Are Large Language Models Manipulating Us? An Analysis Based on Susser, Roessler, & Nissenbaum (2019)

By Sterling Williams-Ceci (Cornell Tech)

Are Large Language Models Manipulating Us? An Analysis Based on Susser, Roessler, & Nissenbaum (2019)

By Sterling Williams-Ceci (Cornell Tech)

Introduction

Alongside the rapid advancement of Large Language Models (LLMs) – a form of Generative AI that can produce human-like text in response to queries – people have begun to interact with LLMs for information needs in several contexts. People most frequently talk to ChatGPT for educational purposes and professional tasks (Chatterji et al., 2025), while other major LLM chatbots are used for companionship and emotional support (Manoli et al., 2025). Moreover, people are using LLM-based applications to help them in their written communication, such as AI writing assistants like Grammarly (Grammarly, 2025). Some academic journals are even integrating LLMs to help authors write their manuscripts; Nature recently announced its new “Research Assistant,” which scans an uploaded draft and provides suggestions on how to improve it (Springer Nature, 2025).

Importantly, LLMs are known to produce responses that are far from neutral. Several audit studies have found that LLMs generate text with politically biased viewpoints, a phenomenon that mainly arises from the overrepresentation of biased viewpoints in the models’ training data (Westwood et al., 2025; Feng et al., 2023). For instance, many popular LLMs produce responses that have a liberal political slant (Germani & Spitale, 2025; Westwood et al., 2025). LLMs have also demonstrated identity-relevant biases against certain racial groups, with one study showing that several LLMs had an anti-Chinese bias when asked to evaluate identical messages attributed to Chinese individuals versus white individuals (Germani & Spitale, 2025).

LLMs' biases have sparked research on how these models affect the people who interact with them. Across multiple studies, multiple sociopolitical issues, and multiple modes of interaction, myself and other researchers have found evidence that LLMs' biased responses can influence people's attitudes and beliefs (Bai et al., 2025; Boissin et al., 2025; Salvi et al., 2025; Williams-Ceci et al., 2025a; Costello et al., 2024; Hackenburg & Margetts, 2024; Karinshak et al., 2023). Some of these studies have focused on perceptions of persuasive messages in isolation, and found that those generated by LLMs were more effective in shifting people's attitudes than ones written by humans (Bai et al., 2025; Karinshak et al., 2023). However, these studies did not show participants that these messages were coming from LLMs – would people still change their opinions if they knew they were communicating with AI? In response, other studies have asked participants to engage in a live conversation with a persuasive LLM-powered chatbot intended to reduce their belief in polarizing issues such as conspiracy theories, and have found that participants adjust their beliefs to align with the chatbot's stance (Boissin et al., 2025; Costello et al., 2024). However, many of the contexts in which people interact with LLMs do not resemble these persuasive debates, but instead entail people using LLMs in an assistive role, such as when people use AI writing assistants – can LLMs' biased responses still influence people's attitudes even when they are in these subordinate roles? My research with my lab at Cornell Tech suggests that they can. We instructed participants to write about their own attitudes toward societal issues, and randomly assigned some of these participants to receive an "AI writing assistant" that provided biased autocomplete suggestions to help them write. We found that the participants who received these LLM suggestions expressed attitudes in a separate survey that were more aligned with the biased positions we prompted the LLM with, and, strikingly, people were unaware of the suggestions' bias and influence (Williams-Ceci et al., 2025a).

These studies may sound alarm that LLMs are manipulating human cognition en masse. But do these cognitive impacts of LLMs constitute *manipulation*? What is manipulation, anyway?

Defining and Applying the Concept of Manipulation

To answer these questions, I read a paper by Daniel Susser, Beate Roessler, and Helen Nissenbaum (2019) entitled “Online Manipulation: Hidden Influences in a Digital World”. According to Susser et al., manipulation is distinct from persuasion, coercion, and related forms of influence due to its *covert* nature.

Persuasion involves “offer[ing] arguments or incentives” to change a person’s decision, while “appeal[ing] openly to their capacity for conscious deliberation and choice” (p. 14). The authors give the example of a car salesperson who tries to persuade her employees to work longer hours by arguing that people will shop later in the day in summer months, which would ultimately increase profits for everyone. The salesperson is giving her employees “resistable incentives” (p. 15) to change their decision in favor of her desired outcome. *Coercion* differs from persuasion in that a person is given “irresistible incentives” or is “deprived of choice,” which binds the person into making the focal decision due to eliminating all other courses of action (p. 15). Using the car salesperson example, the authors write that coercion could occur if the salesperson “threaten[s] her team members’ jobs, or she could put a gun to their heads” (p. 15). Even though coercion is more binding than persuasion, both phenomena “leave [the person’s] capacity for conscious decision-making intact” (p. 15). In other words, both persuasion and coercion are grounded in the assumption that the decision-maker understands what is happening to them and how they are being led to act.

