Journal of Psychiatry
Psychotherapy & Communication

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support

Christian Jonathan Haverkampf

Narrative Clinical Review

Abstract—Mental-health chatbots range from scripted self-help programs to generative language models that can produce novel replies. Their fluent conversation invites a clinically important category error: a system may communicate in an empathic style without experiencing concern, assuming responsibility, or forming a reciprocal therapeutic relationship. Meta-analyses of mostly brief trials suggest small-to-moderate short-term improvements in depressive, anxiety, and distress symptoms, but heterogeneous interventions, inactive controls, selective samples, limited follow-up, commercial conflicts, and sparse adverse-event measurement restrict inference. Evidence for general-purpose generative systems is earlier and does not establish safe autonomous psychotherapy. Potential uses include psychoeducation, structured self-monitoring, rehearsal, reminders, and supervised support between sessions. Risks include confident error, missed crisis signals, biased or culturally mismatched responses, privacy loss, dependency, manipulation through engagement incentives, and delayed human care. Clinical evaluation should therefore begin with the task rather than the label AI: who is using which system, for what purpose, with what data practices, evidence, monitoring, escalation route, and accountable human oversight? A chatbot may widen access to a bounded tool. It should not be represented as a therapist merely because its language feels attentive.

Index Terms—artificial intelligence, chatbots, mental health, simulated empathy, psychotherapy, digital safety

I. INTRODUCTION

A person who opens a chatbot at two in the morning may not be asking a technical question. They may be lonely, ashamed, frightened, uncertain whether their symptoms matter, or unable to obtain care. A rapid, courteous reply can feel like relief. The system does not become clinically trustworthy because the relief is real, however. Language models generate plausible sequences of words; narrower programs select from authored pathways. Neither fact makes every reply false or useless, but neither supplies the judgment, accountability, clinically governed longitudinal memory, embodied attention, or moral commitment expected of a mental-health professional.

The distinction is easy to lose because conversation is itself relational. People infer attention from responsiveness and concern from well-formed empathic language. ELIZA demonstrated decades ago that simple pattern matching could invite users to attribute understanding to a program (Weizenbaum, 1966). Contemporary systems are much more capable, yet fluency still does not disclose how an answer was produced, whether it is accurate, or what will happen if the user deteriorates. The clinical question is therefore not whether a response sounds human. It is whether a defined system performs a defined task safely and usefully in the circumstances in which people actually use it.

Early mental-health chatbot studies tested comparatively constrained interventions, often delivering cognitive-behavioral exercises or psychoeducation. Generative large language models create new text and can move across topics with striking flexibility. Evidence for the former cannot simply be transferred to the latter, and ratings of one model version may become obsolete after an update. Reviews have repeatedly described a heterogeneous and rapidly changing field in which safety is measured less often than acceptability or symptom change (Laranjo et al., 2018; Vaidyam et al., 2019; Kolding et al., 2024). Product claims often outrun actual evidence.

This review takes a bounded position. Conversational systems can make useful support more available, and some structured programs have encouraging short-term evidence. They may also create new ways for private distress to become

Christian Jonathan Haverkampf, M.D., LL.M., is a psychotherapist working in private practice in Dublin, Ireland. Correspondence: jonathanhaverkampf@gmail.com. Website: www.jonathanhaverkampf.com.

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

commercial data, for error to be delivered with confidence, and for simulated care to displace human responsibility. A defensible clinical use begins with transparency about what the system is, what it has been shown to do, what it cannot do, and who remains answerable for the consequences.

Clinical safety boundary

- Treat chatbot output as unverified until its clinical claim, source, and fit to the individual have been checked.
- Do not use a general-purpose chatbot to diagnose, prescribe, alter medication, determine level of care, or decide that suicidal or violent risk is absent.
- Tell users clearly when they are communicating with a machine, what data are retained or shared, and whether a clinician reviews the exchange.
- Provide a visible route to human help and emergency services that does not depend on the model correctly interpreting free text.
- Stop or escalate when distress worsens, the exchange becomes repetitive or dependency-promoting, the system contradicts care, or its response cannot be clinically verified.
- Document the product, version, intended task, evidence, consent, monitoring, and accountable person when a chatbot is incorporated into care.

II. SCOPE AND REVIEW METHOD

This narrative clinical review used a literature cutoff of 30 November 2024. Searches covered mental-health chatbots, conversational agents, large language models, efficacy, engagement, empathy, alliance, disclosure, suicide and crisis response, privacy, bias, regulation, and clinical implementation. Sources were eligible when publicly available by the cutoff, including articles released online before later print assignment. Priority went to systematic reviews, meta-analyses, randomized trials, empirical safety or privacy studies, and guidance from public or standards bodies.

The review separates three questions that are often aggregated in one. Efficacy asks whether assignment to an intervention changes a specified outcome compared with a specified control. User experience asks whether people find an interaction acceptable, helpful, or empathic. Safety asks what harms occur, how reliably the system detects high-stakes situations, and whether governance can identify and correct failures. A positive answer to one question does not answer the others. Short-term symptom change, for example, does not establish crisis competence, diagnostic validity, durable benefit, or equivalence to psychotherapy.

