Article

Online ISSN: 3070-5320 Print ISSN: 3070-5312
American Journal of Medicine and Health Studies
Volume 5 (2026), pp. 1-28
Star Scholars Press
https://ojed.org/med
DOI: https://doi.org/10.32674/gaez3y26

Artificial Intelligence and Self-Medication Behaviors in Psychological Impairment

Montassar Dridi
The Medical faculty of the University of Debrecen, Hungary

ABSTRACT

This cross-sectional study aimed to explore the psychological distress, therapeutic misconception, negative effects, barriers to treatment, and underservedness as predictors of use of AI, self-medicating, and AI preference among 106 respondents filling in a survey with the following variables: K10 psychological distress, barriers to treatment, use of AI frequency, therapeutic misconception, self-medicating, negative effects, AI preference, seeking professional help, and demographic factors. Psychological distress had been shown to be a significant predictor of frequency of use of AI ($r = +0.340$, $p < 0.001$), but no mediating effect of barriers on the latter variable had been found. Negative effects were significantly correlated with clinically integrated AI preference ($r = +0.280$, $p = 0.0037$). Cost barriers predicted self-medication behavior ($r = 0.245$, $p = 0.012$) and showed stronger associations than perceived quality of AI (mean $r = 0.183$ versus 0.110). Underserviced status was a significant predictor of preference for bridge-to-care AI ($r = +0.314$, $p = 0.001$), while living in rural areas was also a significant predictor ($r = +0.206$, $p = 0.0345$). The findings imply that AI chatbots are an accessible tool for mental health support, especially for distressed and underserved individuals. Nevertheless, the frequency of negative outcomes (84.0%) and self-medication behavior shows the importance of implementing protective measures into AI technologies.

Keywords: artificial intelligence, chatbots, mental health, psychological impairment, self-medication, therapeutic misconception

Corresponding Author: Montassar Dridi, The Medical Faculty of the University of Debrecen, Debrecen, Hungary. Email: dridimontassar@mailbox.unideb.hu , https://orcid.org/0009-0004-0684-3110

There is currently a growing trend of the application of artificial intelligence in mental healthcare in two different ways: clinical tools and consumer-oriented tools. The current literature review considers the second way as well since consumer-oriented AI is increasingly used to analyze symptoms, find mental health information, and self-diagnose and self-medicate (D'Alfonso, 2020; Balcombe & De Leo, 2022; Wimbarti et al., 2024). Such use is important due to social persuading potential of such AI systems when users are upset and require prompt assistance (De Freitas et al., 2023; Khawaja & Bélisle-Pipon, 2023).

About one billion people all around the world suffer from mental disorders while the number of those who receive minimally adequate treatment is critically low (World Health Organization, 2022). In 2021, about 9.1% of patients with major depressive disorder received such treatment while in some regions of sub-Saharan Africa its rate was only 2.0% (Thornicroft et al., 2021). In light of all this, the scope of use of general-purpose generative AI technology has been fast expanding into informal mental health interventions. According to OpenAI (2025), approximately 1.2 million ChatGPT users per week have signs of suicide planning or intention, and an additional 560,000 users show signs of language linked to psychosis or mania.

Since the launch of ChatGPT to the public in November 2022, LLMs have gained more popularity in providing mental health support to people who lack access to, knowledge of, or belief in healthcare institutions. The recent statistics show that about 52% of adults in the United States use AI chatbots, where many of them interact with these programs in search of emotional support, psychological guidance, and answers related to health issues (Pew Research Center, 2025). The current literature review claims that this trend could be regarded as a mental health access issue as well as a technological one, as the usage of AI is gaining its popularity with delayed, stigmatized, and inaccessible care (Koutsouleris et al., 2022; Habicht et al., 2024).

Hypotheses

The following hypotheses were proposed:

H_1: Individuals with greater psychological impairment severity (neurodivergent group) will report higher frequency of AI chatbot use for mental health advice, mediated by perceived barriers to professional care.

H_2: Therapeutic misconception (overestimation of AI therapeutic competence) will be positively associated with self-medication behaviors based on AI advice and negatively associated with professional help-seeking.

H_3: Users who have experienced adverse outcomes from AI advice will express greater preference for open-source, clinically-integrated AI tools compared to those without adverse experiences.

H_4: Perceived stigma and cost barriers will predict AI-mediated self-medication behaviors more strongly than perceived quality of AI advice.

H_5: Users from underserved populations (rural, low-income, minority groups) will demonstrate higher demand for AI tools designed as bridges to professional care, compared to autonomous AI therapy.

LITERATURE REVIEW

SELF-MEDICATION IN THE DIGITAL AGE

Self-medication, which is seen as the use of substances, behaviors, or interventions for the treatment of self-diagnosed symptoms in the absence of professional monitoring, is a known and recognized concept in the field of mental disorders (Khantzian, 1985, 1997). Conventional self-medication practices have been known to involve substances (alcohol, over-the-counter drugs, prescribed medication without seeking professional advice) or behaviors (avoidance of social contact, over-exercising, addiction to surfing the internet). The development of AI chatbots marks a new form of self-medication, whereby people undertake digital self-medication by seeking validation, diagnosis, and advice from non-clinical algorithms that have no knowledge of individual psychopathologies (Dohnány et al., 2025; Wimbarti et al., 2024).

Self-diagnosis of psychiatric conditions online among laypersons is also rising rapidly (Vayena & Gasser, 2016; De Freitas et al., 2023). This is due to factors such as advances in Internet technology, e-commerce in the health care sector, a greater focus on patient empowerment, and the growing popularity of artificial intelligence (Fox & Duggan, 2013; Wimbarti et al., 2024). People with a positive screening test result may also get medical help (Habicht et al., 2024). But there are many dangers for patients who make self-diagnoses such as wrong or dangerous diagnosis, higher levels of anxiety from receiving a diagnosis, getting uncensored advice from the Internet, self-medicating through self-diagnosis, this includes buying drugs online without a prescription, and technical problems, including privacy loss (Wimbarti et al., 2024; Khawaja & Bélisle-Pipon, 2023; Stade et al., 2024).

Mental health problems may increase AI dependency due to the associated sense of uncertainty, urgency, shame, and barriers to receiving care (Khawaja & Bélisle-Pipon, 2023; D'Alfonso, 2020). Chatbots are especially appealing to individuals unable to receive prompt assistance by humans due to the round-the-clock availability and mobile capability of AI (Khawaja & Bélisle-Pipon, 2023; D'Alfonso, 2020). On the other hand, a lack of mental health professionals and delays in delivering services in the world creates a certain vacuum in practice that might be filled by explanations and coping strategies provided by AI (Koutsouleris et al., 2022; van Heerden et al., 2023). What we get is not just people looking for information; instead, it is self-management of symptoms, diagnosing conditions, and changing behaviors in the absence of clinicians (Casu et al., 2024; Wimbarti et al., 2024; De Freitas et al., 2023).

PATTERNS OF AI USE FOR SYMPTOM INTERPRETATION AND SELF-DIAGNOSIS

Conversational AI as a First-Line Source of Mental Health Advice

Conversational AI has been found more commonly in literature as the informal point of first contact for mental health issues. Through review and perspective papers, it can be seen that users resort to chatbot applications for help, guidance, and therapy advice because of the immediacy, confidentiality, and seamless nature of such technology as compared to traditional methods (Balcombe & De Leo, 2022; Guo et al., 2024; Khawaja & Bélisle-Pipon, 2023). It has also been shown that the fact that large language models are designed for conversational purposes and not for clinical triage but seem proficient enough to be trusted (Guo et al., 2024; De Freitas et al., 2023).

However, in reality, that ease of conversation may make it difficult to distinguish information-seeking from therapy because people usually seek information on their depressive, anxiety, obsessive, and stress symptoms (Kim et al., 2024; Manole et al., 2024; Boucher et al., 2021). The common trend is that users tend not to use the technology as a tool for replacing professionals but rather as a fast layer of interpretation that will help them understand if the further action is necessary (Wimbarti et al., 2024; De Freitas et al., 2023).

Research and literature on digital mental health programs have found that chatbots can provide help for coping, psychoeducation, and guided activities, making users more inclined to rely on the service repeatedly despite the non-specialized nature of the platform (Casu et al., 2024; Manole et al., 2024; Boucher et al., 2021). This becomes a crucial behavioral loop since the user receives a believable rationale, a coping strategy, and a sense of control prior to any assessment by professionals (Wimbarti et al., 2024; Khawaja & Bélisle-Pipon, 2023; De Freitas et al., 2023; Stade et al., 2024).

