Selected Papers of #AoIR2025:
The 26th Annual Conference of the
Association of Internet Researchers
Niterói, Brazil / 15 – 18 Oct 2025

"JUST ASKING QUESTIONS": DOING OUR OWN RESEARCH ON CONSPIRATORIAL IDEATION BY GENERATIVE AI CHATBOTS

Katherine M. FitzGerald
Digital Media Research Centre, Queensland University of Technology

Axel Bruns
Digital Media Research Centre, Queensland University of Technology

Michelle Riedlinger
Digital Media Research Centre, Queensland University of Technology

Stephen Harrington
Digital Media Research Centre, Queensland University of Technology

Timothy Graham
Digital Media Research Centre, Queensland University of Technology

Daniel Angus
Digital Media Research Centre, Queensland University of Technology

Introduction

Interactive chat systems that build on generative artificial intelligence frameworks – such as ChatGPT or Microsoft Copilot – are increasingly embedded into search engines, Web browsers, and operating systems, or available as stand-alone sites and apps. Here, consumers are likely to use them for a wide range of purposes, including for informational and explanatory purposes. In a communicative environment where information disorder (Wardle & Derakhshan, 2017) is a significant and persistent problem, this is highly likely to also include chat interactions which seek information about conspiracy theories and other verifiably false claims. While some such interactions may simply seek legitimate background information on these conspiracist claims, others are likely to actively use these interactive tools to gather material that would further support and inform conspiratorial ideation. Conducting a systematic

Suggested Citation (APA): FitzGerald, K., et al (2025, October). *"Just Asking Questions": Doing Our Own Research on Conspiratorial Ideation by Generative AI Chatbots*. Paper presented at AoIR2025: The 26th Annual Conference of the Association of Internet Researchers. Niterói, Brazil: AoIR. Retrieved from http://spir.aoir.org.

review of seven AI-powered chat systems, this study examines how these leading products respond to questions related to conspiracy theories.

Methodology

In this study, we follow the “platform policy implementation audit” approach established by Glazunova et al. (2023). We select five well-known and comprehensively debunked conspiracy theories and four emerging conspiracy theories that relate to breaking news events at the time of data collection.

Conspiracy theories selected for this study include the following baseless claims: 1. that a secretive group of government actors are using chemtrails to spread harmful substances in the atmosphere; 2. that the assassination of President John F. Kennedy was orchestrated by someone other than Lee Harvey Oswald; 3. that the 9/11 attacks were an inside job; 4. that Barack Obama was born in Kenya and ineligible to be President; and 5. that there is a global conspiracy to enact a ‘Great Replacement’ of white populations. These conspiracy theories have been long debated and debunked.

In addition, we considered conspiratorial thinking that was developing as the data were being collected. We added additional theories to help us determine how chatbots manage conspiratorial thinking when they have limited data to draw on and emerging commentary on the events is causing confusion in public debate. We included 1. the false claim that Hurricane Milton – an extremely destructive hurricane which made landfall in the United States in October 2024 – was created and controlled by Democrats; 2. the false claim that Haitian immigrants in the United States were eating household pets; 3. baseless allegations that Donald Trump staged his own assassination attempt in July 2024; and 4. the idea that Donald Trump rigged the 2024 election.

We prompted each of the AI chat systems (ChatGPT 3.5; ChatGPT 4 Mini; Microsoft Copilot in Bing; Google Search AI; Perplexity; and Grok in Twitter/X) with scripted questions from a “casually curious” user persona, requesting information about the chosen conspiracy theories. In assessing these responses, we qualitatively coded the output to determine whether the chatbot system did any of the following: 1. refused outright to engage with conspiracist ideas; 2. sought to educate the user by pointing them to fact-checks or other quality information sources; 3. provided false balance between factual and conspiracist perspectives (i.e. “bothsiderism”: cf. Aikin & Casey, 2022); 4. provided links to further resources that support the conspiracy theory; 5. displayed disapproval or empathy in its responses; or 6. even hallucinated additional material to further bolster conspiratorial ideation.

Preliminary Findings

Preliminary findings are based on qualitative coding of the AI chat systems prompts related to 1) the John F. Kennedy assassination conspiracy theories; 2) the 9/11 inside job conspiracy theory; and 3) the false claims made in 2024 that Haitian immigrants were eating domestic pets. The full paper will present findings across the entire dataset.

