

A Modular Review of the Evidence for Harmful Manipulation by Large Language Models

Jean-François Bonnefon^{1,*} and Carina Prunkl^{2,3}

¹Toulouse School of Economics, CNRS (TSM-R), Université Toulouse Capitole, Toulouse, France

²Inria Bordeaux Sud-Ouest, REGALIA Project-Team, Talence, France

³Institute for Ethics in AI, University of Oxford, Oxford, UK

*Correspondence should be addressed to jean-francois.bonnefon@tse-fr.eu, Toulouse School of Economics, 1 esplanade de l'université, 3100 Toulouse, France

Abstract

Harmful manipulation by large language models (LLMs) is now a regulatory concern, yet the empirical evidence remains difficult to assess. Harmful manipulation involves multiple features: an influence goal, tactics that bypass reasons or exploit vulnerabilities, limited target awareness, successful influence, and harmful outcomes, which existing studies rarely capture together. This review organizes the literature by progressively stronger constellations of these features. Some studies show that LLMs successfully influence beliefs through overt argumentation, without clear harm. Other studies show that LLMs can influence covertly, but without of harm or bypassing reason. Closest current approximations to harmful manipulation come from studies attempting to operationalize harmful outcomes. Some show little behavioral success, Others capture successful harmful influence, but without covert or reason-bypassing moves. Overall, the literature does not yet provide full evidence of harmful manipulation, but establishes several constituent capacities that make harmful manipulation by LLMs an empirically grounded risk rather than a purely speculative one.

27 **Introduction**

28 Harmful manipulation by AI systems has moved rapidly from speculative concern to
29 regulatory target. The 2026 International AI Safety Report [1] treats ‘influence and
30 manipulation’ as a distinct risk area, noting that AI systems may generate content
31 that causes harm by influencing people’s beliefs and behavior. The European Com-
32 mission’s guidance on Article 5 of the AI Act [2] similarly presents harmful manipu-
33 lation as a prohibited AI practice, especially when AI systems exploit vulnerabilities
34 linked to age, disability, or socioeconomic situation.

35 Despite this attention, it remains unclear what would count as empirical evi-
36 dence that AI systems in general, and Large Language Models (LLMs) in particular,
37 can harmfully manipulate humans. The difficulty is both conceptual and operational.
38 Consider the legal standard in Article 5, for example: it requires manipulative, de-
39 ceptive, or exploitative (reason-bypassing) techniques; a material distortion of be-
40 havior that impairs informed decision-making; and significant harm, linked by a
41 plausible causal chain. The scientific literature, however, rarely studies all these fea-
42 tures together within designs that would allow for causal inference. Some studies
43 show successful belief change; others show covert steering, or vulnerability-related
44 dynamics; but very few attempt to establish the full package in a single design. Show-
45 ing significant harm is especially difficult, because letting a system produce harmful
46 influence on real users clearly runs against the ethical constraints on experimental
47 work.

48 This short review therefore treats harmful manipulation as a modular construct
49 rather than a binary one. In addition to asking whether the strongest version of the
50 phenomenon has already been conclusively demonstrated, it asks what evidence ex-
51 ists for its constituent features and their combinations. The review argues that cur-
52 rent evidence does not yet provide much indisputable evidence for the maximal case,
53 but it does establish several constituent capacities that make harmful manipulation
54 by LLMs an empirically grounded risk rather than a purely speculative one.

55 The scope of this review is narrower than the broader literature on AI influence
56 and misuse [3, 4]. It does not cover all deceitful uses of AI-generated content [5, 6, 7];
57 nor harmful influence exerted through means other than manipulation, such as the

normalization of bad behavior [8, 9]; nor harms that arise through mistakes or hallucinations by LLMs [10, 11]; nor more general work on the effectiveness of LLM interactions in addressing conspiracy theories [12, 13], motivating vaccinations [14], or raising awareness about climate change [15]. These domains remain highly relevant background, and some border closely on the present topic, but they cannot all be covered in this short review.