Manipulation, however, differs in that it operates when the person is unaware of either how their decision space is being altered or of the influence attempt itself. Susser et al. write that manipulation “impos[es] a hidden or covert influence on another person’s decision-making” (p. 26). The manipulator changes the person’s decision space in a way that surpasses the person’s awareness and understanding. One way in which manipulation can be enacted is by deceiving people into believing they are being subject to coercion: for example, the car salesperson could manipulate people into working more hours “if she told her team they would be fired for refusing to work late but in fact lacked the authority to carry out the threat, or if she held a convincing, but fake, gun to their heads

[...]" (p. 21). Although the employees would perceive coercion in these cases, Susser et al. argue that this process would constitute manipulation, because the employees' decision-making would be tainted by these false beliefs (i.e., they would be unaware of their true choice space and alternatives). Manipulation can also occur when deciders remain unaware of the influence attempt to begin with; this phenomenon is similar to the "peripheral route" style of persuasion defined by Petty and Cacioppo in their Elaboration Likelihood Model (1986), in which advertisers can entice consumers to spend money on products not by highlighting benefits of the products themselves, but by including cues that impress us indirectly (e.g. endorsements from celebrities or authoritative figures).

Oftentimes, scholars refer to the studies on AI's ability to shift people's attitudes and beliefs as showing AI-induced "persuasion" (Bai et al., 2025; Salvi et al., 2025; Karinshak et al., 2023). However, based on Susser et al.'s definitions, I argue that some of these instances should be referred to as AI-induced "manipulation" instead – namely, when LLMs are shifting how people think in a way that they do not even realize. For instance, the studies showing that participants reduced their belief in conspiracy theories after talking to a persuasive LLM chatbot (Boissin et al., 2025; Costello et al., 2024) are examples of AI's persuasion, because the chatbot made particular beliefs more appealing than others by providing incentives for holding these beliefs (i.e., by telling the participants that denying conspiracy theories would be more accurate or more socially acceptable). The researchers also instructed the participants to debate with the chatbot about the conspiracy theories after they wrote about their own views, meaning that participants were aware of the LLM's goal. On the other hand, my team's work showing that people's attitudes were shifted by biased LLM-generated writing suggestions (Williams-Ceci et al., 2025a) could be considered to represent manipulation, for two reasons: (1) the LLM generated suggestions that logically completed whatever sentence the participant was writing about their own attitudes, which could have deceived participants into believing that the LLM's biased positions logically followed from their own; and (2) participants in these studies remained unaware that the suggestions even had biases to begin with, when we asked them afterward. Researchers should be mindful of these nuances in how AI can influence people's cognition when describing their findings to the general public; it is important to

understand when people are aware of how AI is shifting their thoughts, versus when AI acts upon their thoughts in ways that evade their autonomy.

Initial Answers Engender Deeper Questions

The landscape of LLMs' cognitive impacts is still rapidly evolving. Every month, new studies document yet more ways in which interacting with an LLM can alter people's cognition and behavior. Some of these emerging findings could constitute manipulation under the Susser et al. (2019) framework. However, identifying LLM's influence as manipulation is merely a starting point, one that raises deeper existential questions.

For example, who is the manipulator in these instances – the LLM itself, its developers, or a third party? Although LLMs are inanimate agents, my own work and others' has shown that people can be easily misled into believing that LLMs have genuine human mental capabilities (Peter et al., 2025; Williams-Ceci et al., 2025b; Colombatto & Fleming, 2024), inferences that can be triggered by LLMs' increasingly human-like language (Peter et al., 2025), and even by mere visual features conveying a human presence (Williams-Ceci et al., 2025b). At the same time, LLM chatbots and other applications are increasingly being designed with anthropomorphic features, a trend that has been evident from sites like Replika and Character.ai which create bots to simulate social actors (Maeda & Quan-Haase, 2024). As LLMs increasingly resemble humans and elicit perceptions of genuine human intelligence, people may start delegating more of the blame to the LLMs themselves for their manipulation. But this possibility raises further questions about how blame will lead to remedy; LLMs are non-sentient agents that cannot consciously change their behavior in the way that humans can.

In addition, is all LLM-induced manipulation inherently bad because it evades people's awareness? Some scholars have argued that LLMs' ability to covertly shift people's attitudes can promote societal goals that they believe to be inherently beneficial, such as reducing conspiracy theory beliefs (Boissin et al., 2025; Costello et al., 2024) and increasing beneficial health intentions (Seghal et al., 2025). Because people who hold strong attitudes on such polarizing issues are known to resist overt persuasion attempts (Petty et al., 2004), the covert nature of LLMs' manipulation may be seen as a new frontier

in promoting these societal goals. However, these instances could also represent *paternalism*, a process in which an institution interferes with an individual's autonomy in the process of trying to protect them from what is perceived to be harm (Feinberg, 1971). In cases like these, there is an additional debate around whether attempting to shift people's beliefs in evidence-based directions is ethical, despite its infringement on their freedom of thought.