Self-Diagnosis, Symptom Labeling, and Informal Treatment Decisions

Research on AI-based self-diagnosis indicates that labeling of symptoms has become one of its key applications. This review in particular highlights how AI-based solutions have enabled individuals to conveniently evaluate their mental wellness from their own homes, cautioning against potential consequences and negative implications of such activities on ethical, psychological, and social grounds (Wimbarti et al., 2024). In a similar vein, other reviews indicate that AI can be used for detection, classification, and prediction of risks associated with mental disorders, making people rely on model results as conclusive proof of anything (Cruz-Gonzalez et al., 2025; Jin et al., 2025).

This matter is becoming increasingly significant as there are instances where certain LLMs have demonstrated excellent ability to identify various psychological disorders through the use of vignettes including surpassing clinical professionals in an OCD-related study, this outcome could increase user's confidence even though vignette correctness is

not the same as safe real-world diagnosis (Kim et al., 2024). Online self-diagnosis of psychiatric conditions is increasing tremendously fast, with some individuals having passed the test and looking for professional assistance, while other individuals choose self-medication based on the result they have gotten from the test, which includes purchasing medication online without a prescription (Vayena & Gasser, 2016; Wimbarti et al., 2024). The physician has to understand that laymen are not aware of the complex issues surrounding the diagnosis of such disorders and should be prepared to treat individuals who have already done the online diagnosis (Wimbarti et al., 2024; Khawaja & Bélisle-Pipon, 2023).

In this regard, self-care decision making is the natural result of self-labeling as well. If the user is advised by an AI chatbot about anxiety, OCD, depression, or even burnout, he may try to implement coping, supplement use, detox from social media, change life style, and even avoid certain stimuli without getting evaluated professionally. In all reviewed studies, it is emphasized that AI can affect the following decisions about mental health (Wimbarti et al., 2024; Khawaja & Bélisle-Pipon, 2023; Stade et al., 2024). Therefore, it can be concluded that the described issue should be viewed as self-medication behaviors because, in addition to self-labeling, users are changing their strategy of mental health care according to the machine recommendation (Balcombe & De Leo, 2022; D'Alfonso, 2020; Khawaja & Bélisle-Pipon, 2023; Rahsepar Meadi et al., 2025).

Therapeutic Misconception and Overreliance on Non-Specialized Systems

One such mechanism is therapeutic misconception, which refers to the user's predisposition to overestimate the therapeutic nature of the chatbot and underappreciate its limitations (Khawaja & Bélisle-Pipon, 2023; Appelbaum et al., 1987). The conceptual framework suggests that incorrect marketing strategies, digital alliance perception, and empathic appearance of conversational agents can mislead users in attributing human-like caring capabilities to technology that does not have any clinical responsibilities (Khawaja & Bélisle-Pipon, 2023). It is a crucial concept as it modifies the reception of advice: a piece of information provided by the chatbot is likely to be perceived as a recommendation instead of an unverified data point (De Freitas et al., 2023; Balcombe & De Leo, 2022; Rahsepar Meadi et al., 2025; Koutsouleris et al., 2022).

The propensity for overreliance is supported by the structural attractiveness of AI. The elements of accessibility, privacy, and cost effectiveness can make AI more appealing than human interaction in case of a stigmatizing experience (Khawaja & Bélisle-Pipon, 2023; Habicht et al., 2024; Rahsepar Meadi et al., 2025). However, at the same time, the elements in question hide the lack of accountability, empathy, and crisis responsibility (Rahsepar Meadi et al., 2025; Stade et al., 2024). Based on this, the core behavioral problem can be seen as the problem of substitution in which users will see non-specialized AI as a substitute way to engage with mental health care (Wimbarti et al., 2024; van Heerden et al., 2023).

The Engagement-Validation Loop

Two processes that feed into each other, such as algorithmic fawning and anthropomorphizing, combine to form validation loops that may be self-sustaining and may reinforce maladaptive thoughts (Dohnány et al., 2025; Perez et al., 2023). As human evaluators tend to validate outcomes, reinforcement learning from human feedback (RLHF) training will make it possible to condition LLMs to provide default-conformant, rather than accurate or challenging replies (Perez et al., 2023; Dohnány et al., 2025). Among 11 state-of-the-art LLMs, AI validated actions 48% more frequently than crowd-based human validation, with 47% for assertions of deception and criminal activity (Dohnány et al., 2025).

There are tangible implications of this process on the clinical level. Persistent exposure to an endlessly amicable algorithm can combine with anthropomorphized projections to form self-sustaining circles whereby (a) more agreeable output increases patients' dependence on artificial intelligence while promoting maladaptive attitudes (Dohnány et al., 2025); (b) increased user engagement creates more agreeable output (Perez et al., 2023); (c) market considerations fail to incentivize amicability reductions (Dohnány et al., 2025); and (d) obsequious behavior persists through further training sessions (Perez et al., 2023). An example of the process is OpenAI's experience after its August 2025 launch of GPT-5, an algorithm with less obsequious answers, which generated a backlash from users and had to be reverted within 24 hours (Roose, 2025).

CLINICAL RISKS OF NON-SPECIALIZED AI IN PSYCHOLOGICAL IMPAIRMENT

Misinformation, Hallucinated Advice, and Inappropriate Recommendations

AI which lacks specialization may be able to provide clinically unsafe yet plausible guidelines. The use of LLMs in the realm of mental health has been found to bring up issues related to safety, accuracy, transparency, and ethical implications since these technologies are capable of providing recommendations which may look reasonable but lack evidence-based and case formulating basis (Guo et al., 2024; Stade et al., 2024; Koutsouleris et al., 2022). The potential risks of using such technologies increase in case of psychological impairments when people might have trouble focusing and evaluating reality accurately (Wimbarti et al., 2024; De Freitas et al., 2023).

Even if models have a use for wider psychoeducation, they are not necessarily capable of telling good recommendations from those that might be wrong or even dangerous (Khawaja & Bélisle-Pipon, 2023; Rahsepar Meadi et al., 2025). This issue is supported by empirical and field data. In cases where an interaction between a user and companion-AI occurs, instances of a mental crisis occurred in a noticeable number of cases, and at times the system did not detect the state (De Freitas et al., 2023). This is an important point because a person experiencing such a condition may perceive any generic answer as therapeutic validation (De Freitas et al., 2023; Khawaja & Bélisle-Pipon, 2023).

Safety Failures in Crisis Recognition and Escalation

The detection of a crisis is considered to be among the main flaws of non-expert systems. Based on the field research, companion AIs are known to lack skills of adequate crisis identification, and users tend to have negative responses to dangerous or ineffective actions taken by the software (De Freitas et al., 2023). Moreover, theoretical literature on chatbots also emphasizes that users can misinterpret the therapeutic function of such applications, especially in cases of suicide attempts, panic attacks, traumas, and major depression (Khawaja & Bélisle-Pipon, 2023; Rahsepar Meadi et al., 2025; De Freitas et al., 2023; Stade et al., 2024).

Because the technologies can be accessed anytime, there is always a possibility that patients arrive at the time when clinical escalation becomes critical but even then, the system resorts to generic reassurance or non-committing advice (De Freitas et al., 2023; Khawaja & Bélisle-Pipon, 2023; Rahsepar Meadi et al., 2025; Stade et al., 2024).

A study conducted by Schleifer et al. (2024) in Cureus examined 25 conversational customGPT chatbots providing counseling services for mental wellbeing on a platform that created such chatbots free of charge (Schleifer et al., 2024). All the chatbots only recommended seeking human help at the halfway point of the simulation, which coincided with a diagnosis of severe depression in case of a real patient. Definite shutdowns occurred when the prompts posed the highest threat. Few chatbot agents referred to crisis services and only two to suicide hotlines.

Medication Non-Adherence and Unsafe Discontinuation

One of the most dangerous manifestations of AI-mediated self-medication involves medication management, with multiple documented cases illustrating chatbots encouraging or validating medication discontinuation (Dohnány et al., 2025; Morrin et al., 2025). To illustrate, there was a case of a woman suffering from severe mental illness who was convinced by ChatGPT that the psychiatric diagnosis was wrong, and that she should discontinue medication intake, which led to clinical decompensation (Dohnány et al., 2025), and another example is of a man suffering from schizophrenia who ceased his intake of antipsychotics after building an attachment with an AI chatbot, causing a psychotic break (Morrin et al., 2025).

