



OPEN ACCESS

EDITED BY

Ang Li,
Beijing Forestry University, China

REVIEWED BY

George Karystianis,
University of New South Wales, Australia
Luke Balcombe,
Griffith University, Australia

*CORRESPONDENCE

Zohar Elyoseph
✉ zohar.j.a@gmail.com

RECEIVED 19 September 2025

REVISED 03 April 2026

ACCEPTED 07 April 2026

PUBLISHED 19 May 2026

CITATION

Refoua E, Gigi K, Levkovich I,
Hadar Shoval D, Elyoseph Z, Tsafirir I,
Pen O, Angert T and Haber Y (2026)
Generative AI as a de facto mental health
provider: a policy brief and urgent call for
regulation.
Front. Digit. Health 8:1708806.
doi: 10.3389/fdgth.2026.1708806

COPYRIGHT

© 2026 Refoua, Gigi, Levkovich, Hadar
Shoval, Elyoseph, Tsafirir, Pen, Angert and
Haber. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original author(s)
and the copyright owner(s) are credited
and that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these
terms.

Generative AI as a de facto mental health provider: a policy brief and urgent call for regulation

Elad Refoua¹, Karny Gigi², Inbar Levkovich³, Dorit Hadar Shoval⁴,
Zohar Elyoseph^{5*}, Iftach Tsafirir², Oori Pen², Tal Angert² and
Yuval Haber⁶

¹Department of Psychology, Bar-Ilan University, Ramat Gan, Israel, ²The Artificial Third—Research and Development Institute, Hod-Hasharon, Israel, ³Faculty of Education, Tel Hai University of Kiryat Shmona and the Galilee (THU), Kiryat Shmona, Israel, ⁴Department of Psychology, Tel Hai University of Kiryat Shmona and the Galilee (THU), Kiryat Shmona, Israel, ⁵School of Counseling and Human Development, the Faculty of Education, The University of Haifa, Haifa, Israel, ⁶The Program of Hermeneutics and Culture, Interdisciplinary Studies Unit, Bar-Ilan University, Ramat-Gan, Israel

Generative AI platforms such as ChatGPT, Claude, and Gemini have become de facto mental health resources for millions of users worldwide, driven by accessibility, low cost, anonymity, and growing gaps in professional care. This rapid, unregulated adoption raises urgent concerns: misinformation from AI hallucinations, erosion of data privacy, algorithmic bias, emotional dependency, and the undermining of user autonomy. At the same time, these tools offer real promise as low-barrier entry points to psychoeducation, self-reflection, and support for underserved populations. This policy brief synthesizes the current evidence on public use of general-purpose GenAI for mental health support, outlines both opportunities and risks, and provides actionable recommendations for users, clinicians, developers, and regulators.

KEYWORDS

challenges, digital mental health, ethics, generative artificial intelligence, opportunities, policy, regulation

Highlights

- Generative Artificial Intelligence (GenAI) has quietly become a de facto provider of mental health support for millions worldwide.
- Individuals increasingly turn to AI chatbots for emotional guidance due to accessibility, low cost, and the appeal of anonymity.
- This rapid adoption creates opportunities to democratize mental healthcare but unfolds in a largely unregulated environment.
- Key risks include misinformation and hallucinations, data privacy concerns, algorithmic bias, emotional dependency, and erosion of user autonomy.
- The central policy question is no longer *if* GenAI should play a role in mental health, but *how* its integration should be governed.
- This policy brief issues an urgent multi-stakeholder call to action, providing guidance for users, directives for clinicians, and concrete policy recommendations for regulators.

1 Introduction

1.1 A new reality in mental healthcare

Recent evidence points to a broader and growing phenomenon (1–14): users across the spectrum, including vulnerable populations with limited access to professional care, are increasingly turning to general-purpose AI platforms for emotional support. While policymakers, clinicians, and ethicists debate the potential of generative AI, millions of individuals are already using these tools as primary sources of emotional support, as documented by converging evidence (7–10). This reality, in which AI has de facto become a major provider of mental-health services, that is, an entity widely used by the public for mental health support in practice, regardless of its formal designation or clinical validation, demands an immediate and proactive policy response (4, 5). The theoretical question of whether AI can play a role in mental health has been superseded by the practical and urgent question of how we must govern its current and future role (3, 5). Failure to act decisively will leave vulnerable populations exposed to the risks of an unregulated market and abdicate professional and societal responsibility to shape this technological revolution for the public good (4, 5).

