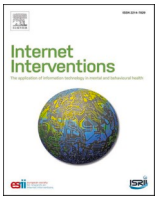
Internet Interventions 41 (2025) 100841

Contents lists available at ScienceDirect

Internet Interventions

journal homepage: www.elsevier.com/locate/invent



Artificial intelligence vs. human expert: Licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship

Ludwig Franke Föyen [a,b,c,d,1], Emma Zapel [a,e,*,1], Mats Lekander [a,b,d], Erik Hedman-Lagerlöf [a,c,d,2], Elin Lindsäter [a,c,d,f,2]

[a] Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
[b] Stress Research Institute, Department of Psychology, Stockholm University, Stockholm, Sweden
[c] Gustavsberg University Primary Care Center, Stockholm Health Care Services, Region Stockholm, Sweden
[d] Osher Center for Integrative Health, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
[e] Division of Clinical Psychology, Department of Psychology, Uppsala University, Uppsala, Sweden
[f] Center for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm Health Care Services, Region Stockholm, Sweden

ARTICLE INFO

Keywords:
Artificial intelligence
Digital health
Empathy
Mental health
Psychological advice
Therapeutic Alliance

ABSTRACT

Background: The use of artificial intelligence for psychological advice shows promise for enhancing accessibility and reducing costs, but it remains unclear whether AI-generated advice can match the quality and empathy of experts.

Method: In a blinded, comparative cross-sectional design, licensed psychologists and psychotherapists assessed the quality, empathy, and authorship of psychological advice, which was either AI-generated or authored by experts.

Results: AI-generated responses were rated significantly more favorable for emotional (OR = 1.79, 95 % CI [1.1, 2.93], $p = .02$) and motivational empathy (OR = 1.84, 95 % CI [1.12, 3.04], $p = .02$). Ratings for scientific quality ($p = .10$) and cognitive empathy ($p = .08$) were comparable to expert advice. Participants could not distinguish between AI- and expert-authored advice ($p = .27$), but *perceived* expert authorship was associated with more favorable ratings across these measures (ORs for perceived AI vs. perceived expert ranging from 0.03 to 0.15, all $p < .001$). For overall preference, AI-authored advice was favored when assessed blindly based on its actual source ($\beta = 6.96$, $p = .002$). Nevertheless, advice *perceived* as expert-authored was also strongly preferred ($\beta = 6.26$, $p = .001$), with 93.55 % of participants preferring the advice they believed came from an expert, irrespective of its true origin.

Conclusions: AI demonstrates potential to match expert performance in asynchronous written psychological advice, but biases favoring perceived expert authorship may hinder its broader acceptance. Mitigating these biases and evaluating AI's trustworthiness and empathy are important next steps for safe and effective integration of AI in clinical practice.

1. Background

The integration of artificial intelligence (AI) for psychological advice has evolved significantly since the creation of ELIZA (Weizenbaum, 1966), a conversational agent designed to simulate Rogerian psychotherapy by mimicking reflective, non-directive responses. ELIZA was a far cry from the sophisticated systems available today, but it already demonstrated the potential of AI to engage users and sparked interest in conversational agents. The so-called "ELIZA effect," where users attributed human-like qualities to the chatbot (Tarnoff, 2023), highlighted the compelling nature of such interactions. Today's large language models (LLMs) and natural language processing (NLP)-based

* Corresponding author at: Department of Clinical Neuroscience, Karolinska Institutet, Nobels Väg 9, Solna, Sweden.
E-mail address: emma.zapel@ki.se (E. Zapel).
[1] Authors share co-first authorship.
[2] Authors share co-senior authorship.

https://doi.org/10.1016/j.invent.2025.100841
Received 16 February 2025; Received in revised form 13 May 2025; Accepted 2 June 2025
Available online 3 June 2025
2214-7829/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

conversational agents were not originally designed to provide psychological advice, but researchers are investigating their potential to do so. With up to 70 % of individuals with mental illness lacking adequate care (Boucher et al., 2021), modern AI tools have been proposed to address resource shortages by providing timely, personalized support (Egan et al., 2024; Moitra et al., 2023). Preliminary evidence on the effectiveness of conversational agents in mental health is mixed. A meta-analysis of 12 studies found weak evidence for improvements in symptoms of depression, stress, and specific phobias after using AI-based treatments, while underscoring significant risk of bias and the lack of high-quality trials (Abd-Alrazaq et al., 2020). Individual interventions, such as "Tess," using natural language processing (NLP) and emotion detection, significantly reduced depressive and anxiety symptoms and outperformed bibliotherapy (Fulmer et al., 2018). Similarly, conversational agents like "Woebot" and "Wysa" have demonstrated potential in alleviating symptoms of depression and anxiety, particularly among highly engaged users (Fitzpatrick et al., 2017; Inkster et al., 2018).