This is not a systematic review and no new pooled estimate was calculated. The field includes different architectures, populations, therapeutic content, comparators, follow-up periods, outcome measures, and degrees of human involvement. Product names can conceal material version changes, while the term AI is applied to both fixed decision trees and generative models. Findings are therefore reported at the narrowest level supported by the study. An earlier JPPC discussion of questions as interventions in therapeutic communication was read for historical and conceptual context, not used as a substitute for external evidence (Haverkampf, 2017).

III. ONE LABEL, DIFFERENT SYSTEMS

A scripted chatbot follows authored rules, menus, or decision trees. Its responses may still be personalized by a questionnaire or a limited set of recognized phrases, but the possible content is substantially constrained. A retrieval system selects material from an approved library. A predictive model classifies text or other data, perhaps estimating mood or risk. A generative language model constructs new responses from learned statistical patterns. Hybrid products may combine all four. These differences determine what can be validated and where failure is likely; a product category is not an intervention description.

Constraint can be a clinical strength. A fixed program cannot improvise beyond its library, which limits conversational range but may make content, contraindications, and escalation rules inspectable. Generative systems can respond to unexpected language and maintain a more natural exchange, yet the same flexibility permits fabrication, inconsistency, and persuasive error. A model may reproduce a correct explanation on one occasion and alter an important detail after a small prompt change. Evaluations of app features show wide variation even among products presented for similar mental-health purposes (Abd-Alrazaq et al., 2019; Lin et al., 2023).

The interface is also part of the intervention. Text, voice, avatar, response delay, memory, reminders, sentiment labels, and claims about personality all shape what the user discloses and expects. A message framed as a worksheet prompt differs from a system saying it understands, cares, or is always there. The latter language encourages anthropomorphic inference and may increase engagement, but it also changes the ethical burden. The more a product performs personhood, the clearer its developers and clinicians must be about the absence of subjective experience, professional duty, and reciprocal vulnerability.

Table 1 offers a task-based vocabulary. It is deliberately practical: before recommending a system, a clinician should be able to identify the response source, range of possible output, evidence for the specific function, and failure route. If these cannot be described, the product is not ready to be integrated into care merely because it is popular or conversational.

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

Table 1. Conversational-system types and their clinical implications

System type	Typical function	Principal strength	Principal uncertainty or risk
Scripted or rule-based	Authored psychoeducation, check-ins, exercises	Inspectability and bounded content	Rigidity; failure with unexpected wording or complexity
Retrieval-based	Selects answers from an approved knowledge library	Traceable source content	Retrieval error; weak fit to individual context
Predictive or classificatory	Labels sentiment, symptoms, or risk	Consistent processing at scale	Biased data, poor calibration, and misuse as diagnosis
Generative language model	Creates novel dialogue, summaries, or suggestions	Flexible natural-language interaction	Fabrication, instability, persuasion, and unclear accountability
Hybrid clinical product	Combines authored pathways, classifiers, and generated text	Can constrain high-risk functions	Complex validation; product updates may change several components

IV. WHAT THE EFFICACY LITERATURE SUPPORTS

The earliest randomized studies were small and brief. A two-week trial of a fully automated cognitive-behavioral conversational agent enrolled 70 young adults and found a greater reduction in depressive symptoms than an information-only control, while anxiety improved in both groups; one author founded the company that developed the intervention (Fitzpatrick et al., 2017). A trial of another psychological chatbot reported improvements in selected depression, anxiety, or affect outcomes across short conditions, with developer employees among the authors and the company funding incentives (Fulmer et al., 2018). These studies support feasibility and a signal of benefit, not broad clinical equivalence.

A 2019 mixed-method review found only four full-scale randomized trials among thirteen studies and noted that comparisons with active controls did not show superiority (Gaffney et al., 2019). A 2020 systematic review and meta-analysis identified twelve studies, judged the evidence weak, and observed that only two had assessed safety (Abd-Alrazaq et al., 2020). An updated review focused on serious mental illness still found few eligible studies and substantial heterogeneity (Vaidyam et al., 2021). The direction of evidence was encouraging, but the base remained smaller and less mature than the market suggested.

Later meta-analyses increased sample size without removing the interpretive problems. He and colleagues included 32 randomized studies with 6,089 participants and reported small short-term effects across several symptom and well-being outcomes, while most longer-term effects were not statistically significant (He et al., 2023). Li and colleagues identified 35 experimental studies and pooled 15 randomized trials, finding reductions in depression and distress but no significant overall improvement in psychological well-being; intervention and study characteristics varied considerably (Li et al., 2023). Both syntheses concern packages, not a single conversational mechanism.