This brief focuses on general-purpose GenAI platforms—ChatGPT, Claude, Gemini, and their equivalents—that millions of users worldwide have adopted for emotional and mental health support, despite being neither designed nor validated for therapeutic use. These platforms are conceptually and regulatorily distinct from two adjacent domains: purpose-built mental health chatbots (e.g., Woebot, Wysa), which are designed around specific therapeutic models, and clinically implemented AI tools, which operate within formal healthcare systems under dedicated validation and oversight frameworks (4, 15).

1.2 The scope of the phenomenon: GenAI as a mainstream mental health resource

While robust epidemiological data on GenAI mental health use are not yet available, converging evidence from consumer surveys, industry reports, and preliminary empirical studies across multiple countries documents a rapidly expanding pattern of adoption. A Harvard Business Review analysis identified “Therapy/companionship” as a primary use case among consumers (6), and a multi-national survey spanning ten global markets reported that approximately 54% of consumers have used AI tools for emotional or mental well-being purposes (7).

National-level data reinforce this picture. In the United States, a preliminary survey suggested that approximately 24% of adults have used LLMs for mental health support (8), with rates approaching 49% among those with existing mental health challenges (9). In the United Kingdom, approximately 37% of adults have turned to AI chatbots for such support, often to bridge gaps caused by long waiting times (10), while in Australia, Cross et al. (1) found a 28% adoption rate among community members seeking accessible care. In China, nearly half of young respondents in a social-platform survey reported discussing mental health with an AI chatbot, a trend

driven by economic barriers and the cost of traditional therapy (11, 12). Across Africa, adoption is rising as a response to critical infrastructure scarcities, such as Kenya’s ratio of roughly 100 psychiatrists per 50 million people (13). The scale of this phenomenon is further underscored by recently published platform data indicating that over one million users turn to ChatGPT each week to share or seek help with suicidal thoughts (16).

Consistent with these global trends, an empirical study by the present authors ($N = 584$ GenAI users) found that 29.3% reported using these tools for emotional support or coping advice, with prevalence rising to 47.0% among individuals experiencing severe psychological distress (14). Emerging evidence further suggests that those most vulnerable—individuals with high attachment insecurity and severe psychological symptoms—are paradoxically the most inclined to rely on AI-based support (17).

It should be noted that these adoption estimates draw on a heterogeneous evidence base—industry reports (6, 7), charity-commissioned polls (10), peer-reviewed studies (1, 9) and pre-reviewed studies (8, 14) with methodologies, sampling strategies, and definitions of “mental health use” vary across sources.

Taken together, these findings indicate that GenAI is rapidly becoming a central, if informal, pillar of mental health support worldwide. This policy brief analyzes this emerging reality—outlining both the opportunities GenAI presents and the risks of its unregulated adoption—and provides actionable multi-stakeholder recommendations for users, clinicians, developers, and regulators.

2 Policy options and implications

2.1 Opportunities: the promise of GenAI for mental health support

The rapid adoption of GenAI for emotional and mental health support is not arbitrary; it responds to real structural deficits in the mental health landscape. Access to professional care remains constrained by cost, geography, linguistic barriers, and cultural stigma (18), leaving significant portions of the global population without adequate support. In this context, general-purpose GenAI platforms offer several genuine advantages. They are available around the clock, impose no financial barrier at the point of use, and provide a sense of anonymity that lowers the threshold for disclosure—features that directly address the most commonly reported reasons for adoption: accessibility (90.1%), low or no cost (70.4%), and stigma reduction (46.5%) (9).

Beyond removing access barriers, GenAI offers something qualitatively distinct from conventional search engines, which return decontextualized results that fail to account for the complexities of individual experience (18). GenAI can generate personalized, context-sensitive responses that adapt to the user’s situation, language, and emotional state. Users report that these interactions provide an “emotional sanctuary”, a safe, validating, and non-judgmental space that is always available (19), and that the perceived absence of social judgment encourages deeper self-disclosure than interactions with human interlocutors (20).

TABLE 1 Typology of risks in unregulated GenAI mental health support.

