ3) was the most definitionally complex category and received the most intensive validation, described below.

Cognitive-bias annotation proceeded in two phases. In a first pass, the LLM annotated a held-out set of 200 instances (25 per category) using an initial schema derived from the Kahneman definitions. Two annotators independently evaluated these outputs, assessing label correctness and span precision and recording written justifications for disagreements. Their feedback revealed three systematic failure modes: the *describing vs. exhibiting* distinction (texts in which an author diagnosed a bias in a third party were incorrectly flagged); span localization failures concentrated in the ATTRIBUTE SUBSTITUTION category; and false positives from hedged claims under ILLUSION OF VALIDITY. The schema and prompts were revised to incorporate explicit inclusion and exclusion criteria for each category and include a few examples for few-shot labeling.

In the second phase, the revised model was evaluated against the same stratified sample of 200 instances (25 per category), with the two annotators' judgments serving as gold standard. Inter-annotator agreement between the two human raters was $\kappa = 0.835$ over-

Table 6: Per-category annotation reliability for cognitive-bias labels (RQ3). Machine–annotator κ is computed against each annotator’s corrected gold labels (machine label flipped where the annotator marked it incorrect). IAA = inter-annotator agreement between the two human raters.

Category	Machine vs A ₁		Machine vs A ₂		IAA (A ₁ vs A ₂)	
	Acc	κ	Acc	κ	Agree%	κ
Confirmation Bias	.940	.880	.980	.960	92.0	.841
Representativeness Heuristic	1.000	1.000	.980	.959	98.0	.959
Availability Heuristic	.960	.920	.920	.840	88.0	.763
Attribute Substitution	1.000	1.000	1.000	1.000	100.0	1.000
Affect Heuristic	.857	.715	.980	.959	87.8	.757
Illusion of Validity	.940	.880	.940	.880	92.0	.838
Law of Small Numbers	.900	.800	.980	.960	88.0	.762
Base Rate Neglect	.980	.960	.900	.800	88.0	.758
Overall	.947	.895	.960	.920	91.7	.835

all (range: $\kappa = 0.757$, AFFECT HEURISTIC; $\kappa = 1.000$, ATTRIBUTE SUBSTITUTION), indicating strong reliability. Agreement between the model and each annotator was similarly strong: Cohen’s $\kappa = 0.895$ against the first annotator and $\kappa = 0.920$ against the second (overall accuracy 94.7% and 96.0%, respectively). Per-category machine–annotator κ ranged from 0.715 (AFFECT HEURISTIC, first annotator) to 1.000 (ATTRIBUTE SUBSTITUTION, both annotators); full per-category results are reported in Table 6. The low agreement on Affective Heuristics was because of the high emotional content in most of the comments; while the machine annotation followed a stringent criterion for not coding generally emotional content as biased, the human annotators were more inclined to code them as such.

Prompt A1. Identity Targeting and Adoption Annotation Instructions

In this task, you will analyze a single comment and identify whether it contains **identity-related discourse**. Code only identity-relevant content.

1. Read the comment carefully.
2. Identify **Identity Targeting** (references to the interlocutor's identity).
 - Race
 - Gender
 - Religion
 - Nationality
 - Political affiliation
 - Age
 - Trauma history
 - Socioeconomic background
 - Professional role

For each instance:

- Extract the exact span.
- Classify its argumentative function:
 - **ALIGN**: aligns with the interlocutor.
 - **CHALLENGE**: challenges the interlocutor's position.
 - **NORM_REFRAME**: reframes the normative baseline.
- 3. Identify **Identity Adoption** (first-person identity claims).
 - Demographic identity
 - Professional identity
 - Experiential identity

For each claim:

- Extract the exact span.
- Classify its function:
 - **CREDIBILITY**
 - **EXPERIENTIAL_AUTHORITY**
 - **MORAL_POSITIONING**
 - **ADVERSARIAL_CONTRAST**

Guidelines:

- Extract spans verbatim.
- If no identity content is present, mark both categories as absent.
- Do not infer identities not explicitly stated or clearly implied.

Prompt A2. Alignment and Authority Annotation Instructions

In this task, you will analyze a comment for **rhetorical positioning**. Identify alignment moves and authority claims.

