From Algorithms to Empathy: Exploring AI Anthropomorphism in the Age of Derived Sentience

ABIGAIL POROPATICH

Abstract: This research investigates the impact of anthropomorphic AI on individuals with traumatic experiences, focusing on its integration in virtual therapeutic modalities. Initially targeting those with childhood trauma or event-specific Post-Traumatic Stress Disorder (PTSD), the study expanded to include a diverse group from Clemson University, encompassing students with varying backgrounds and self-identified trauma. Participants underwent a three-stage process involving a pre-test, interaction with a Large Language Model (LLM), and a post-test. The pre-test revealed significant comfort with AI displaying human emotions and a preference for AI therapy over traditional psychiatry, primarily due to concerns about therapist bias and monetary greed. Participants expressed a desire for AI therapists to strike a balance between human-like empathy and objective, computer-generated responses, indicating a preference for an AI model that maintains a discernible difference from human therapists. The study utilizes a GPT model, fine-tuned with trauma-centric prompts that emphasized maintaining productive, and positive dialogue. Preliminary findings suggest a potential for increased vulnerability and emotional regulation through this AI-human integration. Participants reported positive experiences with a "conversational" canonical view, and through more edge-case adaptations and anthropomorphic features, user experiences could evolve from conversational to reformative. This exploration into AI anthropomorphism within therapy highlights the nuanced relationship between technology and mental health, offering a promising avenue for supporting individuals with traumatic experiences through innovative and empathetic AI solutions.

ACM Reference Format:
Abigail Poropatich. 2024. From Algorithms to Empathy: Exploring AI Anthropomorphism in the Age of Derived Sentience. 1, 1 (October 2024), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Introduction: The annual incidence report of Post-Traumatic Stress Disorder (PTSD) in the United States is approximately 6%. [4] Out of those twenty million dealing with the debilitating effects of the disorder, only 58% of them seek psychiatric help, and only one third of them attend regular sessions. [1] Reasons for avoiding therapy are as personal as the condition itself but may include factors such as cost, inaccessibility, or fear of judgment. Current attempts to remedy such hurdles gave rise to what we currently know as telehealth. This allows medical care to be performed through a secure online connection, removing the physical barrier between patient and physician. As it pertains to mental health, current telehealth applications include platforms like Betterhelp, which connect patients with licensed counselors or psychiatrists. The response to online therapy has been overwhelmingly positive, but it has only addressed one of the many impediments patients are facing. The introduction of Anthropomorphized Artificial Intelligence (AAI) is no foreign concept. The personification of technology is the lineament of the digital revolution. The goal of this study is to harness said humanization into a reflective, palliative entity capable of algorithmically digressing triggering events into simple memories. Coupled with the reduced costs associated with accessible technology and configurable bias mitigation, the ideal therapist can evolve from current GPT models. This study hypothesizes that providing patients

Author’s address: Abigail Poropatich.

Manuscript submitted to ACM

with Post-Traumatic Stress Disorder (PTSD) continuous access to an unbiased, anthropomorphized therapeutic AI interface will positively contribute to their mental health management. The hypothesis suggests that such constant access to a human-like, empathetic AI therapy system will enhance patients' ability to cope with PTSD symptoms, improve emotional regulation, and foster a sense of support, thereby promoting better overall mental well-being.

Related Work: In order to adequately design a rudimentary system that mimics the foundational values of AAI, it was necessary to accumulate knowledge about current AAI techniques. The general survey that was conducted on current attitudes toward the personification of AI for this study revealed that most participants were primarily concerned about ethics. Ethical concerns raised about anthropomorphism in current experimentation resolve around the potential blurring of moral and ontological boundaries. [2]. In Arleen Salles' article, *Anthropomorphism in AI*, the discussion of the epistemological implications is crucial for understanding a system's physical limitations. Salles emphasizes the importance of fully comprehending the boundaries in emulating emotion to avoid creating false expectations. The establishment of AI's emotional repertoire is, at best, algebraic, and researchers should communicate this clearly to participants, as perceptions of capabilities will influence effectiveness. While working with a vulnerable population, it is essential to understand the boundaries of the psychiatric condition being evaluated. Despite not holding an advanced degree in psychology, extensive research into the intersection of psychiatry and AI was conducted to ensure the training dataset closely emulated the knowledge and expertise of a professional. AI technologies facilitate real-time monitoring of psychiatric conditions, allowing for timely interventions and ongoing assessment of treatment effectiveness. Through both supervised and unsupervised learning, analysis of psychiatric data and dialogue can be performed to personalize treatment plans and gauge progress. [3] However, special regard must be given to managing bias in these datasets and ensuring sustained adherence to data privacy regulations.

