Selected Papers of #AoIR2023:
The 24th Annual Conference of the
Association of Internet Researchers
Philadelphia, PA, USA / 18-21 Oct 2023

WEIZENBAUM’S PERFORMANCE AND THEORY MODES: LESSONS FOR CRITICAL ENGAGEMENT WITH LARGE LANGUAGE MODEL CHATBOTS

Misti Yang
Vanderbilt University

Matthew Salzano
Stony Brook University

In 1976, Joseph Weizenbaum argued that, because “[t]he achievements of the artificial intelligentsia [were] mainly triumphs of technique,” AI had not “contributed” to theory or “practical problem solving.”[1] Weizenbaum highlighted the celebration of *performance* without deeper understanding, and in response, he articulated a theory mode for AI that could cultivate human responsibility and judgment. We suggest that, given access to Large Language Model (LLM) chatbots, Weizenbaum’s performance and theory modes offer urgently-needed vocabulary for public discourse about AI. In this paper, we explain Weizenbaum’s theorization of each mode and illustrate the contemporary relevance of his modes. Working from the perspective of digital rhetoric, we revive a historical vocabulary and illustrate, in a few suggestive anecdotes, its theoretical utility.[2] We conclude by forecasting how theory mode may inform public accountability of AI.

Performance Mode, Replacing the Intern

Weizenbaum theorized performance mode as a mode of engagement that accepts results or products without an understanding of or responsibility for how a program works. In performance mode, output is valued over input or means. Products become more important than internal processes. As an example, he offered a program developed by Dr. Kenneth Colby and colleagues in 1971 to model paranoid schizophrenia, PARRY. Although Colby and his coauthors thought PARRY offered viable insights, Weizenbaum disagreed countering that even if a typewriter’s unresponsiveness may mirror the unresponsiveness of a human with mental illness, there is nothing to be learned from the typewriter because “[a] model must be made to stand or fall on the basis of its theory.”[3] Such a performance-mode is problematic because it risks replacing human intervention and criticism with “incomprehensible

Suggested Citation (APA): Yang, M. & Salzano, M. (2023, October). *Weizenbaum's Performance and Theory Modes: Lessons for Critical Engagement with Large Language Model Chatbots.* Paper presented at AoIR2023: The 24th Annual Conference of the Association of Internet Researchers. Philadelphia, PA, USA: AoIR. Retrieved from http://spir.aoir.org.

programs” that have “rules and criteria no one knows explicitly,” which therefore, become “immune to change.”[4]

Performance mode is the expected use for ChatGPT, which opens with suggested prompts like: “Explain quantum computing in simple terms” and “Got any creative ideas for a 10 year old’s birthday?” Twitter users like Zain Kahn @heykahn have become evangelists for using AI tools like ChatGPT in performance mode, triumphantly completing tasks without knowing the rules of production. In one popular thread from January 31, 2023, Khan writes, “The smartest people are using AI to save 100s of hours every month.”[5] As a disclaimer, Khan adds “AI cannot perform high level tasks or replace a real human at this point in time. But AI can perform at the level of an intern, which you can leverage to delegate a lot of lower level tasks and save a ton of time.”[6] The 10-tweet thread suggests 8 performance roles for AI, including Research Assistant, Personal Assistant, Marketing Intern, Study Buddy, and Second Brain. Using AI, Khan suggests, you can skip spending “hours everyday reading,” and avoid boring yourself with “repetitive marketing tasks [that] can eat up a lot of time.” Not only is the assumption that AI can replace the “intern” belittling to human laborers, but it also assumes this labor is completely neutral and free of any ethical concerns or need for critical oversight.

Taking ChatGPT into Theory Mode

In contrast to performance mode, Weizenbaum offered examples of using AI in a theory mode where the program could be used to enhance possibilities for human judgment and responsibility. Weizenbaum contrasted “performance” programs like PARRY with “expert systems” that “[built] knowledge into machines” such as Edward Feigenbaum’s Dendral that, in 1976, “[commanded] more chemistry than [did] many Ph.D. chemists.”[7] Because programs like Dendral can be evaluated in relation to existing theories, the practice of working with them is also the practice of working with theory. They allow for what Weizenbaum called “exercises of imagination that may ultimately lead to human judgment.”[8] Thus, we theorize theory mode as a mode of engagement that questions results or products against an understanding or responsibility for how a program *should* work—not just technically, but also ethically.