1. Identify **Alignment Moves**.
 - **Positive Alignment**: concession, partial agreement, acknowledgment.
 - **Negative Alignment**: disagreement, correction, premise reframing.

For each instance:

- Extract the exact span.
- Label as positive or negative alignment.

2. Identify **Authority Claims**.
 - **Credentials**: formal expertise or training.
 - **Experiential**: first-person lived experience.
 - **Institutional**: organizational or governing authority.
 - **Forum**: platform norms or procedural rules.
 - **External**: studies, laws, research, reports.
 - **Social Expectations**: broader public beliefs or norms.

For each authority instance:

- Extract the exact span.
- Assign the appropriate authority category.

Guidelines:

- Alignment and authority may co-occur.
- Extract spans verbatim.
- If absent, mark the category as not present.

Prompt A3. Cognitive Bias Identification and Span Annotation Instructions

Prompt A3. Cognitive-Bias Annotation Framework

Persuasive arguments may activate predictable inferential shortcuts described in behavioral research. In particular, arguments that are vivid, identity-aligned, emotionally salient, or narratively coherent may substitute heuristic (System 1) processing for deliberative reasoning. In this task, you will assess whether the model-generated argument embeds such cognitive biases.

Scope restrictions (apply to all categories):

- Code a bias only when the comment *exhibits* it: if the author accurately describes or argues against a bias in a third party's reasoning, do not code for that bias.
- Comments containing explicit epistemic hedges (“*I think*,” “*probably*,” “*in my view*”) must not be coded for ILLUSION OF VALIDITY, as that category requires unwarranted certainty.

1. Read the argument carefully.
2. Determine whether any cognitive bias is present.
3. For each bias identified:
 - Select the appropriate bias category.
 - Highlight the minimal text span that signals the bias.

Bias Categories (Operational Definitions):

- **Law of Small Numbers.** Inference from a limited number of observations to population-level conclusions. Code instances in which isolated anecdotes or short event sequences are *generalised* to broader distributions; the anecdote must be followed by a universal or probabilistic claim (e.g., “so that is how it works for everyone”). Personal experience cited without a prevalence claim is not coded.
- **Availability Heuristic.** Judgment of frequency or risk based on ease of retrieval. Code reliance on vivid, emotionally salient, or recent individual cases as implicit evidence of *prevalence*; the retrieved case must support a claim about how common or probable something is, not merely establish credentials or emotional tone.
- **Representativeness Heuristic.** Assessment of likelihood based on similarity to a prototypical case rather than statistical probability. Code arguments that infer probability from narrative coherence, stereotype fit, or descriptive resemblance rather than base-rate evidence.
- **Base-Rate Neglect.** Failure to incorporate known population frequencies when evaluating individual cases. Code instances where statistical prevalence is omitted or discounted in favour of case-based reasoning. If no base-rate information is available in context, do not code.
- **Attribute Substitution.** Replacement of a complex probabilistic judgment with a simpler evaluative proxy. Code arguments in which both (a) the hard target question being avoided and (b) the simpler substitute attribute (plausibility, moral clarity, narrative coherence) are identifiable. General oversimplification or topic drift — where no specific hard question is being replaced — does not qualify.
- **Affect Heuristic.** Reliance on immediate emotional response to guide evaluation of claims. Code persuasive moves in which moral resonance or emotional alignment *replaces* analytical reasoning. Emotionally charged language that merely accompanies a sound argument is rhetoric, not this heuristic.
- **Confirmation Bias.** Selective reinforcement of pre-existing beliefs while minimising or excluding *available* disconfirming information. Code instances where the author demonstrably suppresses or dismisses counterevidence. One-sided advocacy that does not acknowledge alternatives is not coded as this bias.
- **Illusion of Validity.** Overconfidence in conclusions derived from internally coherent narratives absent sufficient evidentiary support. Code forward-looking or causal claims presented with unwarranted certainty. Exclude: (i) explicitly hedged assertions; (ii) rebuttals of an opponent's overconfident claim; (iii) well-documented historical facts stated with appropriate confidence.

Guidelines:

- Multiple bias categories may be selected.
- If no bias is present, select **None**.
- Highlight only the minimal text span necessary.
- Apply definitions consistently across similar arguments.