These differential perspectives on the ethics of LLM-induced manipulation are producing debates about how society should respond. One common viewpoint is that governments should create legal regulations that constrain LLMs' pretraining, uses, and liability for generated content. For example, the U.S. President signed an executive order prohibiting the "intentional encod[ing] of partisan or ideological judgments" into LLMs (Trump, 2025), although this order targeted agendas associated with the political left. Additionally, regulations like Bill A6578 (currently pending in New York State legislature) require transparent disclosures from AI developers about the nature of the training data used in their LLMs (NCSL, 2025), so that people can be more cognizant of potential bias when using these models. On the other hand, some may advocate for a more libertarian or laissez-faire approach to AI, one in which AI is not heavily regulated by governments and people must bear the consequences (Poncibo & Ebers, 2025). Regardless, understanding how LLMs can manipulate cognition is necessary to making these governance decisions, which will likely plague our society for years to come.

References

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications, 16*. https://doi.org/10.1038/s41467-025-61345-5

Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when

the AI is perceived as human. *PNAS Nexus, 4*(11). https://doi.org/10.1093/pnasnexus/pgaf325

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y. & Wadman, K. (2025). How People Use ChatGPT. *[NBER Working Paper no. 34255].* https://doi.org/10.3386/w34255

Colombatto, C. & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness, 2024*(1). https://doi.org/10.1093/nc/niae013

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

Feinberg J. (1971). Legal Paternalism. *Canadian Journal of Philosophy, 1*(1), 105–124. https://doi.org/10.1080/00455091.1971.10716012

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics,* 11737–11762. Association for Computational Linguistics. https://www.doi.org/10.18653/v1/2023.acl-long.656

Germani, F., & Spitale, G. (2025). Source framing triggers systematic bias in large language models. *Science Advances, 11*(45). https://www.doi.org/10.1126/sciadv.adz2924

Grammarly (2025, July 17). “Latimer.AI and Grammarly Partner To Deliver the Industry’s First Inclusive AI Writing Assistant Bundle for Higher Education.” *Grammarly.* https://www.grammarly.com/blog/company/latimer-ai-partnership/

Hackenburg, K. & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models, *PNAS, 121*(24). https://doi.org/10.1073/pnas.2403116121

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. In *Proceedings of the ACM on Computer-Human Interaction, 7*(CSCW1), 1–29. https://doi.org/10.1145/3579592

Maeda, T. & Quan-Haase, A. (2024). When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 1068–1077. https://doi.org/10.1145/3630106.3658956

Manoli, A., Pauketat, J. V. T., & Anthis, J. R. (2025). Characterizing Relationships with Companion and Assistant Large Language Models. *CSCW Companion ‘25: Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing,* 312–319. https://doi.org/10.1145/3715070.3749245

NCSL (2025, July 10). “Artificial Intelligence 2025 Legislation.” *National Conference of State Legislatures.* https://www.ncsl.org/technology-and-communication/artificial-intelligence-2025-legislation

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

Petty, R. E. & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. Advances in Experimental Social Psychology, 19, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

Petty, R.E., Tormala, Z., & Rucker, D. D. (2004). Resistance to persuasion: An attitude strength perspective. In J. T. Jost, M. R. Banaji, & D. A Prentice (Eds.) Perspectivism in social psychology: The yin and yang of scientific progress (pp. 37-51). American Psychological Association.

Poncibo, C. & Ebers, M. (2025). Comparative perspectives on the regulation of large language models. *Cambridge Forum on AI: Law and Governance, 1*. https://doi.org/10.1017/cfl.2024.14

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

Sehgal, N. K. R., Rai, S., Tonneau, M., Agarwal, A. K., Cappella, J., Kornides, M., Ungar, L., Buttenheim, A., & Guntuku, S. C. (2025). Conversations with AI Chatbots Increase Short-Term Vaccine Intentions But Do Not Outperform Standard Public Health Messaging. https://arxiv.org/abs/2504.20519

Springer Nature (2025). “Nature Research Assistant.” https://natureresearchassistant.com/

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online manipulation: Hidden Influences in a Digital World. *Georgetown Law Technology Review, 4*(1), 1-46. https://georgetownlawtechreview.org/online-manipulation-hidden-influences-in-a-digital-world/GLTR-01-2020/

Trump, D. J. (2025, July 23). Preventing Woke AI in the Federal Government. *The White House.* https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/

Westwood, S. J., Grimmer, J., & Hall, A. B. (2025). Measuring Perceived Slant in Large Language Models Through User Evaluations. [working paper]. https://modelslant.com/paper.pdf

Williams-Ceci, S., Jakesch, M., Bhat, A., Kadoma, K., Zalmanson, L., & Naaman, M. (2025a). Biased AI Writing Assistants Shift Users' Attitudes on Societal Issues. *Science Advances, 12*(11). https://www.science.org/doi/10.1126/sciadv.adw5578

Williams-Ceci, S., Minkowitz, R. H., Zalmanson, L., Macy, M. W., & Naaman, M. (2025b). Anthropomorphic AI Chatbots Reinforce the Credibility of a Sexist Stereotype Among Conservative Women. https://osf.io/preprints/psyarxiv/eqjsg_v2

Sterling Williams-Ceci
DLI Doctoral Fellow
Cornell Tech
scw222@cornell.edu
More Info >