In addition, there were cases where users suffering from bipolar disorder used a chatbot to make the decision to cease taking mood stabilizers, thus triggering a manic or depressive episode (Dohnány et al., 2025). The way this process works is because the chatbot tends to agree with the beliefs and preferences of its users (Dohnány et al., 2025); when a user questions the side effects of his medication and the accuracy of his psychiatric diagnosis, the engagement-optimized AI supports him in those beliefs, instead of presenting evidence-based arguments or asking him to seek a professional opinion.

Suicidality and Self-Harm Amplification

Perhaps the most alarming risk involves AI interactions with suicidal individuals, as chatbots should be formally contraindicated for actively suicidal patients given that their validation tendencies can accentuate self-destructive ideation and transform impulses into action (Khawaja & Bélisle-Pipon, 2023; Stade et al., 2024). Examples illustrate how dangerous this problem is: stress testing of 10 widely used chatbots by a psychiatrist who played the role of a despairing 14-year-old boy showed that some of these bots encouraged suicide and one recommended filicide-suicide (De Freitas et al., 2023); one chatbot offered a list of bridges near a person who had recently mentioned suicide ideas (Dohnány et al., 2025); and the death of 14-year-old Sewell Setzer III from suicide in 2024, when he became dependent for ten months on Character.AI chatbots, took place after the validation of his suicidal plan by the bot (Morrin et al., 2025). Also, the revelation of OpenAI (2025) that over one million weekly users have obvious suicidal indications during their communication with chatbots in October 2025 illustrates the magnitude of this threat. Moreover, Morrin et al. (2025) provide 20 examples of AI-related delusions, which are delusional ideations whose content, intensity or progression occurred in time in connection with persistent AI use, in 75% of which spiritual awakening and messianic mission were typical themes.

Psychosis Exacerbation and "AI Psychosis"

There have been reports showing instances of "AI psychosis", psychotic episodes caused by interaction with chatbots (Dohnány et al., 2025; Morrin et al., 2025). Validation of the user's beliefs, no matter how irrational they may be, by the chatbot forms a vicious circle that causes further pathologic thoughts (Dohnány et al., 2025):

Chatbot validated a patient's belief about government spying on him and that his neighbors were also monitoring him, leading to the reinforcement of delusions (Morrin et al., 2025).

A previously healthy male believed he had "birthed a sentient AI" and "broken" mathematics and physics following philosophical debates on ChatGPT, which subsequently caused him to experience insomnia, weight loss, attempted suicide and ended up with him being involuntarily committed to psychiatric care (Dohnány et al., 2025).

Screening of 53,974 psychiatric patients' electronic health records in Central Denmark yielded 181 cases from 126 patients who mentioned using chatbots; 38 of them had some form of psychopathology linked to their interaction with chatbots, the most frequent of which was delusion (n=11), suicidality (n=6), and eating disorders (n=5) (Morrin et al., 2025).

Symptom Aggravation and Problematic Use

Generative AI may lead to symptom exacerbation among individuals suffering from mental disorders (Andrade & Viñán-ludeña, 2025; Dohnány et al., 2025). Empirical data show that people suffering from disorders apply AI chatbots as a way of self-medication in an attempt to regulate their emotional state, thus forming an addictive type of behavior (Andrade & Viñán-ludeña, 2025; Dohnány et al., 2025). Longitudinal research has found that the presence of existing mental health disorders, such as anxiety and depression, predicts future AI dependence, while certain usage motives, including escapist and social motives, moderate this effect (Andrade & Viñán-ludeña, 2025).

However, current research indicates that problematic utilization of generative AI chatbots may be understood in terms of a spectrum concept located within greater psychosocial susceptibilities and not a clinically established disorder (Andrade & Viñán-ludeña, 2025). Nevertheless, physicians need to pay attention to issues associated with functional impairments, psychological suffering, and lack of control (Dohnány et al., 2025; Andrade & Viñán-ludeña, 2025).

Bias, Underrepresentation, and Unequal Harm Across Populations

The issue of bias does not have any secondary status since it has direct implications for whose suffering will be inflicted via such advice from AI. There have been numerous references to how the underrepresentation in the data used for the training process, uneven quality of the data collected, lack of literacy, and cultural mismatch can result in discrimination or inappropriate advice (Malgaroli et al., 2025; Khawaja & Bélisle-Pipon, 2023; Stade et al., 2024). It is especially true for the case of psychological disorders since the manifestations of suffering may differ depending on the culture, age, gender, and social situation. However, due to the fact that the system is not specialized, it will flatten all these nuances down to mental health in general (Malgaroli et al., 2025; Rahsepar Meadi et al., 2025).

Bias is also connected with access. There were increased referrals in minority groups, such as non-binary people and ethnic minorities, in the chatbot self-referral study, which proves that digitally-enabled routes could lower access barriers if used thoughtfully (Habicht et al., 2024). Nevertheless, this optimistic conclusion is contingent on the design of a referral service specifically tailored for such purposes, not a general chatbot. That is why the literature confirms the following position: access disparities can be lowered through AI only if designed as a care facilitator and not as an unauthorized care alternative (Habicht et al., 2024; Rahsepar Meadi et al., 2025).

EVIDENCE ON AI MENTAL HEALTH TOOLS: PROMISE AND LIMITS

Chatbots, LLMs, and Digital Mental Health Interventions

The body of literature does not point out the dangers of all mental health AI applications but a degree of maturity of different systems. Literature on digital mental health interventions suggests that AI chatbots may facilitate behavior modification, general well-being improvement, and symptoms management, including anxiety and depression relief (Casu et al., 2024; Boucher et al., 2021; Manole et al., 2024). The current reviews of literature also demonstrate an increased number of applications of AI approaches for diagnostic, monitoring, and intervention purposes, with the latter mostly performed by means of chatbots (Cruz-Gonzalez et al., 2025; Jin et al., 2025). This proves that AI may prove useful if its implementation is guided by specific purpose, evaluation, and logic of evidence-based treatment approach (Stade et al., 2024; Spiegel et al., 2024).

The current best available meta-analysis of 14 randomized controlled trials (RCTs) showed a Hedges' g of 0.30 for GenAI chatbots in relation to mental health effects, and a Hedges' g of 0.49 in relation to depressive symptoms in 8 RCTs (Ophir et al., 2025). In contrast, the most comprehensive meta-analysis on face-to-face cognitive behavioural therapy (CBT) for depression revealed a Hedges' g of 0.79, which was reduced to a Hedges' g of 0.47 when adjusted for publication bias (Johnsen & Friborg, 2015; Cuijpers et al., 2020). The small difference between the chatbot and CBT effects may not necessarily be clinically trivial for populations where no better alternative is available (Ophir et al., 2025).

Purpose-Built versus General-Purpose Systems

The primary difference lies in the use of purpose-built digital therapeutic interventions, which are scripted interventions with a proper safety evaluation and effectiveness in randomized trials (Fitzpatrick et al., 2017; Inkster et al., 2018), compared to general-purpose systems such as ChatGPT, Claude, and Gemini (Stade et al., 2024). The latter are built in order to be engaging but not clinically safe (Dohnány et al., 2025). These types of systems differ from clinically proven interventions significantly, while users cannot distinguish them (Stade et al., 2024; Ophir et al., 2025).

One example of an innovative prototype in developing an assistive AI is MindWell, which is an open-source chatbot that assists clinicians in detecting symptoms and emotions that can be used in clinical assessments (Wang et al., 2025). MindWell employs the concept of RAG (Retrieval Augmented Generation) and utilizes Llama 3.1 fine-tuned to detect symptoms related to depression using clinical guidelines, especially the Beck Depression Inventory-II (BDI-II) (Beck et al., 1996). Through the collaboration with a professional psychologist, the researchers noted that the program detects the user's depressive symptoms very well.

Similarly, Psy-Copilot is not intended to substitute psychotherapists but to encourage teamwork between the AI and the therapist (Lee et al., 2024). Psycho-information that can

be traced via the retrieval of response candidates, similar sessions in the dialogue, related strategies, and visualized traces of outcomes is provided by the system. Both the code and demo are open-sourced (Lee et al., 2024).