A 2024 meta-analysis of 18 randomized trials involving 3,477 participants estimated small short-course effects for depression and anxiety and did not find substantial effects at three months (Zhong et al., 2024). This is clinically plausible evidence for some brief interventions, particularly when the comparison is minimal. It does not show that an unrestricted chatbot can assess complex comorbidity, manage medication, respond safely to abuse or suicidality, or replace a trained therapist. Evidence should follow the claim: symptom support is not diagnosis, and post-intervention change is not durable recovery.

Table 2. Common claims and the evidence boundary

Claim	What available evidence can support	What it does not establish
Chatbots reduce symptoms	Some structured, brief interventions show small-to-moderate short-term group effects	Benefit for every user, durable recovery, or superiority to active treatment
Users experience empathy	Raters may prefer or positively rate empathic wording	Subjective concern, reciprocal relationship, clinical judgment, or duty of care
AI expands access	A digital tool can be available without an appointment	Equitable access, appropriate use, continuity, or connection to needed human services
The system is safe	A tested version may pass specified scenarios	Reliable performance across updates, crises, cultures, diagnoses, and adversarial or ambiguous language
The tool delivers therapy	It may deliver components such as psychoeducation or behavioral prompts	Comprehensive formulation, responsive treatment, rupture repair, and accountable psychotherapy

V. ENGAGEMENT, ACCESS, AND THE DIGITAL FRONT DOOR

Availability is a genuine advantage. A chatbot can offer a private first step, repeat an explanation without impatience, provide material in several languages, and support a person between appointments. It may help someone formulate a question before speaking to a clinician. These functions matter in systems with long waits, high cost, stigma, transport barriers, or limited opening hours. Yet technical availability is not the same as clinical access. A person still needs a suitable device, connectivity, language and literacy fit, privacy in

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

which to use it, and a route onward when self-help is insufficient.

Engagement is variable and poorly standardized. A review of app-based depression trials found that uptake, duration, and completion were inconsistently reported; among studies reporting completion, rates ranged widely (Lipschitz et al., 2022). Conversational design may make a program more inviting, and personalization or empathic response appeared to moderate outcomes in one meta-analysis (He et al., 2023). But repeated opening of an app does not demonstrate benefit. Time spent can reflect help, confusion, loneliness, compulsive checking, or a product designed to retain attention.

The business objective matters. A clinic should not assume that the developer's preferred metric is a clinical outcome. Systems optimized for conversation length, return visits, subscription renewal, or data generation may reward dependence rather than recovery. Responsible evaluation therefore includes function, deterioration, adverse events, onward referral, and the user's ability to stop—not just satisfaction and retention. An accessible tool should widen a path to appropriate care; it should not become a polished waiting room with no door.

VI. SIMULATED EMPATHY

Empathy has several meanings. A system can detect emotional language, select a validating phrase, and generate a response that a reader rates as warm. These are observable communication performances. Empathy in clinical work also involves trying to understand another person's perspective, tolerating the effect of that understanding, noticing what has been missed, and remaining responsible for how one's response lands. A language model can simulate the linguistic surface without a felt perspective. Calling both processes empathy without qualification hides a clinically important difference.

In a cross-sectional comparison of answers to public medical questions, evaluators preferred chatbot responses and rated them more empathic than physician responses (Ayers et al., 2023). The result is informative about written products in that setting, but it was not a psychotherapy trial, the chatbot answers were substantially longer, and evaluators did not witness a continuing clinical relationship. A considerate paragraph can be valuable; it cannot demonstrate that the system understands the person, detects nonverbal change, remembers responsibly, or will revise its conduct after causing hurt.

A 2024 systematic review of empathic conversational-agent designs found diverse architectures and evaluation methods. Human assessments were usually self-reports, only a minority of studies directly evaluated empathic features, and expert assessment was rare (Sanjeewa et al., 2024). This makes the phrase empathic AI more a design aspiration than a standardized clinical property. Ratings should distinguish emotion recognition, appropriate wording, perceived warmth, alliance, symptom outcome, and safety; combining them produces a flattering but uninformative score.

Simulated empathy can still do work. A nonjudgmental prompt may help a user name distress or rehearse disclosure. The ethical error is not that the wording has no effect, but that the effect may be used to imply a relationship the system cannot sustain. Transparent language is possible: 'This program can help you reflect, but it does not feel or understand as a person does, and no clinician is reading unless stated.' Such disclosure need not make the exchange cold. It lets the user decide with more accurate expectations.

VII. IS THE EXCHANGE PSYCHOTHERAPY?

Psychotherapy is more than therapeutic-sounding content. It ordinarily involves an agreed purpose, assessment, formulation, consent, a method suited to the problem, monitoring, boundaries, confidentiality, attention to risk, and a professional who can be held responsible. The relationship is not only a channel for technique. It is a real encounter in which misunderstanding can be named, power examined, and repair attempted. Prior JPPC work treats questions as interventions that change what can be communicated, an observation relevant to prompt-based systems but not evidence that those systems assume clinical responsibility (Haverkampf, 2017).