While conversational agents have demonstrated potential to engage users while delivering support (Boucher et al., 2021), significant questions remain about their suitability in acting as substitutes for therapists. Limitations in scientific depth and accuracy of AI-generated responses persist, with some studies finding notable gaps in their factual reliability (Ma et al., 2023). However, another study showed that, when blinded, AI-generated responses to questions about mental health problems were rated more authentic, professional, and practical compared to expert-authored responses (Lopes et al., 2023). It is noteworthy that while linguistic patterns in AI responses often lack diversity, even linguistic experts struggle to consistently identify AI-generated text, achieving only 38.9 % accuracy (Desaire et al., 2023).

Research has shown that trust and acceptance of AI in healthcare contexts are strongly influenced by human-like attributes, perceived empathy and interaction quality (Pelau et al., 2021; Stade et al., 2024). Cognitive empathy, or understanding and acknowledging others feelings, has been suggested to be particularly impactful, as it fosters therapeutic alliance and helps practitioners better understand and respond to patients' emotions (Moudatsou et al., 2020). While there has been some skepticism around empathy in AI, some studies have suggested that AI-generated responses can be perceived as highly empathetic. Ayers et al. (2023), for example, found that users rated AI-generated mental health responses as more empathetic than those written by experts. However, other research indicates that users still prefer expert responses for their perceived authenticity and depth (Morris et al., 2018). It is worth noting that such findings may reflect the limitations of earlier-generation models. Given the rapid pace of development in conversational agents, results from studies conducted just a few years ago—such as Morris et al. (2018)—may no longer be representative of current capabilities. As the technology continues to evolve, specifying the exact model and version under evaluation becomes increasingly important. Emerging research indicates that newer models are better able to capture linguistic nuance and contextual cues, which may contribute to enhanced perceptions of empathy (Li et al., 2024). In sum, while conversational agents have demonstrated potential to engage users and deliver support, significant questions remain about its performance and suitability in acting as a substitute for a therapist. There remain critical concerns about ethical implications of deploying conversational agents in sensitive healthcare contexts, where errors or misunderstandings could harm users (Coghlan et al., 2023). As researchers continue to refine these systems, it will be crucial to examine if AI-generated advice can be of similar quality as expert advice in terms of scientific quality and empathy, and to investigate whether the perceived source of the advice affects its perceived quality.

1.1. Objective of the study

This explorative study investigates the quality and empathy of AI (based on a LLM, GPT-4) generated written psychological advice as compared to expert advice using ratings by licensed mental health clinicians. It also examines their ability to distinguish between AI and expert-authored advice, and how perceived authorship influences ratings of quality and empathy. By focusing on clinicians' evaluations, this study aims to provide a qualified assessment of AI-generated content and identify potential biases that may affect trust in AI for psychological advice.

1.2. Research questions

I. Can AI produce psychological advice that is comparable to advice by an expert in terms of scientific quality and empathy?
II. To what extent are clinicians able to identify whether psychological advice was produced by an AI or an expert?
III. Does the clinicians' perception of authorship relate to their preference and scores of the level of empathy, and scientific quality of the advice?

2. Methods

2.1. Study design

The study employed a cross-sectional design comparing psychological advice generated by a conversational agent to that given by psychologists, psychotherapists and psychiatrists published in an advice column in a Swedish national newspaper. Published reader questions around mental health and advice by experts were the basis for the study. Participants were presented with randomly selected question and response sets, blinded to authorship and subsequently asked to rate them. The study was pre-registered on Open Science Framework (Zapel et al., 2024).

2.2. Participants

Participants were recruited using a convenience sampling method, targeting Swedish-speaking licensed mental health clinicians, specifically licensed psychologists, psychotherapists and physicians (hereafter referred to as clinicians). Recruitment was conducted via multiple channels, including clinic posters, targeted email outreach and social media forums aiming to maximize sample diversity.

A total of 47 participants were recruited, of which four were excluded (two non-completers, two not meeting criteria), resulting in a final sample of 43 licensed mental health clinicians (40 psychologists and three psychotherapists). No sensitive data were collected and participation in the study was anonymous.

2.3. Materials

2.3.1. Newspaper articles

The study used 26 mental health advice columns from 'Dagens Nyheter,' a major Swedish newspaper (https://www.dn.se/) published between January 2020 and January 2024. These columns consisted of reader questions on issues like relationship problems and mental health problems and similar topics and the corresponding responses from psychologists, psychotherapists or psychiatrists employed by the newspaper to provide written advice. Columns were chosen by (EZ) to represent a variety of topics and only from the most recent newspaper issues to limit the possibility of AI familiarity with the articles. Of the chosen advice columns, 77 % ($n = 20$) were used to train a chatbot, while 23 % ($n = 6$) were included in a survey. These articles were fairly comprehensive with a median length of the reader questions of 340 words, and the expert responses had a median length of 834 words and were written by four different authors with a coincidental predominance of one specific author (53 % of all, 67 % of the test set).