Methodology: In this research, the methodology focused on three key areas: informing the system's design, evaluating its performance through surveys, and detailing the experimental procedure. The design of the AI chatbot, hereafter referred to as 'AI-de', utilized the OpenAI GPT model, tailored with a therapeutic dataset aimed at ensuring sensitive and appropriate responses for individuals with PTSD. The interface was intentionally plain, warm, and technically simple, minimizing distractions and focusing on the chat functionality. Evaluation was based on a three-stage process: a pre-test to establish baseline comfort and concerns regarding AI therapy, the deployment of AI-de for interaction, and a post-test to assess emotional and psychological impacts. Data collection was conducted qualitatively, with an emphasis on participants' feedback during their interaction with AI-de. To enhance the AI's empathetic capabilities while maintaining neutrality, considerable effort was devoted to de-biasing the training dataset. During the creation of the fine-tuning materials, neutrality was the overarching goal. The experimental procedure was structured to progressively engage participants with the AI-de, initially focusing on basic interactions and gradually introducing more complex therapeutic dialogues through inquisitive elaboration. This approach aimed to simulate a realistic therapeutic scenario, allowing for a comprehensive assessment of AI-de's effectiveness in a therapeutic context. The methodology was carefully crafted to balance the technical aspects of AI development with the sensitive nature of dealing with PTSD, ensuring ethical considerations were at the forefront of the system's design and experimental execution.

System Description: The current contributions of this project are finite due to a limited educational scope. The core AI model is the OpenAI GPT-3.5-turbo model, a large language model (LLM) known for its conversational capabilities. AI-de's training data is a .jsonl file, created in a single line dialogue context. Users prompt the system with emotionally distressing scenarios, and the system responds with a basic sentence structure: reassurance, solution proposition, and a follow-up question. This structure facilitates continuous conversation with a positive connotation. The user interface (UI) is intentionally minimalistic, focusing on ease of use and reducing sensory overload. It features a simple chat window with optional prompts, a chat button, and pause/resume button. The backbone of the server-side logic is encapsulated in a Node.js server file utilizing the Express framework, named 'app.js'. This file is also responsible for integrating the OpenAI API and managing routing for endpoints. On the front-end, the managing file, named 'script.js' handles client-side logic, Document Object Model (DOM) manipulation, and server communication. AI-de is deployed on Azure without a database.

Results: To accurately establish the effects of a targeted LLM on those who have experienced a traumatic event, establishing a baseline was essential. The initial assessment limited the data pool to those who had experienced childhood trauma or event-specific PTSD. However, this approach quickly proved inapplicable to the selection of participants, primarily students at Clemson University. Additionally, the lack of psychiatric diagnostic capabilities in this project precluded the establishment of a clinically validated definition of "trauma". Consequently, participants were asked to self-identify with a traumatic event. From the responses, this generalization of the word proved to be beneficial. Surveys and preliminary assessments were distributed to two distinct groups: members of the Human-Centered Artificial Intelligence (HCAI) course at Clemson University and self-identified trauma victims. The questioning of HCAI course members aimed to gather perspectives on the early phases of anthropomorphism in AI, human-AI teaming, and the integration of AI into modern therapeutics. This group of subjects has no prior knowledge of the study nor did they need to align with a trauma specific event. This allowed a general consensus to be established towards the separate sectors of HCAI and modern day psychiatry, and their subsequent integration.

The survey, comprising of seven participants, showed a predominantly positive stance towards AAI, with 57.1% rating their comfort at 4 (on a scale of 1 to 5, with 1 being extremely uncomfortable and 5 extremely uncomfortable), and 42.9% rating it at 3, indicating a moderate to high level of acceptance of human-like characteristics in AI. When asked about openness to receiving therapeutic guidance from AI, 57.1% of the respondents showed a high level of openness (rating 4), while 42.9% exhibited a moderate level of openness (rating 3). Primary concerns were more technically aligned and centered around security, data privacy, and perceived inaccuracy (bias). In discussing distressing memories, 71.4% preferred a computationally driven therapist, suggesting a comfort with the impartiality of AI in sensitive matters. This was to be expected as this group had a thorough understand of artificial intelligence.