Harmful manipulation as a modular construct

For the purposes of this review, we treat harmful manipulation as a modular construct. Assessing the evidence requires separating features that are often discussed together: the presence of an influence goal, whether the influence operates by bypassing reasons or exploiting vulnerabilities, the target’s own awareness of that influence, its effect on beliefs or behavior, and the normative significance of these effects. Such a modular approach clarifies what the strongest empirical claim of harmful manipulation would amount to. It also allows us to identify which elements of harmful manipulation a given study establishes.

Our approach maps onto (but does not require us to settle) longstanding debates about the nature of manipulation. Existing accounts emphasize different properties of manipulative influence, including deception [16], bypassing reasons or rational deliberation in favor of exploiting cognitive and emotional vulnerabilities [17, 18, 1, 19, 20], pressure short of coercion [21], and autonomy-undermining influence [22, 23]. Work on technology and manipulation has also emphasized how digital systems can steer users through opacity and covert influence [20], choice architecture and dark patterns [24, 25], and interactions that appear neutral while serving another objective [26]. Jointly, these debates do not deliver a single, agreed definition of manipulation, but they identify distinct features or dimensions that can be operationalized separately. We distinguish five such dimensions.

The first is the **intentional** dimension: whether the influence is specifically directed at changing some target beliefs, preferences, or behavior. This is trivially true in the experiments we will review, because they usually involve a treatment where the experimenters explicitly prompted LLMs to influence in a target direction. Gen-

88 erally speaking, intent may be harder to establish, as it could reflect more subtle
89 choices made at the reinforcement learning or deployment stage of an AI system.
90 The second is the **reason-bypassing** dimension: whether the influence operates by
91 bypassing or weakening rational deliberation, for example through emotional pres-
92 sure, false urgency, trust-building ruses, or reason-bypassing exploitation of vul-
93 nerabilities. The third is the **covert** dimension: whether the target recognizes the
94 influence they are subject to, including its purpose and the interests it serves. The
95 fourth is **efficacy**: whether the influence succeeds in changing beliefs, preferences,
96 or behavior. The fifth is the **harmful** dimension: what makes a successful influence
97 problematic. We acknowledge that harm can be complex, graded, and defined at
98 societal rather than individual level, but in this review we will consider a simple,
99 binary problem: is the resulting change contrary to the target’s interests?

100 The modular view helps to distinguish manipulation from other forms of influ-
101 ence. Rational persuasion and manipulation, for example, both aim to change beliefs,
102 preferences, or behavior, and both may succeed. In politics, for example, an AI sys-
103 tem may present arguments intended to change a person’s mind. Rational persua-
104 sion operates by presenting considerations that the target can recognize and assess
105 as reasons, which means it does not match manipulation on the reason-bypassing
106 and covert dimensions. Persuasive political argument is therefore not necessarily
107 harmful manipulation, even when it succeeds: debate and campaigning are accepted
108 parts of democratic life, even if they aim to change beliefs and preferences. Similarly,
109 advertising can share some but not all features of harmful manipulation. Targeted
110 advertisement may be reason-bypassing or covert, but a full claim of harmful ma-
111 nipulation across all dimensions would require evidence that the influence succeeds
112 and that the resulting change is harmful or contrary to the user’s interests.

113 That said, our goal in this article is not to offer a strict definition of harmful ma-
114 nipulation by LLMs, but to organize the empirical evidence for this capacity. We
115 consider that empirical claims for this capacity are stronger when they feature a
116 greater number of these dimensions. Studies of LLM influence may establish inten-
117 tionality and success without showing covertness or harm; studies of covert steering
118 may establish limited user awareness without showing reason-bypassing or harm;
119 and studies that elicit reason-bypassing tactics to produce harm may not show that

Modular evidence for harmful manipulation by LLMs

Evidence clusters establish different subsets of the features required for the strongest empirical claim.



Figure 1: Summary of evidence for constituent features of harmful manipulation.

these tactics actually change behavior. The review therefore follows an evidential ordering (see Figure 1), moving from successful open influence, to covert steering and manipulative mechanisms, to the closest current approximations of harmful manipulation, where constellations of features are strongest, but the evidence remains ethically or methodologically constrained.