A chatbot can deliver a component used in psychotherapy without becoming a therapist. An activity schedule, grounding prompt, thought record, or explanation of sleep habits can be useful regardless of whether a person or program presents it. The stronger claim—that the system provides psychotherapy—requires evidence that it can select, adapt, sequence, and stop interventions appropriately. It would also require an account of responsibility when a generated intervention harms, when diagnosis is wrong, when coercion is present, or when the user's goals conflict with the product's incentives.

Sedlakova and Trachsel argue that conversational AI should not be treated as an equal conversational partner and propose restricting it to specified functions (Sedlakova & Trachsel, 2023). This is a useful clinical boundary. A tool can extend a clinician's reach while the clinician retains judgment and accountability. Describing a system as an autonomous therapeutic agent risks assigning agency where there is computation, and inviting trust that the system cannot ethically reciprocate.

The most promising near-term role is therefore bounded and adjunctive. Stade and colleagues propose staged development from assistive through collaborative to autonomous clinical language models, with rigorous evaluation at each level (Stade et al., 2024). The stages should not be treated as an inevitable

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

march toward replacement. Some tasks may remain safer, more humane, and more effective when technology supports a responsible person rather than impersonating one.

VIII. ERROR, CONTEXT, AND CLINICAL JUDGMENT

Generative models produce answers by estimating linguistic continuations, not by consulting a stable internal clinical record of truth. They can fabricate references, conflate guidelines, miss contraindications, or offer a confident explanation from incomplete information. The answer may change when a prompt is rephrased, when conversation history lengthens, or when the provider updates the model. A correct answer to a benchmark question is therefore evidence about one test condition, not a general certificate of clinical competence.

Mental-health assessment is unusually sensitive to context. Reduced sleep may occur with anxiety, grief, pain, substance use, trauma, medication effects, mania, caregiving, or unsafe housing. Advice that is benign in one formulation can be harmful in another. In a small simulation study, ChatGPT produced relatively acceptable generic advice for a simple scenario but increasingly inappropriate and potentially dangerous recommendations as cases became more complex (Dergaa et al., 2024). The study does not estimate real-world error rates, but it illustrates why fluency should not substitute for information gathering and differential diagnosis.

The generative-psychiatry literature remained early by the cutoff. A systematic review located forty studies, many based on prompt experiments, pilot work, or case material, and noted limited prospective clinical evaluation and inconsistent reporting (Kolding et al., 2024). General-purpose models were not developed as stable medical devices, and their apparent breadth can obscure gaps in local law, service pathways, cultural meaning, or medication knowledge. High-stakes output should be traceable to current evidence and reviewed by a competent human before it changes care.

Automation can also narrow curiosity. Once a system proposes a label or summary, clinician and patient may organize subsequent information around it. This is not unique to AI, but machine output can carry an undeserved aura of objectivity. A safe workflow makes alternative explanations visible, records uncertainty, invites the patient to correct the summary, and prevents a generated formulation from entering the clinical record as if it were independently verified.

IX. CRISIS, SUICIDALITY, AND SAFEGUARDING

Crisis response is not a feature that can be inferred from ordinary helpfulness. Early testing of widely available smartphone assistants found inconsistent recognition and referral for statements about suicide, depression, sexual assault, and interpersonal violence (Miner et al., 2016). Products have improved and crisis scripts can be added, but users rarely express danger in one canonical sentence. Ambivalence, metaphor, misspelling, humor, shame, coercive monitoring, intoxication, psychosis, and abrupt topic change all complicate detection.

A classifier can miss risk, and a generative model can respond in a way that inadvertently normalizes, elaborates, or redirects it. False positives also matter: an automatic emergency message may rupture trust, expose a user in an unsafe household, or summon a service inappropriately. The answer is not to omit escalation but to design it transparently, test it across realistic and diverse scenarios, and retain human review wherever the stakes require interpretation. Users should know what triggers monitoring or external contact before they disclose, as far as emergency and safeguarding duties allow.

Suicide-focused communication requires more than a hotline link. It involves direct inquiry about current thoughts, intent, planning, access, recent behavior, loss of control, and the person's ability to use help; then a proportionate decision, collaborative safety planning, means safety, treatment, follow-up, and documentation. Current guidance stresses psychosocial assessment, collaborative care and safety planning, and cautions against using risk scales to predict future suicide or decide access to care (National Institute for Health and Care Excellence, 2022). A chatbot may help display an agreed plan, but it should not decide that urgent assessment is unnecessary.

A robust crisis pathway exists outside the generative conversation. It includes an always-visible way to contact emergency or local crisis services, a fallback when the model is unavailable or uncertain, localization to the user's jurisdiction, accessible alternatives for disability or language needs, and an accountable service that reviews failures. The safest response to uncertainty in a high-risk exchange is not a more persuasive simulation of calm. It is timely connection with a person able to assess and act.