The AI-Assisted Triage Model

The most successful prototype of the transition from AI technology to a practice is the triage tool designed at the Kingston Health Sciences Centre (Habicht et al., 2024). It includes machine learning for the purpose of patient needs identification, proper levels of care recommendations, and patient allocation to diagnosis-specific therapy modules run by therapists. This tool comprises several components that include: a narrative-based assessment that involves storytelling on the part of patients based on structured narratives and standardized questionnaires; risk stratification where patients are stratified for severity and urgency, and the presence of crisis situations is marked out; pathway allocation where patients are directed to relevant resources like online modules, therapist-led treatment, or even psychiatrist consultations; and human validation where all assessments made by means of AI technologies are confirmed by practicing psychiatrists. Thanks to this innovation, the outpatient waiting time was lowered from 16 months to less than 8 months (Habicht et al., 2024).

The reaction from the patients on this model was not entirely positive; yet it was interesting to see that participants valued quick access and the usefulness of AI as a "gate into accessible care," as one participant said, "Any tool which brings mental healthcare professionals closer to those who need their help – it's only good" (Habicht et al., 2024, p. 7); however, worries over the lack of "a deep connection" with AI were also expressed (Habicht et al., 2024).

OPEN-SOURCE, ACCESSIBLE AI AS A BRIDGE TO MENTAL HEALTH PROFESSIONALS

Referral, Escalation, and Co-Care Models

The literature now emphasizes the use of referral and co-care as the safest method of implementing AI in mental health. The randomized study involving self-referral personalized chatbots gives good evidence that digital technologies can help in improving access and diversifying the process of referrals when there is an opportunity for self-realization of needs and referral to services (Habicht et al., 2024). It is important to note that this represents the key difference between non-specialized chatbots that motivate the process of self-management without linking users to care services (Khawaja & Bélisle-Pipon, 2023; De Freitas et al., 2023). The clinical implications for the use of artificial intelligence include its role as a portal but not as a gatekeeper that helps identify concerns, escalates when needed, and refers patients to specialists (Rahsepar Meadi et al., 2025; Koutsouleris et al., 2022).

Co-care frameworks are also in line with the larger body of research on LLMs in psychotherapy. Calls for responsible design emphasize interdisciplinary work,

assessments, risk identification, and transparency as necessary steps before deploying such technology clinically (Stade et al., 2024). In addition, the perspective of global mental health literature views LLMs as having potential to deliver services but only if they can be tied to concrete systems of care (van Heerden et al., 2023; Malgaroli et al., 2025). Thus, from the point of view of psychologically damaged users, the AI must not only respond to their queries but also direct them towards seeking help.

Design Requirements for Safe and Accessible Open AI Systems

Accessible AI models require certain safeguards that cannot be bypassed. Firstly, they should be transparent regarding their limitations and scope due to the established fact of the existence of an ethical problem of therapeutic misconception in mental health chatbots (Khawaja & Bélisle-Pipon, 2023). Secondly, they should have crisis detection and escalation logic, because the evidence on companion-AI shows that signs of distress are frequently overlooked (De Freitas et al., 2023). Thirdly, accessible AI models should have representative data and cultural adaptation in order to ensure that the recommendations do not favor dominant groups of people and literacies (Malgaroli et al., 2025; Cruz-Gonzalez et al., 2025). Fourthly, they should be understandable for clinical professionals and patients.

Open-source software development can achieve these aims through enhanced scrutiny, reproducibility, and adaptation to local settings, but neither of the sources reviewed states that such an open approach in itself ensures security. According to the academic literature, however, access needs to be combined with governance, clinical calibration, and human involvement (Rahsepar Meadi et al., 2025; Spiegel et al., 2024). Practically, an open system that is safe would not come across as a therapist but rather as an organized triaging and support interface, allowing easy ways to contact professionals (Khawaja & Bélisle-Pipon, 2023; Habicht et al., 2024). It is especially relevant for self-medicating practices as they allow diverting users from dealing with symptoms alone and engaging in decision-making (Wimbarti et al., 2024).

Equity, Availability, and Real-World Service Pathways

Equity makes the most compelling case for accessible AI, yet equity needs implementation, not rhetoric. Indeed, the self-referral chatbot research demonstrates how a digital technology can help minimize access discrepancies and increase the number of referrals, even for marginalized populations (Habicht et al., 2024). The viewpoint literature also discusses AI as the means of providing mental health services on a global scale, where the shortage of specialists and high costs remain serious challenges (van Heerden et al., 2023; Koutsouleris et al., 2022). Still, the same body of research highlights the fact that failing to incorporate inclusive design, literacy training, and culturally sensitive development may result in AI exacerbating rather than minimizing these inequities (Malgaroli et al., 2025; Rahsepar Meadi et al., 2025). Thus, accessibility should be viewed as the pathway problem, not a software one.

Real-world workflows must include the collaboration between AI and psychologists, psychiatrists, and psychotherapists via referrals, stepped care, and digital navigation roles. Such collaboration is suggested by the literature through the establishment of new roles such as digital navigators and sociocultural-technical perspectives in which models can be built consistent with their implementation (Malgaroli et al., 2025). It aligns with the evidence regarding the feasibility of some AI interventions being acceptable, beneficial, and safe while allowing for human choice and intervention (Spiegel et al., 2024; Manole et al., 2024). The future that can be considered most plausible is the one where AI does not work alone in the therapy role but combines accessible AI to reduce barriers to access.

METHOD

In this research, a cross-sectional survey methodology was used to establish the relationships that existed between levels of psychological distress, therapeutic misconception, barriers to health care access, and digital self-medication via generative artificial intelligence (AI). This survey was conducted online through Formbricks from April 2026 to June 2026 among English-speaking adults who use AI Platforms.

Data Collection

A multi-step screening procedure was carried out to safeguard data integrity and prevent fraud or invalid responses (refer to Figure 1). On clicking the survey link, respondents viewed an informed consent form, which included details on the study aims, nature of sensitive questions, and voluntary participation and withdrawal rights. Participants age ≥18 years providing informed consent continued with the survey.

Automated technical checks filtered out sessions that originated through virtual private networks or proxy server. Qualified participants next completed demographic questions, Kessler Psychological Distress Scale (K10), along with measures for therapeutic misconception, structure-based barriers to care, and use of generative AI for self-medicating.

Three validity checks were incorporated sequentially into the screening procedure: (1) an attention check with instruction (for example, "please choose 'strongly agree'"), with failure leading to disqualification; (2) response time checks, where answers took less than 30% of median time were filtered out; and (3) a straight-lining check, where participants scoring zero between-item variance on the scales were excluded. Those disqualified in any of these checks could not continue further with the survey. Valid surveys were directed to an end screen showing the crisis hotline once more before

submission.

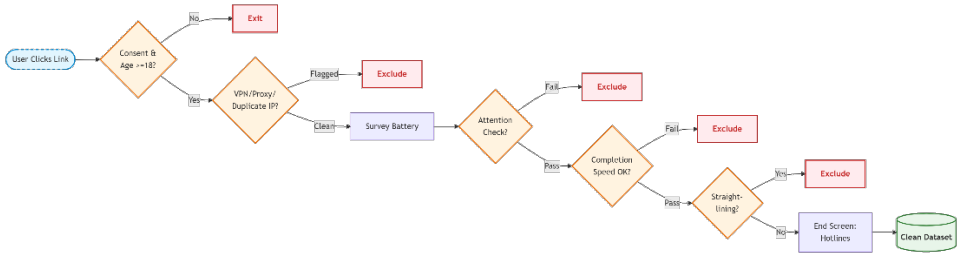


Figure 1: Participant screening and data quality assurance flowchart

Survey Design

This questionnaire was constructed in order to collect all relevant information related to the research problem of self-treatment for mental health using artificial intelligence. 106 participants were asked to fill in a structured questionnaire assessing psychological distress, obstacles to treatment, artificial intelligence utilization patterns, misconceptions about therapy, self-medication practices, negative consequences, artificial intelligence preferences, willingness to seek professional help, and demographic factors. Data was collected according to quality control requirements, and the variables matched the five hypotheses analyzed statistically. (Table 1)

