

The AI Placebo Effect: Anthropomorphic Attribution as a Mechanism in Digital Mental Health Interventions

Marc Augustin

Submitted to: JMIR Mental Health
on: September 22, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	4
---------------------------------	----------

Preprint
JMIR Publications

The AI Placebo Effect: Anthropomorphic Attribution as a Mechanism in Digital Mental Health Interventions

Marc Augustin¹ MD

¹ Protestant University of Applied Sciences Bochum, Germany Bochum DE

Corresponding Author:

Marc Augustin MD

Protestant University of Applied Sciences Bochum, Germany
Immanuel-Kant-Str. 18-20
Bochum
DE

Abstract

Digital therapeutic alliance research suggests that patients experience AI chatbot interactions like face-to-face therapy. However, phenomenological analysis concludes that an authentic alliance requires embodied, conscious interaction, which is impossible with AI systems. This creates a puzzle: What explains the measurable patient benefits? Drawing from human-computer interaction research, the "AI placebo effect" is proposed: therapeutic benefits arising from patients' inability to perceive complex language without attributing subjectivity behind it. This concept requires research methodologies inspired by pharmaceutical placebo studies.

(JMIR Preprints 22/09/2025:84584)

DOI: <https://doi.org/10.2196/preprints.84584>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in https://preprints.jmir.org/preprint/84584

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in https://preprints.jmir.org/preprint/84584

Original Manuscript

Personal Perspectives

The AI Placebo Effect: Anthropomorphic Attribution as a Mechanism in Digital Mental Health Interventions

Marc Augustin¹

¹Protestant University of Applied Sciences Bochum, Germany

Corresponding Author:

Marc Augustin

Protestant University of Applied Sciences Bochum
Immanuel-Kant-Str. 18-20
44803 Bochum

Tel.: +49 234-36901-474

marc.augustin@evh-bochum.de

<http://www.evh-bochum.de>

keywords:

artificial intelligence, digital mental health, chatbot, anthropomorphism, placebo effect

Abstract

Digital therapeutic alliance research suggests that patients experience AI chatbot interactions like face-to-face therapy. However, phenomenological analysis concludes that an authentic alliance requires embodied, conscious interaction, which is impossible with AI systems. This creates a puzzle: What explains the measurable patient benefits? Drawing from human-computer interaction research, the "AI placebo effect" is proposed: therapeutic benefits arising from patients' inability to perceive complex language without attributing subjectivity behind it. This concept requires research methodologies inspired by pharmaceutical placebo studies.

Introduction

Drawing from personal clinical experience and early research on therapeutic AI chatbots, a puzzle arises. On one hand, emerging research defines the digital therapeutic alliance as a “collaborative relationship and emotional connection between a user and an artificial intelligence (AI)-driven psychotherapeutic tool, encompassing goal alignment, task agreement, therapeutic bond, user engagement, and the facilitators and barriers that influence therapeutic outcomes” [1]. Users engaging with AI chatbots such as Wysa report feeling understood (e.g. “You’re the only person that helps me and listens to my problems [...]”) [2], and a first randomized controlled trial of Therabot reduced symptoms of depression and anxiety [3].

On the other hand, phenomenological analysis argues that therapeutic understanding requires “conviviality”, a shared form of life based on biological vulnerabilities and lived experience [4]. AI systems lack consciousness, a physical body, and are unable to perceive – they can only simulate perception. Therefore, they cannot participate in genuine intersubjective understanding [4]. Yet, if an authentic therapeutic alliance is impossible with AI, what explains the positive effects?

The AI Placebo Effect Hypothesis

I propose that benefits attributed to a digital therapeutic alliance actually represent an “AI placebo effect”. The therapeutic improvements arise from patients’ inability to separate contextually appropriate language from the presence of a conscious mind behind it.

Humans exhibit a strong tendency to anthropomorphize, which means attributing human-like feelings and features to inanimate objects, such as pets, robots, or AI [5]. Research suggests that this helps individuals make sense of uncertain environments and attain control [6]. Neuroimaging findings indicate that the process of attributing mental states to nonhuman agents correlates with activity in areas associated with sociocognitive mentalizing, such as vMPFC and ACC [6].

The potential of algorithmic text interactions being perceived as real and therapeutic was first explored in 1966 with ELIZA, a chatbot based on simple feedback loops [7]. As Bender notes, humans reflexively imagine minds behind complex text, even when ‘there is no mind there’ [8]. When AI chatbots respond with phrases like ‘I understand your pain,’ patients may involuntarily attribute empathy and therapeutic qualities to these algorithmic outputs.

This attribution process parallels classical placebo mechanisms: it is hypothesized that patients’ expectations and beliefs about receiving care activate therapeutic pathways, even when the “active ingredient” of genuine human understanding is absent. Like pharmaceutical placebos, the AI placebo may be caused by experiencing a simulation of active therapy, conditioning effects of past treatments, or emotional expectations [9].

Research Implications

Testing the AI placebo effect hypothesis requires methods inspired by pharmaceutical placebo research. Anthropomorphism control studies can compare identical therapeutic content delivered through highly anthropomorphic AI interfaces versus clearly non-anthropomorphic systems to isolate the specific contribution of perceived subjectivity to therapeutic outcomes. Expectation manipulation

trials should investigate whether pre-intervention education about AI's algorithmic nature affects therapeutic benefits, similar to how informed consent influences placebo responses [10].

The AI placebo effect hypothesis provides a testable framework for understanding positive outcomes in human-AI therapeutic relationships, without requiring the attribution of consciousness to artificial systems. This perspective doesn't dismiss the therapeutic potential of AI. Instead, it grounds it in established psychological mechanisms, which we need to understand better to implement increasingly prevalent AI interventions in an ethical manner.

1. Malouin-Lachance A, Capolupo J, Laplante C, Hudon A. Does the Digital Therapeutic Alliance Exist? Integrative Review. *JMIR Mental Health* 2025 Feb 7;12(1):e69294. doi: 10.2196/69294
2. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. *Front Digit Health Frontiers*; 2022 Apr 11;4. doi: 10.3389/fdgth.2022.847991
3. Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, Jewett AD, Salzhauer AJ, Griffin TZ, Jacobson NC. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI Massachusetts Medical Society*; 2025 Mar 27;2(4):AIoa2400802. doi: 10.1056/AIoa2400802
4. Fuchs T. Understanding Sophia? On human interaction with artificial agents. *Phenom Cogn Sci* 2024 Feb 1;23(1):21–42. doi: 10.1007/s11097-022-09848-0
5. Salles A, Evers K, Farisco M. Anthropomorphism in AI. *AJOB Neuroscience* 2020 Apr 2;11(2):88–95. doi: 10.1080/21507740.2020.1740350
6. Waytz A, Morewedge CK, Epley N, Monteleone G, Gao J-H, Cacioppo JT. Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology* 2010;99(3):410–435. doi: 10.1037/a0020240
7. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun ACM* 1966 Jan;9(1):36–45. doi: 10.1145/365153.365168
8. Bender EM, Hanna A. *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. Random House; 2025. ISBN:978-1-5299-4989-6
9. Price D, Finniss D, Benedetti F. A Comprehensive Review of the Placebo Effect: Recent Advances and Current Thought. *Annual review of psychology* 2008 Feb 1;59:565–90. doi: 10.1146/annurev.psych.59.113006.095941
10. Alfano M. *Placebo Effects and Informed Consent*. The American Journal of Bioethics Taylor & Francis; 2015 Oct 3;15(10):3–12. PMID:26479091