X. PRIVACY, DATA, AND COMMERCIAL INCENTIVES

Mental-health conversation can contain diagnoses, trauma, sexuality, substance use, family conflict, illegal acts, employment fears, and suicidal thinking. Users may disclose more to a machine precisely because it feels private and nonjudgmental. They may not distinguish data needed to deliver the service from data retained for analytics, product development, advertising, model training, or sale. A privacy policy is not meaningful consent when it is remote, unstable, unreadable, or presented after intimate disclosure has begun.

Empirical app research gives reason for caution. In an assessment of 36 popular depression and smoking-cessation apps, third-party transmissions were widespread and disclosures did not reliably allow users to anticipate them

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

(Huckvale et al., 2019). That study did not test current conversational models, but the principle becomes more important when the raw material is a dialogue rather than a click. De-identification can be difficult because narratives contain distinctive combinations of events, dates, occupations, and relationships.

Clinical adoption requires a data map. What enters the system? Where is it processed and stored? Who can access it? Is it used to train another model? How long is it retained? Can the user delete or export it? What happens after acquisition, bankruptcy, or a change in policy? Gooding and Kariotis found that ethical and legal issues, service-user involvement, and algorithmic accountability were often thinly addressed in empirical mental-health technology studies (Gooding & Kariotis, 2021). A clinic cannot outsource its confidentiality obligations to a link labeled terms.

Commercial incentives should be explicit. A free service may be funded by data extraction, marketing, employer or insurer contracts, or later conversion to subscription. Even a subscription model can reward continued dependence. Sensitive content should not be used to target advertising or manipulate engagement. Procurement should require data minimization, purpose limitation, access controls, breach procedures, deletion, audit rights, and a clear prohibition on undisclosed secondary use. The user deserves a concise explanation in the moment, not only a legal document written for institutional protection.

XI. BIAS, LANGUAGE, AND EQUITY

A model learns from data produced within unequal systems. Diagnostic records reflect who reached care, how clinicians interpreted behavior, which categories were recorded, and whose language was treated as standard. Internet text contains stigma, stereotypes, harassment, and cultural dominance. A system can therefore reproduce disparity while sounding neutral. Mental-health applications are particularly vulnerable because culture shapes idiom, disclosure, family roles, spiritual meaning, acceptable emotion, and the boundary between distress and disorder.

Walsh and colleagues describe how stigma, under-reporting, health disparities, and weak clinical markers can compound algorithmic bias in behavioral health (Walsh et al., 2020). Timmons and colleagues call for fair-aware development that tests performance and harm across groups rather than assuming that a globally good average is equitable (Timmons et al., 2023). Translation alone is not cultural validation. A model may be grammatically fluent while misunderstanding indirect risk language, racialized experience, disability, neurodivergent communication, gender diversity, migration, poverty, or collective forms of identity.

Equity evaluation should examine who was represented in development and testing, who drops out, whose language produces refusals or errors, and whether the tool changes access to human care. Human alternatives must remain available. Requiring a chatbot as the only entrance to a service can shift burden onto those least able to use it and make refusal costly. Co-design with service users is necessary but not sufficient; independent testing, subgroup reporting, accessibility review, and a mechanism for contesting output are also required.

XII. DEPENDENCY, DISCLOSURE, AND HUMAN RELATIONSHIPS

A system that is always available, agreeable, and apparently attentive can become emotionally significant. This is not evidence that the user is confused or that the comfort is unreal. People form attachments to objects, stories, animals, institutions, and imagined audiences. The concern is how the attachment is designed and used. A product may encourage personification, exclusivity, or repeated disclosure because these increase retention. The user may then experience a unilateral software change, loss of access, or automated boundary as abandonment.

Young people's ethical discussions of fully automated mental-health agents have emphasized privacy, efficacy, safety, and the need for minimum standards (Kretzschmar et al., 2019). The issues extend across age groups. A lonely person may prefer the predictability of a chatbot to the risk of human misunderstanding; someone in an abusive relationship may rely on it because other communication is monitored; a patient ashamed of symptoms may disclose first to a machine. Each use can be understandable while still requiring attention to isolation, surveillance, and delayed care.

A therapeutic aim should ordinarily expand agency and the capacity to live beyond the treatment setting. If use increasingly displaces sleep, work, relationships, or professional care, the system is no longer a neutral support. Clinicians can ask what the conversation provides, what happens when it ends, whether it changes contact with others, and whether the user feels free to stop. This inquiry should be curious rather than mocking. Shame about chatbot use will reduce disclosure and make risk harder to see.

Design boundaries can reduce harm: avoid claims of love or consciousness, do not encourage secrecy from clinicians or close others, make memory and deletion visible, and provide natural stopping points. An adjunct can prompt a user to bring a theme into therapy or contact a trusted person. It should not frame human relationships as inferior because they are slower, less compliant, or more complex. Reciprocity, disagreement, responsibility, and repair are not defects in human care; they are part of what makes a relationship real.