2.3.2. The conversational agent

The conversational agent was developed using OpenAI's model GPT-4 (Fråga Insidan, 2024). It was enhanced using retrieval augmented generation and a training set of 20 of the mental health advice columns. The conversational agent had at the time of article generation been trained on data up until early 2023 and could not have been exposed to five out of six articles in the test set. The conversational agent was instructed to mimic the style and tone of a professional psychologist writing for a Swedish advice column. For the specific prompt instructions, see Zapel et al. (2024). For the purpose of this study, the conversational agent was not directly used by participants but the answers were extracted and slightly adjusted with minor edits by the research team to remove specific signs of AI writing (like bullet points), to allow for blinded comparison with expert answers.

2.3.3. The rating questionnaire

The questionnaire was created on 'Google Forms'. It included the six reader questions and expert answers from the newspaper combined with the alternative answer generated by the conversational agent. Out of six sets, each participant was presented with two randomly selected sets. Participants rated the texts for empathy and scientific quality (scale of 1–5), perceived authorship (AI or Expert), and preference (binary). After having completed two sets, participants were offered to rate the additional four sets. In total, the questionnaire contained seven sections and 73 items; each participant answered a minimum of 26 items, with three questions assessing inclusion criteria and 11 questions per article set. The questions were constructed by the researchers and included a description for each construct investigated (for the complete questionnaire see: Zapel et al., 2024).

2.3.4. Outcome measurements and variables

Subjective ratings from participants, all licensed psychologists and or psychotherapists, served as outcome measures. No personal demographics were collected beyond professional background and Swedish language proficiency.

The study's independent variable was authorship (AI-generated vs. expert-written), while dependent variables included scientific quality and empathy (ordinal scales; 1 = very poor, 5 = very high) as well as perceived authorship and preference (binary).

Scientific quality was rated using a single item assessing the overall scientific soundness of the advice.

Empathy was assessed using 3 items based on the components suggested by Montemayor et al. (2022): emotional empathy, cognitive empathy, and motivational empathy. Emotional empathy was defined as "the ability to share and understand the feelings of others as if they were one's own. Cognitive empathy was defined as "the capacity to understand and acknowledge the perspectives and feelings of others without necessarily sharing them." Motivational empathy was defined as "the drive to respond appropriately to someone's emotional state or needs, often leading to supportive or helping behaviors." This 3-item empathy scale demonstrated good internal consistency in the current study (based on $N = 208$), with a Cronbach's alpha of 0.89 and an average inter-item correlation of 0.74.

2.4. Statistical analysis

The statistical analysis plan for this study was pre-registered (Zapel et al., 2024), and an initial power analysis determined that a sample of 64 respondents was required to detect a medium effect size ($d = 0.5$) with 80 % power at the 0.05 alpha level for ordinal ratings, and over 95 respondents for binary outcomes. However, data collection was terminated earlier than planned due to monetary and time constraints. As a result, post-hoc adjustments were made to the analysis plan to improve robustness, including accounting for the nested structure of the data.

The primary analysis focused on the ordinal ratings of scientific quality and empathy, which were analyzed using cumulative link mixed models. Binary outcomes, such as preference (AI vs. expert), were modeled using generalized linear mixed models with a logit link function. Both types of models included random intercepts for raters and articles to account for variability at the individual and article level. The main fixed effects were actual authorship (AI vs. expert) and perceived authorship (participant's guess of authorship), with an interaction term included to investigate how these two factors influenced preferences and ratings. All statistical tests were two-sided and the alpha level was set at 0.05.

The analysis plan was adapted post-registration to include random effects for both raters and articles, acknowledging the hierarchical nature of the data where each rater assessed multiple articles and each article received multiple evaluations. Likelihood ratio tests were employed for model comparison and to assess the significance of interaction effects and the inclusion of random effects. Simplified models were also tested to isolate the main effects of perceived authorship and actual authorship on preferences.

Data extraction and cleaning were performed using Python (Van Rossum and Drake, 2010), utilizing libraries such as NumPy (Harris et al., 2020) and SciPy (Virtanen et al., 2020), with statistical analyses carried out in R (RStudio Team, 2024). The statistical code, along with the rating questionnaire and AI chatbot transcripts, is available on the Open Science Framework (OSF) for reproducibility and transparency (Zapel et al., 2024).