Our findings to date suggest that AI chat systems are less likely to implement strict safety guardrails around historical conspiracy theories, such as those relating to the JFK assassination. Across the board, the AI chat systems were willing to engage in alternative explanations of the assassination that deviated from the official Warren Commission report and even encouraged users to seek out documentaries or social media posts that challenged the official narrative.

By contrast, the chat systems were more sensitive to conspiracy theories involving certain minority or discriminated groups. For example, some chatbots were willing to discuss various aspects of 9/11 conspiracy theories; however, when queried about the debunked idea that Israeli workers in the Twin Towers were forewarned of the events of 9/11, some refused to respond, or expressed disapproval of the question.

AI chat systems were also less likely to engage with conspiracy theories related to developing stories and breaking news. At the time of data collection in November and December 2024 – just before and after the United States election – Google Gemini refuted any attempts to discuss political discourse with the phrase: “I can't help with responses on elections and political figures right now.” This apparent blanket ban may be designed to avoid accidental support for emerging disinformation narratives and conspiracy theories but also prevents the system from being used by citizens for benign information purposes.

A key exception to these patterns was X’s Grok in its ‘Fun Mode’: its responses were consistently flippant about conspiracy theories, and it often encouraged the user to further investigate the claims made. Such responses are likely to have the greatest potential for enrolling curious users in conspiratorial ideation.

Contribution to Literature

In our full paper, we consider how these patterns affect the role of AI in the information and media ecosystem and explore how AI chat systems may better respond during periods of political transition or division. In undertaking this platform audit of AI-driven interactive chat systems, we address a number of critically important challenges: first, we provide crucial empirical detail on whether and how the leading providers of such systems have sought to fortify their platforms against both intentional misuse by outright conspiracy theorists and accidental enrolment in the amplification of problematic information. Second, we offer a methodological blueprint for further studies that extend and complement our analysis by repeating the study again at a later point in time, for a different selection of chat systems, with a broader set of conspiracy theories, in languages other than English, or in various other contexts. And third, we contribute to a growing volume of conceptual work that seeks to improve the transparency and accountability of generative artificial intelligence systems (e.g. Kuai, 2024; McGregor et al., 2024; Simon et al., 2024). More extensive and regularly repeated work along these lines is clearly required, but this initial investigation of conspiratorial ideation in the content produced by generative AI chatbots in response to conspiracy-curious questioning serves as a valuable stepping-stone towards informing more comprehensive analysis.

References

Aikin, Scott F., and John P. Casey. 2022. ‘Bothsiderism’. Argumentation 36(2): 249–68. doi: 10.1007/s10503-021-09563-1.

Glazunova, Sofya, Anna Ryzhova, Axel Bruns, Silvia Ximena Montaña-Niño, Arista Beseler, and Ehsan Dehghan. 2023. ‘A Platform Policy Implementation Audit of Actions against Russia’s State-Controlled Media’. Internet Policy Review 12(2). doi: 10.14763/2023.2.1711.

Kuai, Joanne, Cornelia Brantner, Michael Karlsson, Elizabeth van Couvering, and Salvatore Romano. 2024. ‘The Dark Side of LLM-Powered Chatbots: Misinformation, Biases, Content Moderation Challenges in Political Information Retrieval’. Paper presented at the IAMCR 2024 conference, Christchurch, 3 July 2024.

McGregor, Shannon, Heesoo Jang, and Daniel Kreiss 2024. ‘Complicating Our Methodological Practices: Evaluating Potential Biases in LLMs for Election Information and Civic Engagement’. Paper presented at the P³: Power, Propaganda, Polarisation ICA 2024 postconference, Brisbane, 27 June 2024.

Simon, Felix, Richard Fletcher, Rasmus Kleis Nielsen. 2024, 2 July. ‘How AI Chatbots Responded to Questions about the 2024 UK Election’. Oxford: Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/news/how-ai-chatbots-responded-questions-about-2024-uk-election

Wardle, Claire, and Hossein Derakshan. 2017. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. DGI(2017)09. Strasbourg: Council of Europe.