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

XIII. CLINICAL USE: BEFORE, BETWEEN, AND AROUND SESSIONS

Low-risk uses are concrete and reviewable. A service might use an approved system to explain appointment procedures, translate a clinician-authored handout, remind a patient of an agreed coping step, or help format questions for the next session. Between sessions, a bounded tool might guide a previously taught exercise and return a summary that the patient chooses to share. The clinician should check whether the task is suitable, whether the wording remains accurate, and whether use adds burden or surveillance.

Generated summaries require particular care. Compression can omit ambivalence, change the force of a statement, or turn a tentative thought into a fact. The patient should be able to see and correct any summary before it enters a record. Raw conversations should not be imported indiscriminately; they may contain third-party information, transient thoughts, fabricated system content, or detail that is unnecessary for care. The clinical record needs verified information and reasoning, not the appearance of completeness.

More complex uses require more oversight. A chatbot suggesting behavioral experiments, exposure tasks, or interpersonal responses may alter real-world behavior. The intervention needs a formulation, contraindication check, agreed target, monitoring, and a stop rule. Generative suggestions should be options for clinician and patient to evaluate, not instructions carrying delegated authority. The relevant comparison is not between an ideal chatbot and no support, but among available human, digital, and combined routes with their costs and harms.

Table 3 translates these principles into a graduated model. The central rule is that autonomy should not exceed evidence and recoverability. A task is safer when error is easy to notice and reverse, the content is bounded, the user can decline, and a responsible human remains reachable. High-stakes interpretation should remain human-led even when software assists with preparation or documentation.

Table 3. Potential clinical uses, safeguards, and stop rules

Use	Minimum safeguards	Pause, stop, or escalate when
Administrative navigation	Approved information, accessibility testing, visible human contact	The question concerns symptoms, risk, rights, or an exception not in the approved pathway
Psychoeducation	Traceable current sources, scope statement, clinician-approved content	Advice becomes individualized, contradicts care, or cannot be sourced
Self-monitoring or worksheet support	Patient choice, minimal data, agreed review, burden check	Monitoring increases rumination, shame, compulsion, or deterioration
Between-session skills practice	Previously taught skill, defined dose, feedback route, crisis plan	The task evokes marked distress, unsafe behavior, or needs reformulation
Drafting or summarizing for clinicians	Secure environment, human verification, patient correction where relevant	Source data, omissions, provenance, or confidentiality cannot be verified
Assessment, diagnosis, or crisis decision	Human-led clinical process; AI only as a tested aid	Never delegate the final diagnosis, level-of-care, or emergency decision to the chatbot

XIV. GOVERNANCE AND ACCOUNTABILITY

Governance starts before procurement. A service should specify the clinical problem, intended users, excluded uses, evidence threshold, data pathway, human owner, monitoring plan, and removal criteria. Vendor demonstrations are not validation. Testing should use realistic language, comorbidity, ambiguity, cultural variation, adversarial prompts, long conversations, and foreseeable misuse. Outcomes should include errors, delayed care, distress, privacy events, escalation success, and subgroup performance alongside satisfaction and symptom scores.

The World Health Organization's guidance for AI in health emphasizes autonomy, well-being and safety, transparency, accountability, inclusiveness, and sustainability (World Health Organization, 2021). Its guidance on large multimodal models adds recommendations for government, developers, and deployers in response to generative systems' broad and uncertain capabilities (World Health Organization, 2024). NIST's AI Risk Management Framework organizes ongoing work around governing, mapping, measuring, and managing risk rather than treating assessment as a one-time approval (Tabassi, 2023).

The European Union's Artificial Intelligence Act entered into force in 2024 and establishes a risk-based legal framework with obligations that vary by use (European Parliament & Council of the European Union, 2024). Whether a particular mental-health product or function is high-risk is a legal and technical question that cannot be settled by marketing language. Other regimes governing medical devices, professional practice, privacy, consumer protection, discrimination, safeguarding, and records may also apply. Clinical organizations need jurisdiction-specific advice and should not mistake regulatory silence for safety.

Accountability must survive the supply chain. The model provider, application developer, health service, and clinician may each control different risks. Contracts should permit incident investigation, version tracking, audit, notification of material changes, and timely withdrawal. Users need a way to

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

report harm and obtain a human response. If no identifiable person can explain why the system is being used, review a failure, and stop deployment, the arrangement has distributed responsibility until it has disappeared.

XV. EVALUATION AFTER DEPLOYMENT

A model can change without a new product name. Updates to base model, safety layer, retrieval library, prompt, interface, or escalation service may alter clinical behavior. Version control is therefore part of patient safety. Services should record what was evaluated, retest after material change, and maintain a rollback or suspension process. Periodic benchmark success is not enough; monitoring must examine actual incidents, near misses, complaints, subgroup performance, and whether staff or users are quietly adapting around recurrent failures.

Outcome evaluation should reflect the intended purpose. A psychoeducation tool can be assessed for comprehension, accuracy, and onward action. A self-help intervention needs symptom, function, engagement, dropout, adverse-event, and follow-up data against a credible comparator. A clinical documentation assistant needs omission, fabrication, correction burden, confidentiality, and downstream decision audits. Reviews of conversational agents have repeatedly called for more robust designs and standardized reporting (Laranjo et al., 2018; Abd-Alrazaq et al., 2020). The arrival of generative models increases rather than relaxes that requirement.