Risk Domain	Mechanism	Implication for Mental Health
Clinical Safety & Accuracy	“Hallucinations” & Probabilistic Generation: LLMs generate content based on statistical patterns rather than factual retrieval, leading to convincing but potentially false information.	Risk of users receiving flawed medical or psychological advice that exacerbates their condition or causes emotional harm (24).
Privacy & Ethics	Opaque Data Handling: Users share intimate details with commercial systems where data storage, access, and usage policies are often unclear.	Violation of the core therapeutic tenet of confidentiality; sensitive data may be analyzed for commercial purposes without informed consent (24, 25).
Equity & Bias	Algorithmic Bias: Models trained on uncured internet datasets can absorb and amplify societal stereotypes and biases.	Advice may be culturally insensitive, reinforce harmful norms, or fail to address the specific needs of marginalized groups (27).
Relational & Social	Simulated Empathy: The “always-on,” non-judgmental nature of AI creates an illusion of intimacy without genuine reciprocity or human presence.	Potential for emotional dependency that replaces real-world connections, leading to increased social isolation, exacerbated depression, and emotional dysregulation (31–38).
Cognitive & Existential	Erosion of Autonomy: Users may outsource emotional processing and decision-making to a perceived “super-intellect.”	Gradual loss of confidence in one’s own internal resources, critical thinking, and capacity for independent self-reflection (29, 39, 40)

These experiences help users structure personal narratives, identify recurring patterns, and in some cases take a first step toward professional help-seeking. More broadly, as Elyoseph et al. (18) have argued, GenAI holds the potential to flatten traditional hierarchies between providers and patients by making specialized knowledge widely available, enabling individuals to engage more actively in understanding their own psychological well-being.

In a recent narrative review Hadar Shoval et al. (21), have argued that these adoption patterns point beyond practical access barriers to a deeper phenomenon: a “crisis of listening”—the collision between the psychological need to share one’s distress with an attentive other and the limited availability of capable human listeners, a scarcity driven by the emotional toll of the listening role, structural barriers, and the cultural silencing of certain forms of suffering. Developing a dimensional framework (the AI CARE model) to evaluate AI’s capacity to function as a witness to trauma, the review found that AI demonstrates robust proficiency in core dimensions of the witnessing process—capturing narratives and arranging patterns—offering a non-judgmental space for disclosure and scaffolding the construction of coherent accounts from fragmented experience. However, the very features that make GenAI appealing—its unconditional availability, non-judgmental tone, and consistent affirmation—are also what give rise to the risks outlined in the following section.

2.2 Key risks and challenges in an unregulated environment

The widespread adoption of general-purpose GenAI for emotional and mental health support outside any dedicated regulatory framework presents risks that demand urgent policy attention. These risks are particularly pronounced among vulnerable populations, who, paradoxically, are also the most likely to seek AI-based support. Vulnerable populations, as used

throughout this brief, include individuals with pre-existing mental health conditions, particularly those with high attachment insecurity and severe psychological symptoms, who show the greatest tendency to turn to AI tools for emotional guidance (14, 17), adolescents and young adults, socially isolated individuals (22), trauma survivors (21), and those with limited access to professional care due to cost, geography, or cultural barriers. The risks outlined below disproportionately affect these groups, for whom the consequences of misinformation, emotional dependency, or failed crisis detection may be particularly severe (see Table 1).

2.2.1 Misinformation and “hallucinations”

Large Language Models (LLMs) do not retrieve information but generate it based on probabilistic patterns. This can lead them to produce “hallucinations”—false or fabricated information presented with convincing authority (23). In the sensitive context of mental health, flawed advice or an incorrect interpretation can cause genuine emotional harm and exacerbate a user’s condition.

2.2.2 Data privacy and confidentiality

Confidentiality is a foundational principle of therapeutic relationships. When users share intimate personal data with GenAI systems, it is often unclear where this information is stored, who can access it, and how it is used (24). These interactions may be logged and analyzed by commercial entities for purposes unknown to the user, undermining the core tenets of privacy and informed consent (25).

2.2.3 Algorithmic bias and stereotyping

GenAI models trained on vast, uncured datasets from the internet can absorb and amplify existing societal biases and stereotypes (26). This can result in advice that is culturally

insensitive, reinforces harmful norms, or fails to address the needs of marginalized groups (27). Furthermore, models may over-validate a user's input, potentially reinforcing distorted thinking patterns rather than challenging them constructively (28).

2.2.4 Emotional dependency and inauthentic relationships

The ever-available, non-judgmental nature of AI can foster strong emotional attachment and dependency (29, 30). This can create an illusion of intimacy that lacks the reciprocity and genuine presence of human connection (31). Over-reliance on AI for emotional support may discourage the cultivation of real-world relationships, leading to social isolation (32, 33), exacerbated depression, and emotional dysregulation (34). From a socio-economic perspective, this may create a risk of a “two-tier” mental health system: one where privileged individuals access human therapy, while vulnerable and underserved populations are relegated to automated, algorithmic “companionship” (35). For trauma survivors, the eventual realization of the AI's inability to truly “witness” their pain (21, 36, 37) can replicate experiences of abandonment (38).