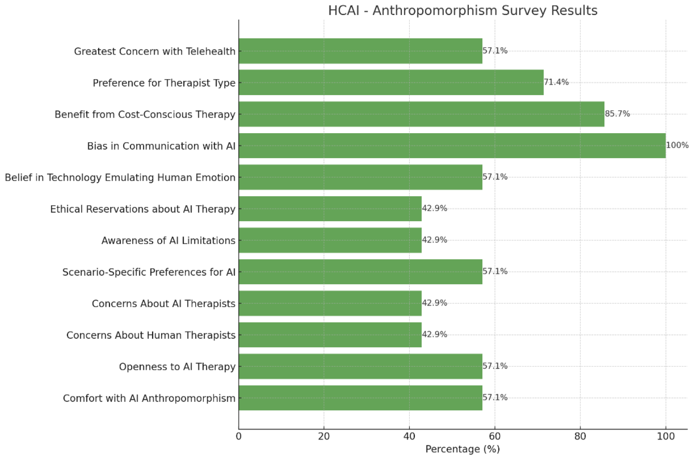

Figure 1: Results of the HCAI - Anthropomorphism Survey with selected participants from the HCAI course at Clemson University

For the second group, individuals were selected based on their willingness to disclose experiencing a trauma amenable to exploration or treatment within a therapeutic context. There was no specific definition about the severity or duration of said event. Participants agreed to a three stage study: a pre-test, deployment of the LLM and subsequent recording of their interactions with it, and a post-test. The pre-test would establish a baseline of their level of comfort with AI-de, the impact of their distressing memories, if they were under current duress, and any bias that were of concern. The survey, comprising of 11 responses, showed that 72.7% of participants never sought professional counseling services for their trauma. The various extenuating circumstances impeding them were as follows: financial constraints, time limitations, and past negative experiences with therapy. Most participants expressed a degree of comfort with AI displaying human-like characteristics, provided there was still a discernible difference. Concerns were noted about the lack of empathy, the inability to understand complex emotions, and fears regarding privacy and data security. The diverse range of opinions on AI therapy underscored the need for a nuanced approach in the design of AI-de.

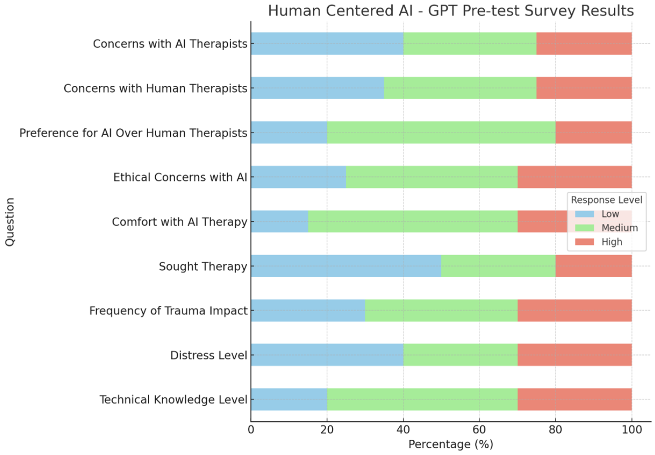

Figure 2: Results of the Human Centered AI - GPT Pre-test survey with self-identified trauma survivors

Due to the second group's commitment to interacting with AI-de, a post-test was essential to evaluate their attitudes towards the technology. When participants were asked how conversing with the artificial intelligence differed from traditional therapy, 75% agreed that knowing they were interacting with a technological entity made the conversations less impactful. However, 50% noted a positive experience with the prompting format of AI-de's responses. The remaining 50% were neutral, feeling that they were not 'listened to.' When asked about any ethical concerns that might have arisen during communication, none were reported.

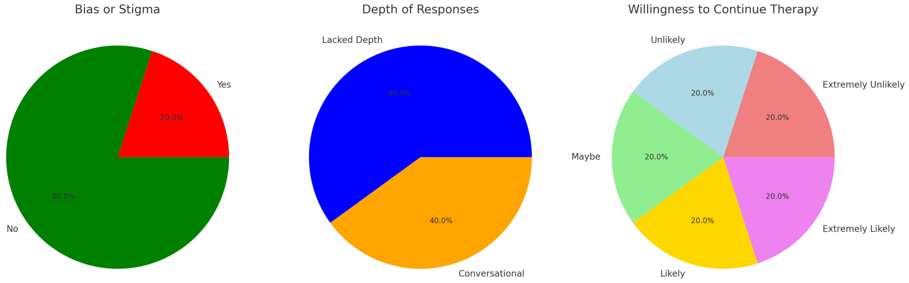

Figure 3: Results of the Human Centered AI - GPT Post-test survey with self-identified trauma survivors