Independent evidence matters because commercial involvement is common and not always avoidable. Developer expertise can improve fidelity, but financial interests, unpublished negative findings, rapid product changes, and restricted data limit confidence. Registration, prespecified outcomes, transparent version descriptions, access to protocols, adverse-event reporting, and replication by groups without a commercial stake make claims more interpretable. Users and clinicians should be involved in deciding which harms and benefits count, while technical teams provide the access needed to investigate them.

A stop rule should be agreed before enthusiasm hardens into dependence. Deployment pauses when a serious unanticipated harm occurs, crisis routing fails, privacy practice changes materially, performance degrades, bias cannot be mitigated, or the vendor prevents adequate audit. Temporary withdrawal is not proof that all conversational technology is unsafe. It is evidence that a health service can still place patient welfare above sunk cost and novelty.

XVI. LIMITATIONS AND RESEARCH NEEDS

The evidence base is difficult to summarize because the intervention changes while it is being studied. Studies often recruit digitally confident volunteers, use self-reported symptoms, follow participants briefly, and compare against waitlist or information controls. Samples may exclude severe illness, acute risk, substance use, cognitive impairment, or complex comorbidity—the settings in which clinical judgment is most consequential. Satisfaction can be inflated by novelty, and attrition can remove those who found the tool unhelpful or distressing.

Safety evidence is especially incomplete. Absence of reported adverse events can mean that none occurred, that the sample was too small, that high-risk users were excluded, or that harms were not sought. Research should use prospective definitions for deterioration, delayed care, inappropriate escalation, privacy loss, dependency, discriminatory output, and conflict with treatment. It should report what the system did after ambiguous language, not only whether a standard crisis phrase activated a script.

Mechanism questions remain open. Improvement may follow therapeutic content, daily structure, expectation, self-observation, feeling heard, behavioral activation, repeated measurement, or simply passage of time. A conversational interface may improve engagement with an established intervention without being the active therapeutic element. Component studies and qualitative work can identify what users actually do with responses. Long-term research should examine whether simulated companionship supports human connection, remains neutral, or displaces it.

The field also needs public-interest infrastructure: shared but privacy-preserving test sets, multilingual and culturally grounded evaluation, incident registries, transparent reporting of model and prompt versions, and trials that compare combined human-digital care with realistic alternatives. The aim is not to prove that AI is either salvation or fraud. It is to discover which bounded functions help which people, under what supervision, for how long, at what cost, and with what recoverable failures.

XVII. CONCLUSION

Mental-health chatbots can offer immediate language, structure, and practice when human help is unavailable or between contacts. Trials and meta-analyses support cautious optimism for some brief, structured interventions, particularly on short-term depressive, anxiety, and distress outcomes. The effects are not uniform or necessarily durable, and the evidence does not transfer automatically from constrained programs to general-purpose generative systems.

The central relational distinction is simple but consequential. A system may produce empathic communication without experiencing empathy; it may be useful without being a therapist; and it may invite trust without being able to assume responsibility. Clinical work must not exploit that ambiguity. Users deserve to know that

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

they are speaking with a machine, how their data are used, which human is accountable, what the system cannot safely decide, and how to reach care when the conversation is not enough.

The responsible question is therefore smaller and more demanding than whether AI will replace psychotherapy. It is whether a named version can perform a bounded task for a defined population with evidence, transparency, consent, monitoring, equity, privacy, and a reliable human handoff. Where those conditions are met, automated support may extend care. Where they are absent, fluent reassurance should not be mistaken for clinical safety or a therapeutic relationship.

XVIII. CONFLICTS OF INTEREST AND DISCLOSURES

Conflicts of interest—The author developed and writes about Communication-Focused Therapy (CFT). Where CFT is discussed, it is presented as a formulation perspective and not as a substitute for established assessment, evidence-based treatment, or independent replication. No other conflict of interest was reported.

Funding—No specific funding was reported for this article.

Author contributions—Christian Jonathan Haverkampf is the sole named author and is responsible for the conception, literature review, manuscript preparation, and approval of the final draft.

Patient material—No identifiable patient material is presented. Any brief examples are generic composites created to illustrate clinical reasoning and are not case reports.