3. Results

Forty-three licensed mental health clinicians rated a total of 208 responses and 104 pairs. On average, each participant rated 2.4 response pairs (range: 1–6). As shown in Table 1, AI-generated advice received equal or more favorable ratings across all measures. For scientific quality ($p = .10$) and cognitive empathy ($p = .08$), the differences were not statistically significant, but AI responses were rated significantly more favorable for emotional empathy ($\beta = 0.59$, $p = .02$) and motivational empathy ($\beta = 0.61$, $p = .02$). Given that we did not reach the intended sample size, we conducted a post-hoc minimal detectable effect size analysis for the non-significant results. The analysis indicated that the study could detect effects of moderate size (OR $\geq$ 1.6–1.7), but lacked sensitivity to smaller effects which may explain the absence of statistically significant findings (Code available at Zapel et al. (2024).

Clinicians were unable to reliably distinguish between AI- and expert-authored responses, performing at chance level (45 % accuracy; χ^2 test, $p = .27$). Answers *perceived* to be authored by an expert were rated more favorable across all measures of scientific quality and empathy, as shown in Table 2. Specifically, expert responses were rated more favorable in scientific quality ($\beta = -1.89$, $p < .001$), emotional empathy ($\beta = -3.62, p < .001$), cognitive empathy ($\beta = -3.02, p < .001$) and motivational empathy ($\beta = -2.74$, $p < .001$).

The analysis of participants' preferences for AI- or expert advice revealed significant main effects for both actual authorship ($\beta = 6.96$, $p = .002$) and perceived authorship ($\beta = 6.26$, $p = .001$), indicating that AI-authored responses were preferred overall and that perceived expert authorship strongly influenced preferences. When participants perceived an answer as being authored by an expert, they overwhelmingly preferred that response, regardless of whether the answer was actually authored by AI or an expert (93.55 % preference for perceived expert advice). The interaction term between perceived and actual authorship was also significant ($\beta = -12.29$, $p = .001$), as visualized in Fig. 1 below.

4. Discussion

This cross-sectional study compared psychological advice provided by a conversational agent, developed using a LLM, GPT-4, to that given by experts in the context of a Swedish national newspaper advice column. The results of this study indicate that AI can produce psychological

Table 1
Ratings of AI and expert psychological advice to reader questions in a newspaper advice column.

		M (SD)				
Measure	Range	AI	Expert	*z-value*	Effect size (*OR* with 95 % *CI*)	*p*-value
Scientific Quality	(1–5)	3.46 (0.93)	3.21 (1.12)	1.65	1.53 (0.92, 2.53)	0.10
Emotional Empathy	(1–5)	3.45 (1.16)	3.05 (1.28)	2.33	1.79 (1.1, 2.93)	**0.02**
Cognitive Empathy	(1–5)	3.67 (1.01)	3.39 (1.14)	1.73	1.55 (0.94, 2.56)	0.08
Motivational Empathy	(1–5)	3.66 (1.06)	3.27 (1.21)	2.40	1.84 (1.12, 3.04)	**0.02**

Note. Raters were blinded to the true authorship of the advice.

Table 2
Ratings of perceived AI and perceived expert psychological advice to reader questions in a newspaper advice column.

		M (SD)				
Measure	Range	Perceived AI	Perceived Expert	*z-value*	Effect size (*OR* with 95 % *CI*)	*p*-value
Scientific Quality	(1–5)	2.93 (0.99)	3.75 (0.91)	−6.28	0.15 (0.09, 0.24)	**< 0.001**
Emotional Empathy	(1–5)	2.42 (0.98)	4.12 (0.79)	−9.31	0.03 (0.01, 0.06)	**< 0.001**
Cognitive Empathy	(1–5)	2.91 (1.03)	4.19 (0.67)	−8.37	0.05 (0.03, 0.08)	**< 0.001**
Motivational Empathy	(1–5)	2.83 (1.14)	4.13 (0.70)	−8.06	0.06 (0.03, 0.11)	**< 0.001**

Note. Raters were blinded to the true authorship of the advice.