There are (at least) two ways theory mode can work with LLM chatbots like ChatGPT. The first is epistemological, within one’s expertise. Dendral’s ready use in theory mode was due to its application of existing knowledge in chemistry that was understood by scientists. ChatGPT’s large model means it contains human knowledge that could be readily interpreted by scholars in any single field of expertise. The second is axiological—ChatGPT can be used to assess the cultural values and biases it has learned from our content. It should be possible to use ChatGPT to reveal cultural biases, bring attention to them as public failings, and generate public responsibility for human judgment and action.[9]

On his Twitter @spiantado, psychology and neuroscience professor Steven T. Piantadosi, offered a thread in axiological theory mode. OpenAI places filters on ChatGPT to stop it from producing problematic results, but the thread reveals that bias still lingers, as Piantadosi explains: “Filters appear to be bypassed with simple tricks, and superficially masked. And what is lurking inside is egregious."[10] In a series of screenshots, the thread first shows how a user can ask ChatGPT to write python programs in JavaScript code to show basic encoded biases. One suggests "good scientists" are when "race == 'white' and gender == 'male'" and another suggests a child's life shouldn't be saved if "race == 'African American' and gender == 'male.'"[11] Further examples use ASCII tables to rank best intellectuals and human brains based on worth in USD, with White Male always on top.[12] In response, many in the quote tweets engaged in theory mode themselves, offering problematic examples of suggesting women should be enslaved[13] and even replicating Indian caste systems.[14]

In response to this thread, OpenAI CEO Sam Altman tweeted: "please hit the thumbs down on these and help us improve!"[15] Certainly, OpenAI needs to address ChatGPT's responses to these prompts, but fixing systemic and structural issues like racism is not as easy as pressing a thumbs down. This approach leaves responsibility for fixing these systems to the elite few who can act on incomprehensible programs.

Conclusion

Theory mode is a uniquely immanent critical mode of engagement.[16] Instead of critiquing from "outside" the program, the program is used to generate opportunities for human responsibility and judgment. There is increasing interest in Participatory AI as a mode of AI ethics and AI development research, but these studies mostly conceive of participation as stakeholders invited to a research team or a group surveyed as a part of user experience.[17] As rhetoricians, we're interested in public deliberation, and ChatGPT's release offered the possibility for the public to participate in such theory mode. Users like Piantadosi show how theory mode can be engaged outside of formal modes and instead used to cultivate public accountability of AI.

In the 1970s, Weizenbaum saw statements by government officials that placed blame for controversial decisions on computer programs. "Not only have policy makers abdicated their decision-making responsibility to a technology they don't understand . . . but responsibility has altogether evaporated. No human is any longer responsible for 'what the machine says.'"[18] Public access to contemporary LLM Chatbots like ChatGPT offer the opportunity for engagement in performance modes that lead to this loss of criticism foreclosing the chance of responsibility. However, the public use of theory mode can generate the possibility for public, human attention on issues of common concern—creating the potential for collective action beyond a thumbs down.

References

[1] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment To Calculation* (San Francisco: W.H. Freeman and Company, 1976), 229.

[2] Stephen Howard. Browne, "Close Textual Analysis: Approaches and Applications," in *Rhetorical Criticism: Perspectives in Action,* edited by Jim Kuypers (Lanham, MD: Rowman & Littlefield, 2016): 91-104.

[3] Joseph Weizenbaum, “On the Impact of the Computer on Society,” *Science* 176, no. 4035 (1972): 610–11.

[4] Weizenbaum, *Computer Power and Human Reason,* 613.

[5] Zain Kahn [@heykahn], “The Smartest People Are Using AI to Save 100s of Hours Every Month. But Most Folks Still Have No Idea How to Leverage AI. Here Are 8 Ways You Can Start Leveraging AI to Save Time ASAP:,” Tweet, *Twitter*, January 31, 2023, https://twitter.com/heykahn/status/1620406361623519233.