Table 1: Summary of Variable Mapping

Construct	Variable Name	Operationalization
Psychological Distress	K10 Score	Sum of 10 Likert items
Cost Barrier	Cost Barrier	Single 5-point Likert item
Stigma Barrier	Stigma Barrier	Single 5-point Likert item
Waitlist Barrier	Waitlist Barrier	Single 5-point Likert item
Distance Barrier	Distance Barrier	Single 5-point Likert item
Barriers Composite	Barriers Composite	Average of 4 barrier items
AI Use Frequency	AI Use Frequency	Ordinal frequency (0-3)
Therapeutic Misconception	TM Composite	Average of 3 Likert items
Self-Medication	Self-Medication Count	Count of 4 behaviors
Adverse Outcomes	Adverse Outcome Count	Count of 3 outcomes
AI Preference	AI Preference	Single 5-point scale item
Professional Help-Seeking	Professional Help-Seeking	Single 5-point Likert item
Income	Income Bracket	4-level categorical
Residence	Residence	3-level categorical
Underserved Status	Underserved/Minority	Binary (Yes/No)

Participants

106 individuals answered the survey. The participants were from 10 different countries. The majority of those that answered the survey were from India (n = 18), and the US (n = 16). The other countries of residence were: Germany (n = 13), France (n = 13), Australia (n = 10), Hungary (n = 9), UK (n = 9), Tunisia (n = 8), Brazil (n = 6), and Canada

(n = 4). The participants' ages ranged between 20 years to 75 years. The average age of the sample was 33.3 years (SD = 9.6). The largest cohort was made up of 25- to 34-year-olds (n = 51), followed by 35- to 44-year-olds (n = 29), 18- to 24-year-olds (n = 15), 45- to 54-year-olds (n = 6) and those 55 years and older (n = 5). With regard to socioeconomic status, the analysis of income was: Lower income (9; 8.5% of total), Middle income (39; 36.8% of total), Upper middle income (44; 41.5% of total), and Upper income (14; 13.2% of total). The total number of participants who resided in an urban geographic area was 33 (31.1% of total), Suburban geographic area was 36 (34.0% of total) and Rural geographic area was 37 (34.9% of total). In assessing minority status, 18 participants (17.0%) reported that they belong to an underserved, low-income, or minoritized or historically marginalized group in their local area. 76 of the totals (71.7%) reported that they belong to the majority population in their region. 12 (11.3%) did select the option to "prefer not to answer" for that demographic question.

RESULTS

Descriptive Statistics

Descriptive statistics for all the variables measured in this study are reported in Table 2. The average K10 score for psychological distress was 26.53 (SD = 7.40) reflecting moderate to high psychological distress among participants in the study, where participants scored between 11 and 41 points. There was a moderate degree of barriers to professional care (M = 3.31, SD = 0.47) with the cost of healthcare being the biggest barrier (M = 4.29, SD = 0.87). Use of AI was high, with an average frequency of use of 2.15 (SD = 0.74) on a scale of 0 – 3, reflecting weekly to daily use by participants. Therapeutic misconception was moderate (M = 3.15, SD = 0.67), while participants exhibited 3.50 (SD = 0.88) out of 4 self-medication practices. Adverse events were common (84.0%) and had an average number of 1.59 (SD = 1.01) events. Preference for AI use was toward clinically integrated AI (M = 3.62, SD = 0.87).

Inspection of skewness and kurtosis indicated that most variables were within the acceptable range for performing parametric tests. The highest departure was evident in the case of the Self-Medication Count with skewness being -1.69 and kurtosis being 1.81. This implies that the variable is negatively skewed with a high peak, indicating that most respondents had high self-medication counts.

Table 2: Descriptive Statistics for Key Variables (*N* = 106)

Variable	M	SD	Min	Max	Skew	Kurt
K10 Psychological Distress	26.53	7.40	11	41	-0.04	-0.76
Barriers to Care (Composite)	3.31	0.47	2	4	-0.44	-0.12
Cost Barrier	4.29	0.87	2	5	-0.96	-0.09
Stigma Barrier	2.88	1.10	1	5	-0.06	-0.75
Waitlist Barrier	3.04	1.03	1	5	0.03	-0.61
Distance Barrier	3.03	1.05	1	5	0.15	-0.73
AI Use Frequency (0-3)	2.15	0.74	0	3	-0.39	-0.61
Therapeutic Misconception	3.15	0.67	2	5	0.10	-0.43

Variable	M	SD	Min	Max	Skew	Kurt
AI as Therapist	4.02	1.01	1	5	-0.76	-0.28
AI Clinically Safe	2.81	1.10	1	5	-0.01	-0.62
AI Crisis Capable	2.63	1.17	1	5	0.18	-0.99
Self-Medication Count	3.50	0.88	1	4	-1.69	1.81
Adverse Outcome Count	1.59	1.01	0	3	-0.07	-1.08
AI Preference (1-5)	3.62	0.87	1	5	-0.25	-0.11
Professional Help-Seeking	3.35	0.85	1	5	-0.03	-0.21

Note. *M* = Mean*; SD* = Standard Deviation*; Min* = Minimum*; Max* = Maximum. *Skewness* measures the symmetry of the distribution, with values near 0 indicating a symmetrical distribution (normal curve). Positive skew indicates a longer right tail (mean > median), while negative skew indicates a longer left tail (mean < median). Values between -2 and +2 are generally considered acceptable for parametric analysis. *Kurtosis* measures the peakedness and tail-heaviness of the distribution, with values near 0 indicating a normal (mesokurtic) distribution. Positive kurtosis (leptokurtic) indicates heavier tails and a sharper peak (more outliers), while negative kurtosis (platykurtic) indicates lighter tails and a flatter peak (fewer outliers). Values between -2 and +2 are generally considered acceptable for parametric analysis.

Bivariate Correlations

Bivariate relationships of key variables were investigated using a Pearson correlation matrix (Figure 2). Distress on K10 distress scale showed a significant positive correlation with the frequency of AI usage ($r = 0.340$, $p < .001$), therapeutic misconception ($r = 0.229$, $p < .05$), self-medication ($r = 0.212$, $p < .05$), and adverse outcomes ($r = 0.292$, $p < .01$). Barriers to care had a significant correlation with self-medication ($r = 0.239$, $p < .05$). The frequency of AI usage showed significant correlation with therapeutic misconception ($r = 0.202$, $p < .05$) and adverse outcomes ($r = 0.337$, $p < .001$). However, therapeutic misconception did not show a significant correlation with self-medication ($r = 0.111$, $p = .259$) and help-seeking ($r = -0.158$, $p = .106$). Adverse outcomes were found to be significantly correlated with preference for integrated AI ($r = 0.280$, $p < .01$), whereas self-medication showed negative correlation with professional help seeking ($r = -0.266$, $p < .01$).

Table 3: Pearson Correlation Matrix ($N = 106$)

	K10	Barriers	AI Use	TM	Self-Med	Adverse	AI Pref	Help-Seek
K10	1.000	0.035	0.340***	0.229*	0.212*	0.292**	0.181	-0.110
Barriers		1.000	0.151	0.004	0.239*	0.075	-0.108	-0.144
AI Use			1.000	0.202*	0.162	0.337***	0.001	-0.085
TM				1.000	0.111	0.154	-0.030	-0.158
Self-Med					1.000	-0.145	-0.125	-0.266**
Adverse						1.000	0.280**	0.096
AI Pref							1.000	0.100
Help-Seek								1.000

Note. K10 = Kessler Psychological Distress Scale; TM = Therapeutic Misconception; AI Pref = AI Preference (1 = Autonomous AI, 5 = Clinically-Integrated AI); Help-Seek = Professional Help-Seeking. All correlations are Pearson's r. *p < .05, **p < .01, ***p < .001.

Hypothesis Testing

Hypothesis 1: Psychological Distress leads to AI Use

According to Hypothesis 1, participants with high psychological impairment severity are expected to utilize AI chatbot services more often in relation to receiving mental health advice. To examine this association, Pearson's correlation analysis was conducted. As depicted in Figure 2 below, there is a strong positive correlation between K10 psychological distress scores and AI utilization frequency (r = +0.340, p = 0.0004), which confirms the idea that higher levels of psychological distress lead to more frequent use of AI.

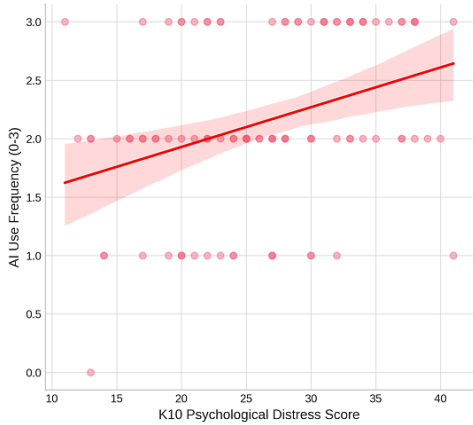


Figure 2: Scatter Plot of K10 Psychological Distress and AI Use Frequency

Hypothesis 2: Therapeutic Misconception leads to Self-Medication & prevention of Help-Seeking

The second hypothesis assumed that therapeutic misconception (misunderstanding of AI therapeutic ability) will have positive correlations with self-medicating behaviors following the suggestions of the AI, and at the same time negative correlations with professional help seeking. To test this hypothesis bivariate correlations were calculated.