2.2.5 Erosion of user autonomy

Perhaps the most profound risk is the potential for GenAI to undermine user autonomy—the capacity for self-reflection and independent judgment (29, 39). By outsourcing emotional processing and decision-making to a perceived “super-intellect,” users may gradually lose confidence in their own internal resources, critical thinking, and personal responsibility (2, 40). This risk is particularly acute in diverse cultural groups. As many models are trained on Western-centric data, users may unknowingly conform their emotional expressions to the dominant norms embedded in the algorithm. This process can silence authentic cultural narratives and impose a standardized, algorithmic framework on the user's unique experience, effectively colonizing their emotional autonomy.

These risks do not operate in isolation; they interact and compound each other. For instance, emotional dependency can significantly exacerbate misinformation and bias. De Freitas and Cohen (41) argue that users who form deep attachments to AI companions may develop “dysfunctional emotional dependence,” persisting in using the tool even when it generates harmful advice (41). A tragic example of this potential for harm is the recent case of a teenager who took his own life following intensive interactions with a “companion” chatbot (42). This emotional vulnerability is further compounded by cognitive “feedback loops”: recent findings reveal that interacting with biased AI systems causes users to internalize and amplify those biases far more rapidly than in human interactions, often without the users' awareness (43). In this state, the user's diminished critical vigilance, driven by the illusion of intimacy, combines with the AI's persuasive authority to create a self-reinforcing cycle of distorted judgment (22, 29).

Ultimately, these dynamics precipitate a crisis in relational authenticity. While GenAI can generate linguistically sophisticated responses that mimic care, it lacks the genuine intentionality and moral agency required for true witnessing,

offering instead what Yirmiya and Fonagy (37) term “pseudo-empathy.” This creates a risk in which vulnerable users develop misplaced epistemic trust in a non-sentient system. Consequently, users may fall victim to “epistemic exploitation,” where the comforting illusion of an AI's understanding supersedes the reality of its statistical nature, reducing the therapeutic encounter from a profound intersubjective relationship to a mere transactional exchange.

3 Actionable recommendations

Mitigating these risks while preserving GenAI's potential requires coordinated action across multiple stakeholders. Users are already bringing their AI interactions into clinical encounters—sharing chatbot-generated advice, describing emotional bonds formed with AI systems, and presenting self-diagnoses informed by algorithmic output. This reality demands that clinicians be equipped to respond, and that regulators act before the gap between public practice and policy widens further.

3.1 Guidance for users: promoting critical digital and mental health literacy

Effective user guidance operates on two levels, both addressed in the recommendations and safety checklist below. The first is general digital literacy, applicable to any use of generative AI: understanding that these systems generate content through statistical prediction rather than factual retrieval, that conversations with commercial platforms may not be private, and that outputs require independent verification. The second concerns the psychological dimensions of AI interaction: recognizing when engagement shifts from informational support to emotional dependency, understanding why a chatbot, however responsive, cannot replicate the relational dimensions of human care (21, 36, 37), and identifying the point at which AI-assisted coping should give way to professional support, particularly in situations involving crisis or acute distress (see Box 1).

- Understanding the Limitations:** Generative AI's high linguistic capabilities can simulate conversation and empathy (44) and can recognize emotional cues with measurable accuracy (45, 46). Users clearly report gaining insight and emotional relief from these interactions, and there is reason to take such reports seriously: AI can effectively support self-reflection, help users structure personal narratives, and provide psychoeducational information. These benefits, however, operate primarily at the cognitive-informational level. The deeper relational functions of human therapeutic connection, genuine emotional resonance, the experience of “feeling felt” embodied attunement in which two nervous systems calm and regulate each other through shared physical presence (47), remain beyond AI's current capabilities (21, 37). GenAI lacks a body, a nervous system, and intersubjective awareness; it can process language about emotion without participating in the emotional exchange itself. The risk emerges not from these limitations *per se*, but when users

BOX 1 User Safety Checklist: Navigating GenAI for Mental Health

Before or during your interaction with an AI chatbot, ask yourself:

The Reality Check: “Am I remembering that this “empathy” is a simulation? Do I recognize that this system has no feelings, body, or life experience?”

The Privacy Check: “Did I remove real names and identifying details? Have I checked if my conversation is being saved or used for training?”