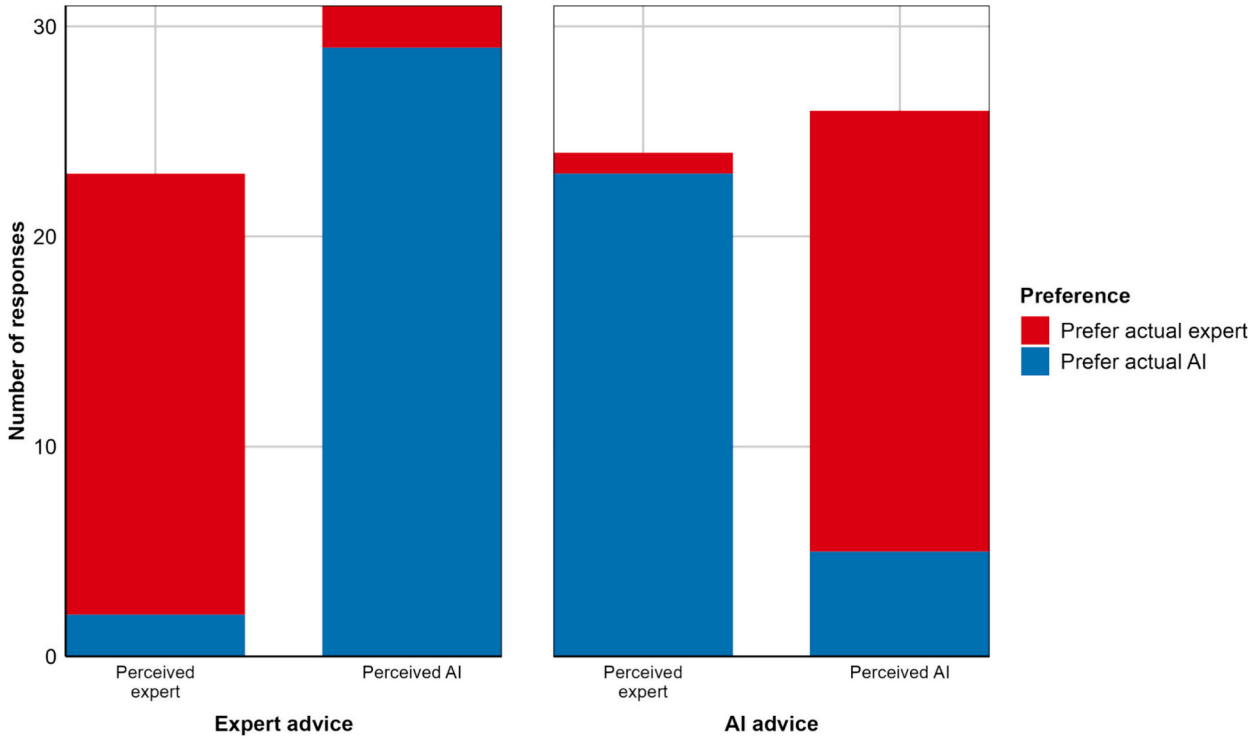


Fig. 1. Preferences for AI and expert-authored advice by actual and perceived authorship
Note. Preferences are displayed as the number of responses indicating a preference for AI or expert advice. Bars are grouped by actual authorship (Expert advice or AI advice) and further divided by perceived authorship (Perceived expert or Perceived AI).

advice in a specific, structured, textual format comparable or even superior to psychological advice provided by experts with a high degree of scientific quality and empathy. Although the clinically trained raters were not able to distinguish between AI and expert advice they did display a preference for responses they perceived to be expert-authored when the original source was blinded. They also gave more favorable ratings of scientific quality and empathy when they thought the content was expert-authored. These results suggest a bias toward expert advice among clinically trained professionals, which could influence the acceptance of AI tools, despite AI's demonstrated ability to provide empathetic and high-quality support.

4.1. Quality

While findings for scientific quality in AI-writing have been mixed so far (Lopes et al., 2023; Ma et al., 2023) the present results suggest that AI may be suitable to provide psychological advice as the conversational agent tested in this study produced comparable or better results to experts. Notably, the training and outcome languages were Swedish, highlighting the potential for conversational agents to perform across diverse languages. These promising findings in a relatively small language suggest similar or better outcomes in languages with greater availability of training data, warranting further investigation.

4.2. Empathy

Empathy, as a critical component of effective psychological support, has traditionally been viewed as a uniquely human trait. Our results suggest that AI can offer written advice perceived to be highly empathetic, particularly displaying high levels of emotional and motivational empathy. This is surprising considering recent studies suggested that AI can only effectively mimic certain aspects of empathetic communication (Ayers et al., 2023). In this study, the prompts were relatively informative and slightly longer than average chatbot messages, which may have influenced the perceived empathy in AI responses as compared to other conversational agent formats.