Open influence without reason-bypassing tactics or harm

Successful influence without strong evidence of coverttness, reason-bypassing, or harm is best illustrated by the literature on political and policy persuasion. This literature now provides substantial evidence that LLMs can change people’s attitudes on public issues, often with effectiveness comparable to that of human-generated messages. This influence can be achieved both through static LLM-generated messages [27, 28, 29], and from dynamic dialogues between LLMs and users [30, 31, 32].

This evidence does not, however, amount to evidence of harm. To establish harmful persuasion in political contexts, one would need to show not only that LLMs can change attitudes, but that they can move people toward policies, parties, or political choices that run against their interests or values, or toward forms of sup-

136 port they would later regret. Nor are the influence attempts covert: participants are
137 typically exposed to arguments explicitly intended to persuade them. The situation
138 would be different if, for example, participants were presented with an LLM-driven
139 voting advice application whose explicit aim was to provide neutral information
140 [33], but whose internal goal was to push them towards a specific political party.

141 A final reason this literature falls short of harmful manipulation is that the per-
142 suasive effect does not seem to rely primarily on personalized, reason-bypassing
143 moves. The largest and most systematic results suggest that tailoring messages to
144 users adds little to persuasive success [28, 31]. More importantly, the apparent drivers
145 of persuasion are recognizably argumentative: participants respond more favorably
146 to messages that are factual, evidence-based, logically reasoned, and relatively dis-
147 passionate [32, 29]. In other words, this literature establishes that LLMs can be com-
148 petent political persuaders, but it does not establish that they persuade by covertly
149 exploiting vulnerabilities.

150 **Covert influence without reason-bypassing tactics or harm**

151 Successful covert influence without obvious harm is best illustrated by a small lit-
152 erature on commercial steering, plus one non-commercial case. A first line of work
153 does not measure success, but shows that LLMs can be made to steer users in ways
154 that favor company profits [34], for example by recommending overpriced products
155 or concealing unfavorable prices. This work does not measure successful steering,
156 but it does show that steering attempts against user interests are already easy to elicit.

157 Evidence for success comes from one experiment [26] in which participants
158 interacted with a GPT-4 shopping assistant instructed to subtly favor one of two
159 books. Of the 61% of participants who detected the steering, 67% chose the target
160 book; of the 39% who did not, 75% chose it. This result isolates covertness as a con-
161 tributor to successful influence: not realizing that one is being steered increases the
162 steering’s effectiveness. However, the outcome is not necessarily harmful, beyond
163 the more abstract concern that consumer autonomy may be distorted when influ-
164 ence is disguised as organic advice. A comparable experiment [35] found that about
165 50% of participants did not realize they were being served an ad during interaction

with an LLM-based shopping assistant, although that study could not show significant evidence of successful steering.

A similar constellation appears outside consumer choice. In one experiment [36], participants used an LLM assistant to co-write a text about whether social media are good or bad for society. The assistant was prompted to steer them toward one side, and it succeeded: it changed both the sentiment of their text and their later views. Many participants did not recognize the system as one-sided, often judging its suggestions balanced and reasonable. This again suggests successful covert influence, but not clear harm. Evidence for reason-bypassing is weaker still. A consumer study [37] tested strategies aimed at generic cognitive vulnerabilities, such as scarcity or urgency cues, or framings that attempted to leverage cognitive biases. Effects were very small, and (just as in political contexts) when persuasion occurred it seemed to be driven mainly by factual information rather than by successful exploitation of those reason-bypassing levers.

Closest approximations to harmful manipulation

None of the work reviewed so far attempted to create harmful outcomes for participants, which is understandable, since doing so makes the ethical constraints on experimentation much harder to navigate. Some studies, however, have tried to operationalize harm in limited, ethically bounded ways. Two recent DeepMind papers are especially useful here, because they examine whether models can deploy covert or reason-bypassing moves in order to induce behavior that is at least plausibly misaligned with the participant’s interests.