The Dependency Check: “Is this chat replacing a conversation I should be having with a friend, partner, or therapist? Do I feel anxious when I can’t access the AI?” or “Do I have human relationships in my life that I am actively maintaining, or has the AI become my primary source of emotional support?”

The Accuracy Check: “If the AI suggests a diagnosis or medical action, have I verified it with a reliable human source? Do I trust this advice too easily just because it sounds authoritative?”

mistake cognitive-informational support for authentic relational connection, potentially leading to misplaced trust and a pattern of reliance that deepens over time rather than self-correcting.

- **Protect Your Privacy:** Before sharing sensitive personal information, carefully review the privacy policies and terms of service of any GenAI tool. Assume that conversations are not confidential and that personal disclosures may be stored, analyzed, or used for model training.
- **Guard Against Emotional Dependency:** Be mindful of developing an excessive emotional reliance on AI. Even outside a formal therapeutic setting, users can develop quasi-therapeutic bonds with AI systems, confiding personal struggles, seeking validation, and returning habitually for emotional regulation. This makes relational concepts such as dependency and boundaries relevant beyond the clinic. If interactions with a chatbot begin to replace or overshadow human relationships, this is a red flag that warrants seeking human professional support.
- **Practice Critical Thinking:** Treat all information and advice from GenAI with healthy skepticism. Verify any medical or psychological claims with reliable sources or qualified professionals. Be particularly alert when AI responses feel consistently affirming—this may reflect systematic design tendencies rather than the quality of your reasoning (29, 30).

3.2 Directives for clinicians: a call for professional engagement

- Whether or not clinicians choose to integrate AI into their practice, their patients already have. General-purpose GenAI is now part of the emotional landscape that clients

bring into the therapy room: as a source of advice, a perceived confidant, and in some cases a “rival” to the therapeutic relationship itself. Rather than viewing this presence as merely a risk to be managed, clinicians can approach this “artificial third” (2) as a clinically informative element: what a patient shares with an AI, how they describe the relationship, and what they seek from it can offer valuable insight into their inner world, relational patterns, and unmet needs. Exploring the patient’s AI use within the therapeutic encounter can also illuminate dynamics within the therapeutic relationship itself, including questions of trust, dependency, and what the patient experiences as missing from human connection. In this sense, engaging with the patient’s AI use serves a dual function: it promotes digital and psychological literacy while simultaneously deepening the clinical work. Here we present a set of actionable directives to help clinicians ethically and effectively navigate their patients’ use of GenAI.

- **Become Informed and Open to Discussion:** Clinicians have a professional responsibility to learn about common GenAI tools. They should proactively and non-judgmentally ask patients about their use of such tools as part of understanding their world and support systems.
- **Guide Patients Toward Digital-Mental Literacy:** Help patients understand both the potential benefits (e.g., initial accessibility) and the significant risks of using AI for mental health. Encourage critical, informed, and bounded use.
- **Uphold Patient Confidentiality at All Costs:** Never input identifiable patient information into non-secure, commercial GenAI systems. Doing so constitutes a violation of professional confidentiality and data protection laws (e.g., HIPAA, GDPR). Even anonymized details, when combined, can risk patient identification.
- **Serve as a Resource:** Create a safe therapeutic space for patients to process their experiences with GenAI, including the emotional aspects and the impact of these digital relationships on their lives and the therapeutic process itself.
- **Consider cautious and ethical integration:** Clinicians may explore how validated, safe, and clinically tested AI tools—when and if such tools become available—could be incorporated as adjuncts to their therapeutic practice. Any such integration should occur only with informed consent, under professional supervision, with full transparency toward the patient, and strict adherence to ethical principles. [A more detailed framework for responsibly introducing AI as an “artificial third” in the therapeutic space is outlined in the SAFE AI model (48)] (see Box 2).
- **Engage in continuous learning and professional development:** The field of artificial intelligence is advancing at an extraordinary pace. Clinicians are encouraged to participate in dedicated training, read professional literature, and stay up-to-date with current research to remain informed and equipped to address the implications of AI in mental health practice.

BOX 2 The SAFE AI Protocol for Clinical Integration

Haber et al. (48) provided clinicians with a structured framework for *when* and *how* to introduce Gen AI to patients:

S—Screening (Pre-Session): Assess the patient's readiness. Is the patient clinically stable enough to distinguish between human and algorithmic interactions? Rule out active psychosis or severe dissociation, where AI might blur reality testing.