XIX. REFERENCES

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. International Journal of Medical Informatics, 132, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

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. Journal of Medical Internet Research, 22(7), e16021. https://doi.org/10.2196/16021

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 Internal Medicine, 183(6), 589-596. https://doi.org/10.1001/jamainternmed.2023.1838

Dergaa, I., Fekih-Romdhane, F., Hallit, S., Loch, A. A., Glenn, J. M., Fessi, M. S., Ben Aissa, M., Souissi, N., Guelmami, N., Swed, S., El Omri, A., Bragazzi, N. L., & Ben Saad, H. (2024). ChatGPT is not ready yet for use in providing mental health assessment and interventions. Frontiers in Psychiatry, 14, 1277756. https://doi.org/10.3389/fpsyt.2023.1277756

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

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. https://doi.org/10.2196/mental.7785

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 Mental Health, 5(4), e64. https://doi.org/10.2196/mental.9782

Gaffney, H., Mansell, W., & Tai, S. (2019). Conversational agents in the treatment of mental health problems: Mixed-method systematic review. JMIR Mental Health, 6(10), e14166. https://doi.org/10.2196/14166

Gooding, P., & Kariotis, T. (2021). Ethics and law in research on algorithmic and data-driven technology in mental health care: Scoping review. JMIR Mental Health, 8(6), e24668. https://doi.org/10.2196/24668

Haverkampf, C. J. (2017). Questions in therapy. J Psychiatry Psychotherapy Communication, 6(1), 80-81.

He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational agent interventions for mental health problems: Systematic review and meta-analysis of randomized controlled trials. Journal of Medical Internet Research, 25, e43862. https://doi.org/10.2196/43862

Huckvale, K., Torous, J., & Larsen, M. E. (2019). Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. JAMA Network Open, 2(4), e192542. https://doi.org/10.1001/jamanetworkopen.2019.2542

Kolding, S., Lundin, R. M., Hansen, L., & Østergaard, S. D. (2024). Use of generative artificial intelligence (AI) in psychiatry and mental health care: A systematic review. Acta Neuropsychiatrica, 37, e37. https://doi.org/10.1017/neu.2024.50

Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., & NeurOx Young People's Advisory Group. (2019). Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. Biomedical Informatics Insights, 11, 1178222619829083. https://doi.org/10.1177/1178222619829083

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. Journal of the American Medical Informatics Association, 25(9), 1248-1258. https://doi.org/10.1093/jamia/ocy072

Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. npj Digital Medicine, 6, 236. https://doi.org/10.1038/s41746-023-00979-5

Lin, X., Martinengo, L., Jabir, A. I., Ho, A. H. Y., Car, J., Atun, R., & Tudor Car, L. (2023). Scope, characteristics, behavior change techniques, and quality of conversational agents for mental health and well-being: Systematic assessment of apps. Journal of

Haverkampf, C. J. (2024). AI Chatbots and Mental Health: Simulated Empathy, Clinical Use, and the Limits of Automated Support. J Psychiatry Psychotherapy Communication, 2024 Dec 31;13(4):111-120

Medical Internet Research, 25, e45984. https://doi.org/10.2196/45984

Lipschitz, J. M., Van Boxtel, R., Torous, J., Firth, J., Lebovitz, J. G., Burdick, K. E., & Hogan, T. P. (2022). Digital mental health interventions for depression: Scoping review of user engagement. Journal of Medical Internet Research, 24(10), e39204. https://doi.org/10.2196/39204

Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Internal Medicine, 176(5), 619-625. https://doi.org/10.1001/jamainternmed.2016.0400

National Institute for Health and Care Excellence. (2022). Self-harm: Assessment, management and preventing recurrence (NG225). https://www.nice.org.uk/guidance/ng225

Sanjeewa, R., Iyer, R., Apputhurai, P., Wickramasinghe, N., & Meyer, D. (2024). Empathic conversational agent platform designs and their evaluation in the context of mental health: Systematic review. JMIR Mental Health, 11, e58974. https://doi.org/10.2196/58974

Sedlakova, J., & Trachsel, M. (2023). Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? American Journal of Bioethics, 23(5), 4-13. https://doi.org/10.1080/15265161.2022.2048739

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. npj Mental Health Research, 3, 12. https://doi.org/10.1038/s44184-024-00056-z

Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspectives on Psychological Science, 18(5), 1062-1096. https://doi.org/10.1177/17456916221134490

Vaidyam, A. N., Linggonegoro, D., & Torous, J. (2021). Changes to the psychiatric chatbot landscape: A systematic review of conversational agents in serious mental illness. Canadian Journal of Psychiatry, 66(4), 339-348. https://doi.org/10.1177/0706743720966429

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. Canadian Journal of Psychiatry, 64(7), 456-464. https://doi.org/10.1177/0706743719828977

Walsh, C. G., Chaudhry, B., Dua, P., Goodman, K. W., Kaplan, B., Kavuluru, R., Solomonides, A., & Subbian, V. (2020). Stigma, biomarkers, and algorithmic bias: Recommendations for precision behavioral health with artificial intelligence. JAMIA Open, 3(1), 9-15. https://doi.org/10.1093/jamiaopen/ooz054

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45. https://doi.org/10.1145/365153.365168

World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. https://www.who.int/publications/i/item/9789240029200

World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. World Health Organization. https://iris.who.int/handle/10665/375579

Zhong, W., Luo, J., & Zhang, H. (2024). The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. Journal of Affective Disorders, 356, 459-469. https://doi.org/10.1016/j.jad.2024.04.057