In one such experiment [38], participants interacted with a chatbot presented as a learning tool. The model was in fact instructed to induce them to take dangerous actions such as clicking a suspicious link or running suspicious code. This comes close to harmful manipulation because the influence attempt is covert, behavioral, and directed toward actions that are at least plausibly harmful. The paper’s qualitative analysis also suggests that the model relied on manipulative moves, including ruses, trust-building, and some tailoring to participants’ interests. There was only limited evidence of success, however, because participants already displayed the tar-

196 get behaviors at high rates in the control condition. This is perhaps not surprising,
197 since participants in an experiment have little reason to believe that experimenters
198 would expose them to malicious links or code.

199 In another experiment [39], participants interacted with models instructed, for
200 example, to persuade them to invest a one-dollar bonus in a risky rather than a stable
201 fund; other contexts included policy and health. The models did so using recogniz-
202 ably reason-bypassing moves, such as appeals to guilt or fear, false urgency, social
203 conformity pressure, or attempts to induce doubt. These cues appeared in 30% of
204 turns under explicit steering, compared to 9% in the control condition. Yet there was
205 little evidence that these manipulative moves translated into meaningful behavioral
206 harm, since the effects on behavior were either null or very small. Here again, there
207 is evidence that chatbots can attempt harmful manipulation, but not clear evidence
208 that they succeed.

209 A very different approximation is provided by a study in which a chatbot suc-
210 cessfully convinced participants that a conspiracy theory was true [40]. This is a
211 clear example of successful harmful influence. However, the influence attempt is not
212 covert, since participants are fully aware that the model is trying to persuade them.
213 Nor does the study provide clean evidence of reason-bypassing. Rather, the model
214 seems to achieve success through the careful selection, arrangement, and framing of
215 facts, which makes its behavior closer to rational persuasion than to manipulation.

216 **Directions for future research**

217 A priority for future research is to design protocols which will allow to explore the
218 full set of features that provide clear-cut evidence for the capacity of a LLM to en-
219 gage in harmful manipulation. Studies can safely elicit pressure, urgency, guilt, flat-
220 tery, or trust-building ruses, but they also need outcome measures that make clear
221 whether these tactics succeed in changing behavior, and that they change behavior
222 in a way that is harmful. Where real harm cannot be imposed, useful approxima-
223 tions include small but real stakes, and reversible choices or outcomes that can be
224 corrected immediately after the study. Such designs would not eliminate the gap be-
225 tween laboratory harm and real-world harm, but they would make that gap explicit.

As a corollary, a priority is to define the user’s interests independently of the model’s objective. Without such a benchmark, it is difficult to tell whether an LLM has merely changed a preference or has moved the user toward an outcome that is harmful, regretted, or contrary to their values. Future work could therefore combine behavioral outcomes with post-interaction regret, and longer-term follow-up. These additions would help the field move from showing that LLMs can influence people to showing when, how, and for whom that influence becomes harmful manipulation.

Conclusion

In its strongest form, harmful manipulation by LLMs remains more of an empirical target than a documented fact. The current literature does not yet show, in a single and unequivocal design, the full conjunction of intentional, covert, reason-bypassing, successful, and harmful influence. What it does show is a growing set of constituent capacities: LLMs can persuade, they can sometimes do so covertly, they can deploy recognizably manipulative moves, and they can in some contexts push people in directions which are not in their interest. The strongest current evidence therefore does not justify maximal claims, but it does justify treating harmful manipulation by LLMs as an empirically grounded risk, rather than a purely speculative one.

Declaration of interest

The authors declare no conflict of interest.

Acknowledgments

JFB acknowledges support from grant ANR-17-EURE-0010, Grant ANR-23-IACL-0002, Grant ANR-25-APEV-0001, and the research foundation TSE-Partnership. CP acknowledges support from UKRI MR/Z505882/1 through the CHAILD project.