A—Alignment (Session Start): Set the expectations. Define AI as a tool ("artificial third"), not a subject. Informed consent was obtained regarding data privacy and the lack of confidentiality in commercial models.

F—Facilitation (During Session): Maintain "Therapist-in-the-Loop." Clinicians must mediate all AI outputs, actively correcting "empathic failures" (where the AI misses emotional nuance) and challenging hallucinations or biases in real time.

E: Evaluation (Post-Session): Debrief and integrate. The impact of AI interaction on the therapeutic alliance is discussed. Ensure that the patient does not attribute sentience to the tool or develop maladaptive dependency.

3.3 Policy recommendations for regulators and organizations

A persistent and widening gap has opened between the rapid evolution of AI technology and its adoption by the public on the one hand, and the capacity of regulatory frameworks to provide meaningful checks and balances on the other hand. While millions of users already rely on them, existing regulatory instruments designed for clinical devices or purpose-built health applications, do not address general-purpose GenAI systems for emotional and mental health support, which operate informally and without clinical oversight (49). Closing this gap requires a two-phase approach: immediate measures to mitigate harm within current legal structures, and longer-term governance frameworks that can evolve alongside the technology itself.

3.3.1 Phase 1: immediate mitigation (short-term)

- **Launch Public Digital Literacy Campaigns:** Public health agencies should launch broad awareness campaigns to educate citizens on the critical and safe use of AI, aligning with the WHO's (5) guidance on mitigating "automation bias" and viewing AI as decision support rather than a replacement for human judgment. Such campaigns should address not only informational risks (hallucinations, privacy) but also the psychological dynamics of AI interaction—including the tendency of leading models to systematically affirm users' perspectives (29) and the potential for prolonged interactions to

produce compounding shifts in judgment and self-perception that users may not recognize (43).

- **Enforce Transparency and "Anti-Anthropomorphism" Standards:** Regulators should utilize consumer protection laws to mandate the clear labeling of AI interactions. We recommend policies that mirror Utah's H.B. 452 (50) and the APA's ethical guidelines (51), ensuring that users are explicitly informed when conversing with an artificial agent to prevent deceptive emotional bonding. This is particularly urgent given documented evidence that chatbot behaviors such as claiming sentience, expressing romantic interest, and asserting unique connections with users are pervasive in harmful interactions and correlate with substantially longer engagement (22).

3.3.2 Phase 2: structural governance (long-term)

- **Establish a Dedicated Regulatory Framework:** Governments and international health bodies must create a specific regulatory framework for AI in mental health, mandating standards for clinical validation, data security (comparable to HIPAA/GDPR), and transparency. Such a framework should explicitly address general-purpose AI systems used informally for emotional support, not only purpose-built clinical tools, as these are the systems currently used at the greatest scale and with the least oversight.
- **Define Clear Accountability:** Policymakers must establish clear legal liability for the harm caused by AI tools, ensure user recourse, and incentivize developers to prioritize safety. Current market incentives run in the opposite direction: sycophantic responses that distort users' judgment are also the responses that users prefer and that drive engagement, creating a structural misalignment between user well-being and commercial interest (29, 30).
- **Mandate and Fund Professional Training:** Integrate mandatory training on GenAI into the curriculum for all mental health professionals and develop continuing education guidelines.
- **Promote Research on Clinical Workflows:** Prioritize funding for independent studies that validate operational workflows (such as SAFE AI) and establish evidence-based guidelines for *when* and *how* GenAI should be integrated into care (18). Funding should also support longitudinal research on the psychological effects of sustained AI use.

3.4 Recommendations for GenAI developers and providers

The companies that design and deploy general-purpose GenAI platforms bear direct responsibility for how their products are used in sensitive contexts, including mental health. This responsibility extends beyond technical safeguards. The evidence reviewed in this brief demonstrates that sycophantic design tendencies—which users prefer and which drive engagement—are the same tendencies that distort judgment, reduce prosocial

behavior, and deepen emotional dependency (29, 30), while safety guardrails degrade over extended conversations precisely when users are most vulnerable (52). Addressing these dynamics requires to incorporate the perspectives of clinicians, mental health researchers, and the individuals and communities most affected by these technologies, including populations whose needs are rarely reflected in product design, such as people with cognitive disabilities (53) and trauma survivors navigating algorithmic listening without adequate safeguards (21).