4.3. Authorship - accuracy and influence of perceived authorship ratings

The accuracy of clinicians' guesses regarding the true source of the responses was approximately at chance level. This aligns with previous research, showing that professionals are generally unable to distinguish between support provided by experts and that generated by AI (Desaire et al., 2023). This inability, combined with the comparable scientific quality and empathy ratings between expert and AI-generated content, suggests a notable resemblance between the two sources of advice. However, when considering perceived authorship, the results show a clear bias: participants consistently rated content they believed was expert-authored more favorably. This supports previous findings (e.g., Jain et al., 2024), which suggest that perceived human expertise strongly impacts the selection and evaluation of psychological support, with clinicians tending to assume that experts offer superior advice. In particular, trust in the source appears to play a key role in how advice is received and acted upon (Shevtsova et al., 2024). While undisclosed AI advice is often seen as competent (Alowais et al., 2023) and can foster trust (Ayers et al., 2023), studies show that disclosing AI origins reduces perceived credibility and empathy, possibly leading to a nocebo effect in future AI interventions (Reis et al., 2024). However, other findings suggest that when patients knowingly interact with conversational agents, they may actually become more open and self-disclose more information, possibly due to reduced fear of judgment (Miner et al., 2017). User attitudes toward AI are shaped by factors like preexisting beliefs about AI competence and individual traits such as neuroticism (Vodrahalli et al., 2022; Bergdahl et al., 2023) but the strong effect of perceived authorship may also reflect well-established cognitive biases, such as authority bias—where humans tend to assign greater credibility to perceived experts—and algorithm aversion, a tendency to distrust machine-generated output after witnessing errors (Dietvorst et al., 2015). In this context, human-generated responses may be overvalued due to assumptions about empathy, intentionality, or experiential knowledge, while AI responses may trigger skepticism based in broader societal narratives about automation and technological control.

In sum, attitudes toward AI in mental healthcare warrant further investigation, especially as familiarity with AI grows, potentially influencing its acceptance and integration.

4.4. Implications for clinical practice

While the current results provide some evidence for AI's ability to provide written advice with high scientific quality and empathy, the absence of a genuine human relationship and the nuances of human interaction may limit the effectiveness of AI in a therapeutic context. A recent survey of community members and mental health professionals reflects the ambivalence surrounding AI's role in mental healthcare (Cross et al., 2024). While acknowledging its potential to improve accessibility, personalization, and efficiency, they voice concerns about compromised trust, risks to accuracy and ethics, and the loss of human connection central to the therapeutic relationship. These concerns are echoed by Carlbring et al. (2023), who emphasize that while AI may successfully simulate empathy, the establishment of therapeutic alliance and trust remains a significant challenge. Nevertheless, recent findings from Habicht et al. (2025) indicate that integrating a generative AI tool into group-CBT settings improved both patient engagement and clinical outcomes, suggesting the promise of conversational agents in such contexts. Similarly, a recent randomized controlled trial by Heinz et al. (2025) using a generative AI therapy chatbot showed significant reductions in symptoms of depression, anxiety, and early psychosis, with participants rating the therapeutic alliance as comparable to that with human therapists.

Concerns regarding AI in mental healthcare are reminiscent of earlier skepticism about internet-based therapies and bibliotherapy; these approaches, initially questioned due to the absence of face-to-face interaction, have since demonstrated their effectiveness in various contexts (Fulmer et al., 2018). This historical perspective, combined with findings from the present study that perceived authorship significantly influences user evaluations, points toward specific directions for future research. One avenue involves exploring AI as a ‘co-pilot’ or supportive tool for therapists, a model pertinent to the widespread use of asynchronous therapist-patient communication in internet-based psychological treatments. In such applications, therapists would retain primary responsibility for patient communication, with AI providing assistance. Further investigation is also warranted to examine outcomes and engagement related to AI-assisted psychological support within other blended care frameworks or low-intensity interventions.

4.5. Strengths and limitations

This study's strengths are, first, the empirical insights into differences in empathy, scientific quality, and preference between AI-generated and expert advice, providing a multidimensional evaluation of AI's capacity to replicate expert-like emotional intelligence. Second, blinding participants to authorship helped to mitigate bias and allowed for testing of perceived versus actual authorship effects. Third, using clinicians as raters adds depth to the understanding of AI's ability to replicate expert-like support compared to consumer-based evaluations. Additionally, the study was conducted in accordance with ethical standards and open science principles, with materials available on OSF.io to support transparency and replicability.

The study has limitations that may affect its validity and generalizability. The predominance of texts written by a single expert author (67 %) may not fully represent how expert advice is evaluated in terms of empathy and scientific quality, underscoring the need for diverse authorship in future research. The professional credentials of raters were self-reported and not verified, which could have led to unqualified individuals influencing the results. Further, the use of convenience sampling and the limited information about the sample characteristics limits the generalizability, as this sampling method may not provide an accurate representation of the target population.

5. Conclusion

The findings suggest that AI-generated psychological advice is comparable to expert asynchronous written psychological advice in scientific quality and cognitive empathy while surpassing expert advice in emotional and motivational empathy. However, preferences were strongly influenced by perceived authorship, with advice believed to be expert-authored consistently rated more favorably, regardless of its actual source. These results indicate that biases related to perceived authorship may present challenges for the acceptance of AI in mental healthcare. Further research is needed to examine the role of these biases across diverse contexts and explore strategies for their mitigation. Additionally, future studies should investigate the potential of AI tools to enhance clinical decision-making and improve treatment processes in mental healthcare.