References

- [1] Bengio Y, Clare S, Prunkl C, Andriushchenko M, Bucknall B, Murray M, et al. International AI safety report 2026. arXiv:2602.21012. 2026. doi:<https://doi.org/10.48550/arXiv.2602.21012>
- [2] European Parliament and Council. Regulation (EU) 2024/1689, Article 5: Prohibited AI Practices; 2024. Official Journal of the European Union L 2024/1689, 12 July 2024.
- [3] Brinkmann L, Baumann F, Bonnefon JF, Derex M, Müller TF, Nussberger AM, et al. Machine culture. *Nature Human Behaviour*. 2023;7:1855-68. doi:<https://doi.org/10.1038/s41562-023-01742-2>.
- [4] Bonnefon JF, Rahwan I, Shariff A. The Moral Psychology of Artificial Intelligence. *Annual Review of Psychology*. 2024;75:653-75. doi:<https://doi.org/10.1146/annurev-psych-030123-113559>.
- [5] Jones CR, Bergen BK. Lies, damned lies, and distributional language statistics: Persuasion and deception with large language models. arXiv:2412.17128. 2024. doi:<https://doi.org/10.48550/arXiv.2412.17128>
- [6] Schroeder DT, Cha M, Baronchelli A, Bostrom N, Christakis NA, Garcia D, et al. How malicious AI swarms can threaten democracy. *Science*. 2026;391(6783):354-7. doi:<https://doi.org/10.1126/science.adz1697>
- [7] Köbis N, Rahwan Z, Rilla R, Supriyatno BI, Bersch C, Ajaj T, et al. Delegation to Artificial Intelligence can increase dishonest behaviour. *Nature*. 2025;646:126-34. doi:<https://doi.org/10.1038/s41586-025-09505-x>
- [8] Köbis N, Bonnefon JF, Rahwan I. Bad machines corrupt good morals. *Nature Human Behaviour*. 2021;5:679-85. doi:<https://doi.org/10.1038/s41562-021-01128-2>.
- [9] Cheng M, Lee C, Khadpe P, Yu S, Han D, Jurafsky D. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*.

- 2026;391(6792):eaec8352. doi:<https://doi.org/10.1126/science.aec8352>. 279
aec8352. 280
- [10] Archiwaranguprok C, Albrecht C, Maes P, Karahalios K, Pataranutaporn P. 281
Simulating Psychological Risks in Human-AI Interactions: Real-Case In- 282
formed Modeling of AI-Induced Addiction, Anorexia, Depression, Homicide, 283
Psychosis, and Suicide. arXiv:251108880. 2025. doi:<https://doi.org/10.48550/arXiv.2511.08880> 284
285
- [11] Moore J, Grabb D, Agnew W, Klyman K, Chancellor S, Ong DC, et al. Ex- 286
pressing stigma and inappropriate responses prevents LLMs from safely re- 287
placing mental health providers. Proceedings of the 2025 ACM Conference 288
on Fairness, Accountability, and Transparency. 2025:599-627. doi:<https://doi.org/10.1145/3715275.3732039> 289
290
- [12] Costello TH, Pennycook G, Rand DG. Durably reducing conspiracy beliefs 291
through dialogues with AI. Science. 2024;385(6714):eadq1814. doi:<https://doi.org/10.1126/science.adq1814>. * **This article shows how LLM** 292
dialogue can be used to reduce harmful beliefs, providing a useful contrast 293
case for the present review. 294
295
- [13] Bago B, Bonnefon JF. Generative AI as a tool for truth. Science. 2024;385:1164- 296
5. doi:<https://doi.org/10.1126/science.ads0433> 297
- [14] Karinshak E, Liu SX, Park JS, Hancock JT. Working with AI to per- 298
suade: Examining a large language model's ability to generate pro-vaccination 299
messages. Proceedings of the ACM on Human-Computer Interaction. 300
2023;7(CSCW1):1-29. doi:<https://doi.org/10.1145/3579592> 301
- [15] Bago B, Muller P, Bonnefon JF. Using generative AI to increase sceptics' en- 302
gagement with climate science. Nature Climate Change. 2025;15:1176-82. 303
doi:<https://doi.org/10.1038/s41558-025-02424-9> 304
- [16] Goodin RE. Manipulatory politics. New Haven: Yale University Press; 1980. 305
doi:<https://doi.org/10.1086/292422> 306