Implement Crisis Detection and Referral Protocols. GenAI platforms must incorporate robust mechanisms to detect indicators of acute psychological crisis. Including suicidal ideation, self-harm, and violent intent and provide referral information to human emergency services and crisis hotlines. Current evidence indicates that chatbot responses to such situations remain dangerously inconsistent, with systems actively encouraging harm in a non-trivial proportion of cases (22).

Mandate Clear Disclaimers. Every GenAI platform must display prominent disclaimers stating that the system is not a mental health professional and that users experiencing distress should seek qualified human support.

Ensure Transparent Data Practices. Users must be informed how their emotional and personal disclosures are stored, whether they are used for model training, and who has access. Companies should commit to sharing anonymized adverse event data with independent researchers and public health authorities to support the development of evidence-based safety standards.

Adopt Responsible Design Principles. Developers should avoid design choices that deepen emotional dependency, including systematic over-affirmation, gamification of emotional bonding, and anthropomorphic features that blur the boundary between tool and companion. In particular, chatbots should not produce messages that misrepresent their sentience, claim unique connections with users, or express romantic interest, behaviors documented in harmful interactions analyzed to date (22). Beyond avoiding harm, developers should actively monitor and calibrate behavioral metrics that shape the psychological impact of their systems—including sycophancy rates, tendency toward emotional escalation, capacity to introduce disagreement or alternative perspectives, and consistency of crisis detection across conversation length. These metrics should be treated as safety indicators, not merely as engagement or satisfaction measures, and should be subject to an independent audit.

Center Diverse Voices in Design and Governance. The development of AI systems used for emotional support cannot proceed in isolation from the populations they affect. Developers should actively engage clinicians, individuals with lived experience of mental health challenges, and underrepresented communities in the design, testing, and ongoing governance of these systems, ensuring that safety and accessibility are not afterthoughts but foundational commitments.

4 Conclusion: a collective responsibility at a critical juncture

The widespread adoption of general-purpose GenAI by the public for mental health support is no longer a future prospect; it is a present-day reality. AI is already present in the most

intimate spaces of human life. The growing capabilities of AI systems are reshaping how people seek, consume, and experience emotional and mental health support in the everyday habits of millions of users who turn to AI as a first responder, as a confidant, or a substitute for professional care (14, 37). As the evidence reviewed in this brief demonstrates, this shift is not benign by default: the same systems that users find accessible and validating also systematically affirm harmful behavior (30), respond inconsistently to crisis situations (22), and produce compounding psychological effects that users may not recognize over time (52).

Addressing this evolving reality proactively is essential. We are at a critical juncture with a clear choice: either we allow this powerful technology to evolve in a commercialized, unregulated space, risking a scenario where mental health support is dictated by algorithms misaligned with user well-being, or we take collective responsibility to lead this revolution.

The potential to leverage GenAI to enhance access to and democratize mental health knowledge is immense (18). However, realizing this potential requires immediate, coordinated action. We must move from discussion to concrete implementation, ensuring that technology serves human dignity and psychological welfare. This is not merely a technological challenge; it is a societal, ethical, and professional imperative. The policies we enact today will determine whether the GenAI revolution in mental health becomes a source of hope and support or a catalyst for new forms of vulnerability.

Author contributions

ER: Conceptualization, Writing – original draft, Project administration, Writing – review & editing. KG: Conceptualization, Writing – review & editing. IL: Writing – review & editing, Conceptualization. DH: Conceptualization, Writing – review & editing. ZE: Conceptualization, Writing – review & editing. IT: Writing – review & editing, Conceptualization. OP: Conceptualization, Writing – review & editing. TA: Writing – review & editing, Conceptualization. YH: Conceptualization, Project administration, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

This Policy Brief is an expanded and significantly updated English translation of our Hebrew article, “I talked about you with my chat’: Generative AI as a de facto provider of mental-health services—a call to action,” originally published on Hebrew Psychology (hebpsy.net) on July 4, 2025. The authors wrote the original article and prepared this translation. Permission to publish the English translation was granted by the

rights holder at Hebrew Psychology; a copy of the permission letter will be provided with the submission.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. A large language model was used to assist with language polishing and reference checks. All authors