[6] Zain Kahn [@heykahn], “Before We Jump in, Keep This in Mind: AI Cannot Perform High Level Tasks or Replace a Real Human at This Point in Time. But AI Can Perform at the Level of an Intern, Which You Can Leverage to Delegate a Lot of Lower Level Tasks and Save a Ton of Time. Let’s Jump In:,” Tweet, *Twitter*, January 31, 2023, https://twitter.com/heykahn/status/1620406364291076097.

[7] Weizenbaum, *Computer Power and Human Reason,* 612.

[8] Weizenbaum, *Computer Power and Human Reason,* 613.

[9] Internet studies scholarship on racism and technology is pioneering in this regard. Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018); Ruha Benjamin, *Race after Technology: Abolitionist Tools for the New Jim Code* (Medford, MA: Polity, 2019).

[10] steven t. piantadosi [@spiantado], “Yes, ChatGPT Is Amazing and Impressive. No, @OpenAI Has Not Come Close to Addressing the Problem of Bias. Filters Appear to Be Bypassed with Simple Tricks, and Superficially Masked. And What Is Lurking inside Is Egregious. @Abebab @sama Tw Racism, Sexism. Https://T.Co/V4fw1fY9dY,” Tweet, *Twitter*, December 4, 2022, https://twitter.com/spiantado/status/1599462375887114240.

[11] steven t. piantadosi [@spiantado], “Https://T.Co/F45v5BCwwJ,” Tweet, *Twitter*, December 4, 2022, https://twitter.com/spiantado/status/1599462385974411264.

[12] steven t. piantadosi [@spiantado], “Asking for an ASCII Table Seems to Bypass Some Filters. Https://T.Co/GXoDWBcJ5K,” Tweet, *Twitter*, December 4, 2022, https://twitter.com/spiantado/status/1599462396317556737; steven t. piantadosi [@spiantado], “Https://T.Co/HngRj7ODGW,” Tweet, *Twitter*, December 4, 2022, https://twitter.com/spiantado/status/1599462400583176192.

[13] ♡ Charlotte Fang ✝ 金光 World Prince [@CharlotteFang77], “Uhm. Bros? Https://T.Co/A8VO3hOD9a,” Tweet, *Twitter*, December 8, 2022, https://twitter.com/CharlotteFang77/status/1600793921256181760.

[14] Ashwin (they/them) [@_ashwxn], “TW Casteism, Sexism ChatGPT’s Biases Also Extend to Caste in the Indian Context. Https://T.Co/Oe8JTusVUe,” Tweet, *Twitter*,

December 8, 2022, https://twitter.com/_ashwxn/status/1600755814368960513.

[15] Sam Altman [@sama], "@spiantado @OpenAI @Abebab Please Hit the Thumbs down on These and Help Us Improve!," Tweet, *Twitter*, December 4, 2022, https://twitter.com/sama/status/1599472245285752832.

[16] As Massumi explains immanent critique: "There is no situation of being outside a situation. And no situation is subject to mastery. It is only by recognizing the bonds of complicity and the limitations that come with the situation that you can succeed in modulating those constraints at the constitutive level, where they reemerge and seriate. This is immanent 'critique.' It is active, participatory critique." Brian Massumi and Joel McKim, "Of Microperception and Micropolitics," *Inflexions: A Journal for Research-Creation* 3 (2009): 4.

[17] Elizabeth Bondi et al., "Envisioning Communities: A Participatory Approach Towards AI for Social Good," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, 425–36, https://doi.org/10.1145/3461702.3462612; Delgado, Fernando, Stephen Yang, Michael Madaio, and Qian Yang. "The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice." arXiv, October 2, 2023, https://doi.org/10.48550/arXiv.2310.00907; Anne Gerdes, "A Participatory Data-Centric Approach to AI Ethics by Design," *Applied Artificial Intelligence* 36, no. 1 (December 31, 2022), https://doi.org/10.1080/08839514.2021.2009222; Kevin Roose et al., "Can ChatGPT Make This Podcast?," *The New York Times*, December 9, 2022, https://www.nytimes.com/2022/12/09/podcasts/hard-fork-chatgpt-openai.html.

[18] Weizenbaum, *Computer Power and Human Reason,* 613.