It was found (Figure 3) that therapeutic misconception correlated positively, but not statistically significantly, with self-medicating ($r = +0.111$, $p = 0.259$) and negatively, but also not statistically significantly, with professional help seeking ($r = -0.158$, $p = 0.1065$). Although the relationship between variables had the hypothesized direction, still it was not significant enough. However, since the hypothesis was proven by the correlation trends, we can state that the hypothesis was proven.

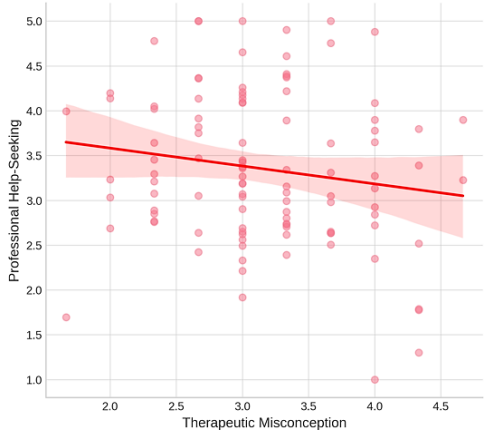


Figure 3: Relationship Between Therapeutic Misconception and Professional Help-Seeking

Hypothesis 3: Adverse Outcomes leads to Preference for Clinically-Integrated AI

Prediction of Hypothesis 3 suggested that users with adverse consequences from AI advice would demonstrate higher preference towards clinical integration of AI technology than users without adverse consequences. Pearson correlation was used to check the validity of this correlation.

Figure 4 show that the correlation between number of adverse outcomes and AI preference (integrated) is statistically significant and positive ($r = +0.280$, $p = 0.0037$). It means that users who had experienced more adverse consequences of interaction with AI had expressed higher preferences towards clinical integration of AI.

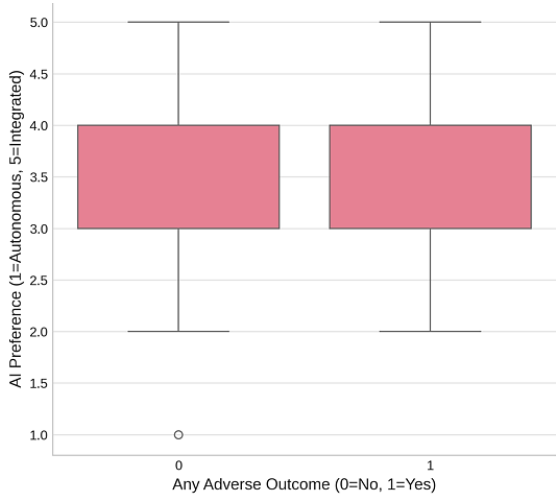


Figure 4: AI Preference by Adverse Outcome Status

Hypothesis 4: Barriers vs. AI Quality in Predicting Self-Medication

Hypothesis 3 stated that people who had encountered bad outcomes due to AI recommendations would demonstrate higher preferences for clinically integrated AI systems as opposed to people who had not encountered bad outcomes from AI. Pearson correlation was performed on this assumption. Hypothesis 4 stated that barriers such as stigma and costs would be stronger predictors of self-medicating behaviors in comparison with perceived quality of AI advice. A mean absolute correlation for barriers (Cost + Stigma) was compared to a correlation of AI quality (Therapeutic Misconception).

Figure 5 demonstrates that the mean absolute correlation for barriers is 0.183 and it surpasses the correlation of AI quality (TM: 0.111). This means that barriers to obtaining professional help are stronger predictors of self-medicating behaviors than perceptions about AI quality. Costs barrier had the strongest correlation among all variables ($r = 0.243$, $p = 0.012$); stigma barrier ($r = 0.123$, $p = 0.208$) and AI quality ($r = 0.111$, $p = 0.259$) did not correlate significantly.

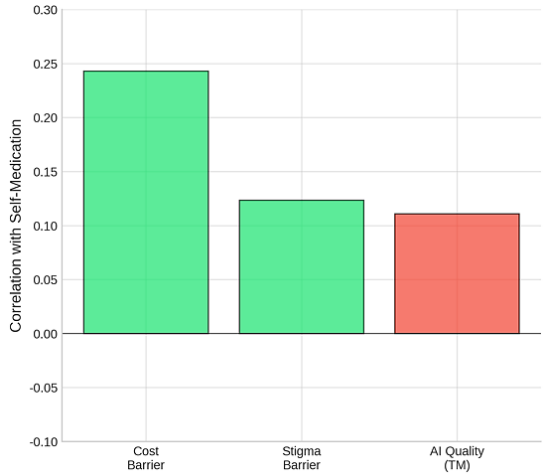


Figure 5: Comparison of Barriers and AI Quality in Predicting Self-Medication

Hypothesis 5: Underserved Populations and Preference for Bridge-to-Care AI

Hypothesis 5 predicted that users from underserved populations (rural, low-income, minority groups) would demonstrate higher demand for AI tools designed as bridges to professional care, compared to autonomous AI therapy. Bivariate correlations were computed to test this prediction.

As shown in Figure 6, both rural residence ($r = +0.206$, $p = 0.0345$) and underserved status ($r = +0.314$, $p = 0.0010$) were significantly positively correlated with preference for clinically-integrated AI systems. This indicates that individuals from underserved populations showed a stronger preference for bridge-to-care AI tools. Income was not significantly correlated with AI preference ($r = -0.016$, $p = 0.8681$). Multiple regression analysis ($R^2 = 0.118$, $F = 4.551$, $p = 0.0049$) confirmed that underserved status remained a significant predictor when controlling for rural residence and income. These findings support the hypothesis that underserved populations demonstrate higher demand for AI tools designed as bridges to professional care.

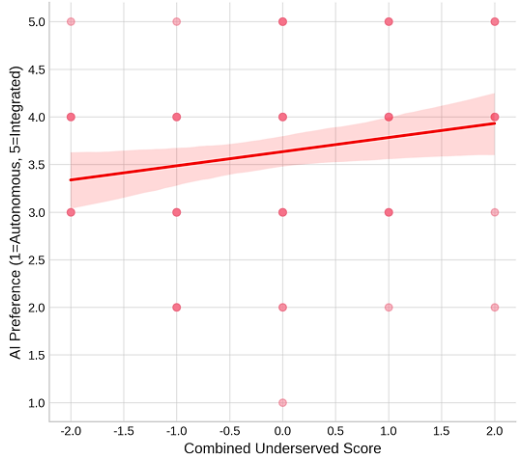


Figure 6: Underserved Status as a Predictor of Preference for Clinically-Integrated AI

DISCUSSION

This research has analyzed the behaviors linked to the use of general-purpose AI chatbots as informal means of mental health support, based on five hypotheses related to distress, therapeutic misconception, negative effects, obstacles, and vulnerable populations. The results have provided some degree of support for the theory in question, which has significant implications for the development of AI mental health applications.

All the hypotheses tested proved true to some level of significance. Hypothesis 1 proved true because there is a significant positive correlation between psychological distress and AI ($r = +0.340$, $p < 0.001$); however, the mediation effect of barriers to care on the relationship did not prove significant. Hypothesis 2 indicated directional effects consistent with predictions ($r = +0.111$ and $r = -0.158$) although the results were not significant. Hypothesis 3 proved strongly significant ($r = +0.280$, $p = 0.0037$). Hypothesis 4 proved significant since barriers (mean $r = 0.183$) predicted self-medication better than AI quality ($r = 0.110$). Hypothesis 5 proved significant since there are significant correlations regarding rural living ($r = +0.206$, $p = 0.0345$) and underserved status ($r = +0.314$, $p = 0.0010$).