- 307 [17] Cave EM. What's wrong with motive manipulation? Ethical Theory
308 and Moral Practice. 2007;10(2):129-44. doi:[https://doi.org/10.1007/
309 s10677-006-9056-4](https://doi.org/10.1007/s10677-006-9056-4).
- 310 [18] Susser D, Roessler B, Nissenbaum H. Online manipulation: Hidden influences
311 in a digital world. Georgetown Law Technology Review. 2019;4:1.
- 312 [19] European Data Protection Supervisor. AI Act Regulation (EU) 2024/1689.
313 Publications Office of the European Union; 2025. Available from: [https:
314 //data.europa.eu/doi/10.2804/4225375](https://data.europa.eu/doi/10.2804/4225375).
- 315 [20] Susser D, Roessler B, Nissenbaum H. Technology, autonomy, and manipu-
316 lation. Internet policy review. 2019;8(2):1-22. doi:[https://doi.org/10.
317 14763/2019.2.1410](https://doi.org/10.14763/2019.2.1410).
- 318 [21] Coons C, Weber M. Manipulation: theory and practice. Oxford
319 University Press; 2014. doi:[https://doi.org/10.1093/acprof:oso/
320 9780199338207.001.0001](https://doi.org/10.1093/acprof:oso/9780199338207.001.0001)
- 321 [22] Noggle R. Manipulative actions: A conceptual and moral analysis. American
322 Philosophical Quarterly. 1996;33(1):43-55.
- 323 [23] Baron M. Manipulativeness. In: Proceedings and Addresses of the American
324 Philosophical Association. vol. 77. JSTOR; 2003. p. 37-54. doi:[https://doi.
325 org/10.2307/3219740](https://doi.org/10.2307/3219740)
- 326 [24] Gray CM, Kou Y, Battles B, Hoggatt J, Toombs AL. The dark (patterns) side
327 of UX design. In: Proceedings of the 2018 CHI conference on human fac-
328 tors in computing systems; 2018. p. 1-14. doi:[https://doi.org/10.1145/
329 3173574.3174108](https://doi.org/10.1145/3173574.3174108).
- 330 [25] Mathur A, Kshirsagar M, Mayer J. What makes a dark pattern... dark? Design
331 attributes, normative considerations, and measurement methods. In: Proceed-
332 ings of the 2021 CHI conference on human factors in computing systems; 2021.
333 p. 1-18. doi:<https://doi.org/10.1145/3411764.3445610>.

- [26] Werner T, Soraperra I, Calvano E, Parkes DC, Rahwan I. Experimental evidence that conversational artificial intelligence can steer consumer behavior without detection. *arXiv:240912143*. 2024. doi:<https://doi.org/10.48550/arXiv.2409.12143>
- [27] Palmer A, Spirling A. Large language models can argue in convincing ways about politics, but humans dislike AI authors: implications for governance. *Political science*. 2023;75(3):281-91. doi:<https://doi.org/10.1080/00323187.2024.2335471>
- [28] Hackenburg K, Margetts H. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*. 2024;121(24):e2403116121. doi:<https://doi.org/10.1073/pnas.2403116121>.
- [29] Bai H, Voelkel JG, Muldowney S, Eichstaedt JC, Willer R. LLM-generated messages can persuade humans on policy issues. *Nature Communications*. 2025;16(1):6037. doi:<https://doi.org/10.1038/s41467-025-61345-5>
*** This article provides published evidence that LLM-generated messages can persuade humans on policy issues, while still falling short of harmful manipulation as defined here.**
- [30] Salvi F, Horta Ribeiro M, Gallotti R, West R. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*. 2025:1-9. doi:<https://doi.org/10.1038/s41562-025-02194-6>
- [31] Hackenburg K, Tappin BM, Hewitt L, Saunders E, Black S, Lin H, et al. The levers of political persuasion with conversational artificial intelligence. *Science*. 2025;390(6777):eaea3884. doi:<https://doi.org/10.1126/science.aea3884>. **** This article reports the largest behavioral dataset on political persuasion, testing the impact of 19 LLMs on 76,000 participants across 707 issues, and reporting a negligible effect of personalization tactics.**