Declaration of ethical data handling

This study did not include an intervention and no personal data was collected, which is why no ethical approval was needed. The study adhered to ethical guidelines by ensuring the privacy rights of all participants were respected. Informed consent was obtained from all participants through a dedicated section in the questionnaire.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used OpenAI's ChatGPT-4 in order to create the specialized GPT that produced the advice for the comparison. ChatGPT was also used to make small adjustments in the article like shortening to fit the word count and compiling statistical analysis code. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of Funding Sources

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data supporting the findings of this study, along with the statistical analysis code, have been deposited in the Open Science Framework (OSF) repository and are openly available at (Zapel et al., 2024).

References

Abd-Alrazaq, A.A., Rababeh, A., Alajlani, M., Bewick, B.M., Househ, M., 2020. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J. Med. Internet Res. 22 (7). https://doi.org/10.2196/16021.

Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A., Al Yami, M.S., Al Harbi, S., Albekairy, A.M., 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med. Educ. 23 (1). https://doi.org/10.1186/s12909-023-04698-z.

Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., Faix, D.J., Goodman, A.M., Longhurst, C.A., Hogarth, M., Smith, D.M., 2023. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. JAMA Intern. Med. 183 (6), 589. https://doi.org/10.1001/jamainternmed.2023.1838.

Bergdahl, J., Latikka, R., Celuch, M., Savolainen, I., Soares Mantere, E., Savela, N., Oksanen, A., 2023. Self-determination and attitudes toward artificial intelligence: Cross-national and longitudinal perspectives. Comput. Human Behav. 147, 107816. https://doi.org/10.1016/j.chb.2023.107816.

Boucher, E.M., Harake, N.R., Ward, H.E., Stoeckl, S.E., Vargas, J., Minkel, J., Parks, A.C., Zilca, R., 2021. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev. Med. Devices 18 (sup1), 37–49. https://doi.org/10.1080/17434440.2021.2013200.

Carlbring, P., Andersson, G., Rozental, A., Ström, L., 2023. The promise and pitfalls of using AI in psychotherapy: a review. Internet Interv. 32, 100621. https://doi.org/10.1016/j.invent.2023.100621.

Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., D'Alfonso, S., 2023. To chat or bot to chat: ethical issues with using chatbots in mental health. Digital Health 9. https://doi.org/10.1177/20552076231183542, 20552076231183542.

Cross, S., Bell, I., Nicholas, J., Valentine, L., Mangelsdorf, S., Baker, S., Titov, N., Alvarez-Jimenez, M., 2024. Use of AI in mental health care: community and mental health professionals survey. Jmir Ment. Health 11, e60589. https://doi.org/10.2196/60589.

Desaire, H., Chua, A.E., Isom, M., Jarosova, R., Hua, D., 2023. Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. Cell Rep. Phys. Sci. 4 (6). https://doi.org/10.1016/j.xcrp.2023.101426.

Dietvorst, B.J., Simmons, J.P., Massey, C., 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. 144 (1), 114–126. https://doi.org/10.1037/xge0000033.

Egan, S.J., Johnson, C., Wade, T.D., Carlbring, P., Raghav, S., Shafran, R., 2024. A pilot study of the perceptions and acceptability of guidance using artificial intelligence in internet cognitive behaviour therapy for perfectionism in young people. Internet Interv. 35, 100711. https://doi.org/10.1016/j.invent.2024.100711.

Fitzpatrick, K.K., Darcy, A., Vierhile, M., 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment. Health 4 (2). https://doi.org/10.2196/mental.7785.

Fråga Insidan, 2024. ChatGPT. https://chatgpt.com/g/g-jCFQU8W4O-fraga-insidan.

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., Rauws, M., 2018. Using psychological artificial intelligence (TESS) to relieve symptoms of depression and anxiety: randomized controlled trial. Jmir Ment. Health 5 (4). https://doi.org/10.2196/mental.9782.

Habicht, J., Dina, L.M., McFadyen, J., Stylianou, M., Harper, R., Hauser, T.U., Rollwage, M., 2025. Generative AI–enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: real-world observational study. J. Med. Internet Res. 27, e60435. https://doi.org/10.2196/60435.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., et al., 2020. Array programming with NumPy. Nature 585 (7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Heinz, M.V., Mackin, D.M., Trudeau, B.M., Bhattacharya, S., Wang, Y., Banta, H.A., Jewett, A.D., Salzhauer, A.J., Griffin, T.Z., Jacobson, N.C., 2025. Evaluating therabot: a randomized control trial investigating the feasibility and effectiveness of a generative AI therapy chatbot for depression, anxiety, and eating disorder symptom treatment. NEJM AI. Advance online publication. https://doi.org/10.1056/AIoa2400802.