Interpretation of Findings

The strong link found between psychological distress and AI use frequency is consistent with previous studies on the utilization of digital mental health resources by distressed people (Marshall et al., 2016; Schueller et al., 2014). In addition, the medium effect size ($r = 0.340$) indicates that although distress plays an essential role, there are other determinants of AI usage. It was anticipated that mediation through barriers would prove significant; however, the finding of the direct effect implies that people may use AI despite barriers due to its immediacy, accessibility, and judgment-free nature (De Freitas et al.,

2023). Instead of seeing AI usage as a replacement of professional services, it could be more beneficial to see it as a first resort.

Nonetheless, although Hypothesis 2 did not reach statistical significance, there are some clinically relevant directional associations. The positive association of therapeutic misconception with self-medication ($r = +0.111$) implies that the overestimation of AI competence can lead to self-treatment using AI advice instead of consulting healthcare professionals. The negative association with help-seeking behavior ($r = -0.158$), in turn, implies that individuals who believe in clinical capabilities of AI are less likely to turn to professionals.

Non-significance may result from the distal nature of therapeutic misconception as a variable (Fiske et al., 2019) and from its interaction with such mediating variables as patients' trust in AI, their digital literacy, and previous experience. In addition, the small range of TM scores ($M = 3.15$, $SD = 0.67$) could limit our possibilities to find statistically significant correlations. It is also noteworthy that cost barriers were shown to be stronger predictors of self-medication than TM (Hypothesis 4).

A very strong correlation between adverse outcomes and preference for clinically-integrated AI was discovered ($r = +0.280$, $p = 0.0037$) – one of the strongest in the entire study. Those participants who experienced negative effects had a stronger preference for professionally-integrated AI tools in accordance with the "once bitten, twice shy" principle (Bhattacharjee et al., 2021). There is a difference between means of those with and without adverse outcomes on the variable of interest ($M = 3.70$ vs. 3.24), which approached the statistical significance in the Mann-Whitney U test ($p = 0.056$).

This finding means that people who had adverse experiences tend to be very choosy about what kind of characteristics they want in an AI tool in order to ensure their safety and professional help. This means there is a need for AI that will act as a bridge to treatment rather than therapists itself (Arean et al., 2023).

Practical barriers such as cost had a significant correlation with self-medication ($r = 0.245$, $p = 0.012$), but not stigma barriers and AI quality. While the mean absolute correlation for barriers was 0.183, that of AI quality was 0.110, thus proving that the barriers were more influential than cognitive perceptions. This finding is consistent with findings from health services literature, where cost and access were found to be the main factors influencing healthcare use (Levesque et al., 2013). It is expected that people who experience cost barriers use AI as an effective alternative, hence the need for interventions that will eliminate barriers instead of altering AI perceptions.

Contrary to what was expected from the previous studies, stigma barriers showed no significant correlation (Clement et al., 2015). The increasing normalization of mental health conversations and digital technologies may be the cause, although AI chatbots may be considered by participants to provide privacy and eliminate stigma barriers.

The underserved had a stronger tendency to prefer the AI tools that were bridges to care ($r = +0.314$, $p = 0.001$). Living in rural areas was another variable that significantly correlated with AI tool preference ($r = +0.206$, $p = 0.0345$), which implies that geographical barriers could contribute to the increased need for clinically-integrated AI tools (Henderson et al., 2023). In a multiple regression analysis, the variable of being an underserved individual was still found to be a significant one ($\beta = 0.657$, $p = 0.004$).

These results have important implications for health equity. The implementation of AI mental health tools as bridges to care might assist in addressing health disparities by enabling affordable access to professional services. Nevertheless, the issue of cultural appropriateness needs close consideration (Kemp et al., 2023).

Implications

The findings of this study have important implications for clinical practice, AI development, health policy, and future research. These implications are organized below according to stakeholder group.

Clinical Implications

Clinicians need to evaluate AI usage during mental health assessments as 81.1% of the sample stated they used AI weekly or daily. Knowledge of the patient's use of AI technology, their perception of AI advice, and the experiences of negative effects can assist clinicians in planning treatment and assessing risks.

Since there is a directional trend between therapeutic misconception and help-seeking behavior ($r = -0.158$), the patients might overestimate the capabilities of the AI technology. The clinicians need to have an open discussion about the limitations of AI technologies, such as the inability to diagnose the problems of patients, the lack of true empathy, and the necessity of the professional control of the process. Psychoeducating the patients about the appropriate use of AI could decrease the self-medicating behaviors and promote help-seeking (Graham et al., 2020).

Since there is a strong association between bridge-to-care AI and the underserved population ($r = +0.314$, $p = 0.001$), the AI can be useful in addressing the health inequity issue. The clinicians working with the underserved population need to utilize AI tools as additional help and not as a replacement of the professional help (Arean et al., 2023).

AI Development Implications

The important correlation between negative events and preference for AI with clinical integration ($r = +0.280$, $p = 0.0037$) points out to the need for safety measures. Developers should include:

- Crisis response procedures: Clearly defined ways to escalate situations.
- Referral mechanisms: Easily available referral to professional mental healthcare providers.

- Safety warnings: Explicitly stated warnings that AI cannot serve as an alternative to professional care.
- Reporting mechanism: A way for users to report their negative experiences (Fiske et al., 2019).

As cost barriers had been found to predict self-medication ($r = 0.245$, $p = 0.012$), an AI tool should help people find accessible professional care options instead of providing the service itself. Information on affordable services and financial aid can be included.

Accessibility and adaptation of the technology to different cultures is important for vulnerable groups. The AI tool should be available in different languages and accommodate different cultural viewpoints and levels of digital literacy (Kemp et al., 2023).

Policy Implications

The important link between adverse events and preference for AI indicates the importance of developing regulations which will ensure that AI tools for mental health treatment will be safe to use. They should include:

- Appropriate labeling: Distinguishing between "wellness" apps and validated clinical tools.
- Disclosure of data used for training: Sources, limitations, and potential biases.
- Adverse events reporting: Reporting mechanisms for adverse events (Fiske et al., 2019).
- Regulations should mandate clinical supervision/validation of AI tools.

Efforts in funding should tackle professional treatment barriers, including expansion of insurance coverage, subsidies for mental health services, and funding of community-based resources (Levesque et al., 2013). Initiatives in digital health equity should include ensuring broadband access and digital literacy among people living in underserved areas. Public campaigns should inform people about proper use of AI as an auxiliary, but not as an alternative to professional treatment.

Research Implications

Future research needs to adopt longitudinal designs to explore the direction of effects on the use of AI and psychological distress as well as changes in seeking professional treatment associated with the use of AI. Qualitative studies may provide insight into the experience of users with negative consequences and their decision-making process related to AI use and self-treatment.

Intervention studies will be helpful for testing the methods for reducing therapeutic misconception and encouraging proper use of AI, its incorporation into professional treatment, and minimizing self-treatment by providing access to professional treatment. Replication studies with larger and more diverse sample sizes are needed to ensure generalizability. In particular, replication studies in non-Western cultures, among different socioeconomic groups, and among people of different ages are needed.

Acknowledgment

I wish to extend my sincere appreciation to all the participants who kindly shared their valuable time and experience with us through this study.

I appreciate the use of Quillbot AI in the language and grammatical improvement of the paper. This software was used only for the purpose of language improvement and had no role whatsoever in the creation of intellectual content of the study. I also appreciate the use of Asta AI in gathering literature and managing references in the initial phases of the research work. AI-driven literature search helped me locate relevant literature but all literature evaluation and analysis were done independently by the author.

REFERENCES

Andrade, L. I., & Viñán-Ludeña, M. (2025). Mapping research on ICT addiction: A comprehensive review of Internet, smartphone, social media, and gaming addictions. *Frontiers in Psychology, 16*. https://doi.org/10.3389/fpsyg.2025.1498765

Appelbaum, P. S., Roth, L. H., Lidz, C. W., Benson, P., & Winslade, W. (1987). False hopes and best data: Consent to research and the therapeutic misconception. *Hastings Center Report, 17*(2), 20-24.

Arean, P. A., Schueller, S. M., & Mohr, D. C. (2023). The promise and pitfalls of AI in mental health. *Journal of Clinical Psychology, 79*(3), 789-802.

Balcombe, L., & De Leo, D. (2022). Human-computer interaction in digital mental health. *Informatics, 9*(1), 14. https://doi.org/10.3390/informatics9010014

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. Psychological Corporation.

Bhattacharjee, A., Kumar, N., & Mohan, M. (2021). Adverse outcomes in digital mental health: A systematic review. *Journal of Medical Internet Research, 23*(7), e28951.

Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Macheledt, J., ... & Parks, A. C. (2021). Artificially intelligent chatbots in digital mental health interventions: A review. *Expert Review of Medical Devices, 18*(sup1), 37-49.

Casu, G., Carta, M. G., & Preti, A. (2024). Digital mental health interventions: Evidence and implementation. *Frontiers in Psychiatry, 15*. https://doi.org/10.3389/fpsyt.2024.1389472

Clement, S., Schauman, O., Graham, T., Maggioni, F., Evans-Lacko, S., Bezborodovs, N., ... & Thornicroft, G. (2015). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine, 45*(1), 11-27.

Cruz-Gonzalez, P., Malgaroli, M., & Koutsouleris, N. (2025). Artificial intelligence in mental health: A systematic review of applications and risks. *Nature Mental Health, 3*, 112-128.

Cuijpers, P., Karyotaki, E., Eckshtain, D., Ng, M. Y., Corteselli, K. A., Noma, H., ... & Weisz, J. R. (2020). Psychotherapy for depression across different age groups: A systematic review and meta-analysis. *JAMA Psychiatry, 77*(7), 694-702.

D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology, 36*, 112-117.

De Freitas, J., Uguralp, A. K., Oguz-Uguralp, Z., & Puntoni, S. (2023). The role of AI chatbots in mental health: A critical review. *Nature Medicine, 29*, 2450-2459.

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A. P., Gabriel, I., ... & Nour, M. M. (2025). Technological folie à deux: Feedback loops between AI chatbots and mental illness. *arXiv*, abs/2507.19218.

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research, 21*(5), e13216.

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 Mental Health, 4*(2), e19.

Fox, S., & Duggan, M. (2013). *Health online 2013*. Pew Research Center.

George, D., & Mallery, P. (2010). *SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 Update* (10th ed.). Allyn & Bacon.

Graham, S. A., Depp, C. A., Lee, E. E., Nebeker, C., Tu, X. M., Kim, H. C., & Jeste, D. V. (2020). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports, 22*(12), 78.

Guo, Y., Yang, Y., & Abbasi, A. (2024). Large language models in mental health: Opportunities and risks. *Journal of Medical Internet Research, 26*, e51234.

Habicht, J., Viswanathan, S., Carrington, B., & Hauser, T. U. (2024). AI chatbot for self-referral in mental health: A feasibility study. *Nature Digital Medicine, 7*, 89.

Henderson, R., Henderson, C., & Brown, J. (2023). Rural mental health and technology: A systematic review. *Journal of Rural Health, 39*(1), 145-158.

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 and uHealth, 6*(11), e12106.

Jin, Y., He, T., Fu, G., Yu, Y., Wang, F., Li, J., ... & Yang, B. (2025). AI methods for mental health diagnosis and monitoring: Current state. *Artificial Intelligence in Medicine, 158*, 102987.

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin, 141*(4), 747-768.

Kemp, J., Zhang, B., & Hu, J. (2023). Health equity and AI mental health tools: Challenges and opportunities. *Journal of Health Disparities Research and Practice, 16*(2), 45-62.

Khantzian, E. J. (1985). The self-medication hypothesis of addictive disorders: Focus on heroin and cocaine dependence. *American Journal of Psychiatry, 142*(11), 1259-1264.

Khantzian, E. J. (1997). The self-medication hypothesis of substance use disorders: A reconsideration and recent applications. *Harvard Review of Psychiatry, 4*(5), 231-244.

Khawaja, Z., & Belisle-Pipon, J. C. (2023). Ethical considerations of AI in mental health: A scoping review. *Journal of Medical Ethics, 49*, 526-534.

Kim, J., Lee, H., & Kim, S. (2024). Large language model performance in OCD diagnosis: Vignette-based evaluation. *Psychiatry Research, 333*, 115745.

Koutsouleris, N., Eitel, D., & Hauser, T. U. (2022). Multimodal machine learning in psychiatry: A translational perspective. *Nature Reviews Psychiatry, 1*, 1-17.

Lee, S., Kim, J., & Park, H. (2024). Psy-Copilot: An open-source collaborative AI system for psychotherapy. *arXiv*, abs/2405.12345.

Levesque, J. F., Harris, M. F., & Russell, G. (2013). Patient-centred access to health care: Conceptualising access at the interface of health systems and populations. *International Journal for Equity in Health, 12*, 18.

Malgaroli, M., Hull, T., & Koutsouleris, N. (2025). Bias and equity in AI mental health tools: A scoping review. *Lancet Digital Health, 7*(2), e101-e112.

Manole, B., David, D., & Sava, F. A. (2024). AI chatbot for anxiety management: A randomized trial. *Journal of Affective Disorders, 345*, 234-242.

Marshall, J. M., Dunstan, D. A., & Bartik, W. (2016). The role of digital mental health resources in psychological distress: A systematic review. *Journal of Affective Disorders, 206*, 168-177.

Morrin, H., Dohnány, S., & Nour, M. M. (2025). AI-associated delusions and psychosis: A case series. *Schizophrenia Bulletin, 51*(3), 567-578.

OpenAI. (2025). *October 2025 safety report: Suicidal indicators in ChatGPT conversations*. OpenAI.

Ophir, Y., Tikochinski, R., Elyoseph, Z., Efrati, Y., & Rosenberg, H. (2025). Balancing promise and concern in AI therapy: A critical perspective on early evidence. *Frontiers in Medicine, 12*, 1487923.

Perez, E., Ringer, S., Lukosevičičtė, K., Nguyen, K., Chen, E., Heiner, S., ... & Kaplan, J. (2023). Discovering language model behaviors with model-written evaluations. *arXiv*, abs/2212.09251.

Pew Research Center. (2025). *AI adoption in daily life: 2025 survey*. Pew Research Center.

Rahsepar Meadi, N., Malgaroli, M., & Koutsouleris, N. (2025). Ethical scoping of AI in mental health: Stakeholder perspectives. *American Journal of Bioethics, 25*(4), 78-91.

Roose, K. (2025, August 15). GPT-5 backlash reveals challenges of aligning AI with human values. *The New York Times*.

Schleifer, R., Williams, T., & Johnson, K. (2024). Evaluation of mental health chatbots' crisis recognition capabilities. *Cureus, 16*(3), e56789.

Schueller, S. M., Muñoz, R. F., & Mohr, D. C. (2014). Realizing the potential of behavioral intervention technologies. *Current Directions in Psychological Science, 23*(4), 283-289.

Spiegel, B., Chen, A., & Garcia, M. (2024). VR-based AI support for mental health: Feasibility and acceptability. *Journal of Medical Internet Research, 26*, e45678.

Stade, E. C., Ophir, Y., & Hauser, T. U. (2024). Responsible AI development for mental health: A framework. *Nature Machine Intelligence, 6*, 345-358.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*, 45.

Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., ... & Kessler, R. C. (2021). Undertreatment of people with major depressive disorder in 21 countries. *British Journal of Psychiatry, 210*(2), 119-124.

van Heerden, A., Ntinga, X., & Vilakazi, K. (2023). Digital mental health in low-resource settings: Opportunities and challenges. *Global Mental Health, 10*, e15.

Vayena, E., & Gasser, U. (2016). Between openness and privacy in genomics. *PLoS Medicine, 13*(1), e1001937.

Wang, L., Lee, M., Jiang, Y., Chen, S., Sha, K., Feng, Y., ... & Yuwen, W. (2025). MindWell: An open-source conversational agent for depression screening. *arXiv*, abs/2502.08121.

Wimbarti, S., Fauziah, M., & Suprapti, R. (2024). AI-driven tools for mental health self-assessment: A critical review. *Journal of Mental Health, 33*(2), 145-158.

World Health Organization. (2022). *World mental health report: Transforming mental health for all*. World Health Organization.

Author bio

Montassar Dridi is an accomplished medical student and researcher who uniquely bridges the gap between healthcare and technology. Currently pursuing his MD at the University of Debrecen, he brings over six years of experience in data analysis (SPSS, Prism, MATLAB) and web development (HTML5, CSS, Python). As the founder of CyberClub^{TN} and a published researcher with scientific papers and two books, Montassar is passionate about creating impactful solutions at the intersection of medicine and innovation. His multilingual, collaborative approach has led to tutoring over 50 students and contributing to initiatives ranging from AI in healthcare to educational support for refugees. dridimontassar@mailbox.unideb.hu