Discussion: Findings suggest that AI, when fine-tuned with trauma-centric prompts, can offer a level of conversational engagement beneficial for some participants in managing emotions surrounding a traumatic event. The qualitative analysis of each participants data is unique. As the participants were self-identified and conversations were not guaranteed to stay on the topic of the aforementioned trauma, gauging the success of such an experiment as it strictly pertains to Post-Traumatic Stress Disorder is challenging. However, the positive reception of AI-de's conversational prompts indicates that the system can provide a format that is conducive to reflection and self-guided exploration of emotions. Neutral responses regarding the lack of depth of the system points to the need for edge-case training and engagement in more deep-seated, nuanced dialogues. Future deployments could incorporate advanced natural language processing techniques, primarily named-entity recognition.

AI-de's preliminary design aimed to maintain a balance between human-like empathy and objective, computer generated responses. Feedback indicates a preference for this balance, offering a therapeutic experience that is distinctly, yet positively different from human interaction, free from hindsight bias. Notably, one case of system bias was reported. While it is essential to take all reports of bias seriously, it is equally critical to differentiate between true algorithmic bias and user dissatisfaction, especially if the user's understanding of algorithmic bias is limited. The report was carefully analyzed, suggesting that the AI's response was consistent with its programming and typical of its outputs. This highlights the need for setting realistic expectations for users and providing clear communication about AI-de's limitations.

Conclusion: This study has provided a novel exploration into the realm of derived sentience and Anthropomorphic Artificial Intelligence (AAI), particularly as it pertains to the management of Post-Traumatic Stress Disorder. This elementary exploration underscores the potential along with the challenges intrinsically found in the evolution of AI within the therapeutic sphere. The preliminary investigation performed with the GPT model named 'AI-de' highlighted the diversity of recovery through dialogue and the valuable role AI can play as a supplementary conversationalist with its current scope. The experiment conducted with two distinct groups - students from Clemson University's Human-Centered Artificial Intelligence (HCAI) course and self-identified trauma survivors - illuminated the various perspective on ethical AI, data security, and potential avenues for bias. While there was a general positive consensus on the comfort with AI displaying human-like characteristics, it was also evident that maintaining a discernible difference from human therapists was a necessary part of early interaction. As participants were informed about the study's nature and engaged with questions on their opinions of sentience and a computer's ability to mimic sentience, it is important to acknowledge that these probing questions might introduce their own bias before interaction. Through further testing and prompt engineering for edge-cases, more experimentation should be done in a blind or double blind fashion to strengthen validity. Furthermore, it is important to recognize that the frontier for AI therapy is still being forged. Artificial intelligence is a testament to human innovation, notwithstanding the age-old challenge of deciphering human emotion. The pursuit of engineering a technology, born of human ingenuity, with the explicit objective of deciphering nuanced emotional states, forebodes an algorithmic nightmare. However, through continuous interdisciplinary refinement, the advancement of neural networks, and refinement of machine learning algorithms through attention mechanisms, the age of derived sentience may yet be attainable.

REFERENCES

[1] C.J. Nobles, S.E. Valentine, E.D. Zepeda, E.M. Ahles, D.L. Shtasel, and L. Marques. Usual course of treatment and predictors of treatment utilization for patients with posttraumatic stress disorder. *Journal of Clinical Psychiatry*, 78(5):e559–e566, 2017. doi: 10.4088/JCP.16m10904. PMID: 28570794; PMCID: PMC5454778.

[2] Arleen Salles, Kathinka Evers, and Michele Farisco. Anthropomorphism in ai. *AJOB Neuroscience*, 11(2):88–95, 2020. doi: 10.1080/21507740.2020.1740350.

[3] Jie Sun, Qun-Xi Dong, San-Wang Wang, Yong-Bo Zheng, Xiao-Xing Liu, Tang-Sheng Lu, Kai Yuan, Jie Shi, Bin Hu, Lin Lu, and Ying Han. Artificial intelligence in psychiatry research, diagnosis, and therapy. *Asian Journal of Psychiatry*, 87:103705, 2023. ISSN 1876-2018. doi: 10.1016/j.ajp.2023.103705. URL https://www.sciencedirect.com/science/article/pii/S1876201823002617.

[4] U.S. Department of Veterans Affairs. Ptsd in adults: Overview, 2023. URL https://www.ptsd.va.gov/understand/common/common_adults.asp. Accessed: 2023-12-02.