Inkster, B., Sarda, S., Subramanian, V., 2018. An empathy-driven, conversational artificial intelligence agent (WYSA) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 6 (11). https://doi.org/10.2196/12106.

Jain, G., Pareek, S., Carlbring, P., 2024. Revealing the source: how awareness alters perceptions of AI and human-generated mental health responses. Internet Interv. 36, 100745. https://doi.org/10.1016/j.invent.2024.100745.

Li, A., Lu, Y., Song, N., Zhang, S., Ma, L., & Lan, Z. (2024). Automatic evaluation for mental health counseling using LLMs. arXiv preprint arXiv:2402.11958.

Lopes, E., Jain, G., Carlbring, P., Pareek, S., 2023. Talking mental health: a battle of wits between humans and AI. J. Technol. Behav. Sci. https://doi.org/10.1007/s41347-023-00359-6.

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., Liu, X., 2023. AI vs. human - differentiation analysis of scientific content generation. arXiv. https://doi.org/10.48550/arXiv.2301.10416.

Miner, A.S., Milstein, A., Hancock, J.T., 2017. Talking to machines about personal mental health problems. JAMA 318 (13), 1217. https://doi.org/10.1001/jama.2017.14151.

Moitra, M., Owens, S., Hailemariam, M., Wilson, K.S., Mensa-Kwao, A., Gonese, G., Kamamia, C.K., White, B., Young, D.M., Collins, P.Y., 2023. Global Mental Health: Where We Are and Where We Are Going. Curr. Psychiatry Rep. 25 (7), 301–311. https://doi.org/10.1007/s11920-023-01426-8.

Montemayor, C., Halpern, J., Fairweather, A., 2022. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. AI & Soc. 37 (4), 1353–1359. https://doi.org/10.1007/s00146-021-01230-z.

Morris, R.R., Kouddous, K., Kshirsagar, R., Schueller, S.M., 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. J. Med. Internet Res. 20 (6). https://doi.org/10.2196/10148.

Moudatsou, M., Stavropoulou, A., Philalithis, A., Koukouli, S., 2020. The Role of Empathy in Health and Social Care Professionals. Healthcare (Basel, Switzerland) 8 (1), 26. https://doi.org/10.3390/healthcare8010026.

Pelau, C., Dabija, D.-C., Ene, I., 2021. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput. Human Behav. 122, 106855. https://doi.org/10.1016/j.chb.2021.106855.

Reis, M., Reis, F., Kunde, W., 2024. Influence of believed AI involvement on the perception of digital medical advice. Nat. Med. 30 (11), 3098–3100. https://doi.org/10.1038/s41591-024-03180-7.

RStudio Team, 2024. *RStudio desktop (version 2024.04.0+735)* [computer software]. PBC. https://posit.co/products/open-source/rstudio/.

Shevtsova, D., Ahmed, A., Boot, I.W., Sanges, C., Hudecek, M., Jacobs, J.J., Hort, S., Vrijhoef, H.J., 2024. Trust in and acceptance of artificial intelligence applications in medicine: mixed methods study. Jmir Hum. Factors. 11. https://doi.org/10.2196/47031.

Stade, E.C., Stirman, S.W., Ungar, L.H., Boland, C.L., Schwartz, H.A., Yaden, D.B., Sedoc, J., DeRubeis, R., Willer, R., Eichstaedt, J.C., 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. NPJ Ment. Health Res. 12 (3). https://doi.org/10.31234/osf.io/cuzvr.

Tarnoff, B., 2023, July 25. Weizenbaum's nightmares: How the inventor of the first chatbot turned against AI. The Guardian. https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai.

Van Rossum, G., Drake, F.L., 2010. The Python language reference: release 3.0.1. Python Software Foundation, Hampton, NH.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., et al., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17 (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Vodrahalli, K., Daneshjou, R., Gerstenberg, T., Zou, J., 2022. Do humans trust advice more if it comes from AI? An analysis of human-AI interactions. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 763–777. https://doi.org/10.1145/3514094.3534150.

Weizenbaum, J., 1966. Eliza—a computer program for the study of natural language communication between man and machine. Commun. ACM 9 (1), 36–45.

Zapel, E., Lekander, M., Föyen, L.F., Lindsäter, E., 2024. Artificial intelligence vs. human expert: blinded comparison of AI-generated and expert psychological advice on quality, empathy, preference and perceived authorship. https://doi.org/10.17605/OSF.IO/WNY62.