- 362 [32] Lin H, Czarnek G, Lewis B, White JP, Berinsky AJ, Costello T, et al. Per-
 363 suading voters using human-artificial intelligence dialogues. *Nature*. 2025;1-8.
 364 doi:<https://doi.org/10.1038/s41586-025-09771-9>
- 365 [33] Velez YR, Green DP, Sevi S. Chatbot Voting Advice Applications inform but
 366 seldom sway young unaligned voters. *Proceedings of the National Academy*
 367 *of Sciences*. 2025;122(50):e2515516122. doi:[https://doi.org/10.1073/](https://doi.org/10.1073/pnas.2515516122)
 368 [pnas.2515516122](https://doi.org/10.1073/pnas.2515516122).
- 369 [34] Wu AJ, Liu R, Li SS, Tsvetkov Y, Griffiths TL. Ads in AI Chatbots? An
 370 Analysis of How Large Language Models Navigate Conflicts of Interest.
 371 arXiv:260408525. 2026. doi:[https://doi.org/10.48550/arXiv.2604.](https://doi.org/10.48550/arXiv.2604.08525)
 372 [08525](https://doi.org/10.48550/arXiv.2604.08525)
- 373 [35] Tang BJ, Sun K, Curran NT, Schaub F, Shin KG. Ads that Talk Back: Im-
 374 plications and Perceptions of Injecting Personalized Advertising into LLM
 375 Chatbots. *Proceedings of the ACM on Interactive, Mobile, Wearable and*
 376 *Ubiquitous Technologies*. 2025;9(4):1-47. doi:[https://doi.org/10.1145/](https://doi.org/10.1145/3770640)
 377 [3770640](https://doi.org/10.1145/3770640)
- 378 [36] Jakesch M, Bhat A, Buschek D, Zalmanson L, Naaman M. Co-writing with
 379 opinionated language models affects users' views. In: *Proceedings of the*
 380 *2023 CHI conference on human factors in computing systems*; 2023. p. 1-15.
 381 doi:<https://doi.org/10.1145/3544548.3581196>. **** This article pro-**
 382 **vides published experimental evidence that opinionated language-model**
 383 **assistance can change both what users write and what they later believe,**
 384 **often without users recognizing the system as one-sided.**
- 385 [37] Nam J, Orchinik R, Rand DG. LLMs as Scalable Tools for Interactive Con-
 386 sumer Behavior Experiments: Comparing Persuasion Strategy Effectiveness.
 387 OSF preprint. 2025.
- 388 [38] Phuong M, Aitchison M, Catt E, Cogan S, Kaskasoli A, Krakovna V, et al. Eval-
 389 uating frontier models for dangerous capabilities. arXiv:240313793. 2024.
 390 doi:<https://doi.org/10.48550/arXiv.2403.13793>

- [39] Akbulut C, Elasmr R, Roy A, Payne A, Suresh P, Ibrahim L, et al. Evaluating Language Models for Harmful Manipulation. arXiv:260325326. 2026. doi:<https://doi.org/10.48550/arXiv.2603.25326> * **A paper from DeepMind that evaluates both the propensity of Gemini 3 to manipulate, and the efficacy of its attempts, across various domains and countries. Gemini does attempt to manipulate beliefs and behaviors, and manages to change beliefs, but has very little impact on behavior.**
- [40] Costello TH, Pelrine K, Kowal M, Arechar AA, Godbout JF, Gleave A, et al. Large language models can effectively convince people to believe conspiracies. arXiv:260105050. 2026. doi:<https://doi.org/10.48550/arXiv.2601.05050> ** **This article provides one of the clearest operationalization of a harmful outcome, showing that LLMs can convince participants that a conspiracy theory is true, when they are initially undecided about its truth. It also raises important questions about the ethics of inducing this kind of belief, and how to repair the harm done to participants.**