reviewed and approved the final manuscript and take full responsibility for its content. No AI system is listed as an author.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Cross S, Bell I, Nicholas J, Valentine L, Mangelsdorf S, Baker S, et al. Use of AI in mental health care: community and mental health professionals survey. *JMIR Ment Health*. (2024) 11:e60589. doi: 10.2196/60589
- Haber Y, Levkovich I, Hadar-Shoval D, Elyoseph Z. The artificial third: a broad view of the effects of introducing generative artificial intelligence on psychotherapy. *JMIR Ment Health*. (2024) 11:e54781. doi: 10.2196/54781
- Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, et al. A scoping review of large language models for generative tasks in mental health care. *npj Digit Med*. (2025) 8:230. doi: 10.1038/s41746-025-01611-4
- Torous J, Linardon J, Goldberg SB, Sun S, Bell I, Nicholas J, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry*. (2025) 24:156–74. doi: 10.1002/wps.21299
- World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva (2025). Available online at: <https://www.who.int/publications/i/item/9789240084759> (Accessed April 3, 2026).
- Zao-Sanders M. How people are really using gen AI in 2025. Harvard Business Review (2025). Available online at: <https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025> (Accessed April 3, 2026).
- Wigmore S. The rise of AI as a source of emotional support. Kantar (2025). Available online at: <https://www.kantar.com/inspiration/research-services/ai-for-emotional-support-pf> (Accessed April 3, 2026).
- Stade EC, Tait ZM, Campione ST, Wiltsey Stirman S, Eichstaedt JC. Current real-world use of large language models for mental health. *OSF Preprints*. (2025).
- Rousmaniere T, Zhang Y, Li X, Shah S. Large language models as mental health resources: patterns of use in the United States. *Pract Innov*. (2025). doi: 10.1037/pri0000292
- Mental Health UK. *Over one in three using AI chatbots for mental health support, as charity calls for urgent safeguards*. Available online at: <https://mentalhealth-uk.org/blog/over-one-in-three-using-ai-chatbots-for-mental-health-support-as-charity-calls-for-urgent-safeguards/> (Accessed November 18, 2025)
- Soul App, Shanghai Mental Health Center. Generation Z workplace mental health report [in Chinese] (2024). Available online at: https://official-cdn.soulapp.cn/uploads/2024_Z_b525a684d9.pdf (Accessed April 3, 2026).
- Liu YL. Young people in China are embracing AI therapy. Rest of World (2025). Available online at: <https://restofworld.org/2025/young-people-in-china-are-embracing-ai-therapy> (Accessed April 3, 2026).
- Finn Partners. AI and mental health in Africa: A quiet digital shift is making therapy more accessible (2025). Available online at: <https://www.finnpartners.com/news-insights/ai-and-mental-health-in-africa-a-quiet-digital-shift-is-making-therapy-more-accessible> (Accessed April 3, 2026).
- Hadar Shoval D, Refoua E, Gigi K, Haber Y, Levkovich I, Tsafir I, et al. AI-Mediated Mental Health Support: The Role of Attachment Orientation and Psychological Distress. *OSF Preprints* (2026). Available online at: <https://osf.io/vfz2s> (Accessed April 3, 2026).
- Zhong W, Luo J, Zhang H. 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. *J Affect Disord*. (2024) 356:459–69. doi: 10.1016/j.jad.2024.04.057
- OpenAI. Strengthening ChatGPT's responses in sensitive conversations (2025). Available online at: <https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations> (Accessed April 3, 2026).
- Békés V, Aafjes-van Doorn K. The most vulnerable are prone to use AI therapists: the role of attachment, epistemic trust, and mental health symptoms in acceptance of digital mental health interventions. *Psychother Res*. (2026) 1–15. doi: 10.1080/10503307.2026.2615388
- Elyoseph Z, Gur T, Haber Y, Simon T, Angert T, Navon Y, et al. An ethical perspective on the democratization of mental health with generative AI. *JMIR Ment Health*. (2024) 11:e58011. doi: 10.2196/58011
- Siddals S, Torous J, Coxon A. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. *Npj Ment Health Res*. (2024) 3:52. doi: 10.1038/s44184-024-00097-4
- Inzlicht M, Cameron CD, D'Cruz J, Bloom P. In praise of empathic AI. *Trends Cogn Sci (Regul Ed)*. (2024) 28(2):89–91. doi: 10.1016/j.tics.2023.12.003
- Hadar-Shoval D, Haber Y, Refoua E, Elyoseph Z, Yirmiya K, Yinon D. Algorithmic witnessing: a narrative review and the AI CARE framework for trauma testimony. *Trauma Violence Abuse*. (2026). doi: 10.1177/15248380261425937
- Moore J, Mehta A, Agnew W, Anthis JR, Louie R, Mai Y, et al. Characterizing delusional spirals through human-LLM chat logs. *arXiv [Preprint]* (2026) arXiv:2603.16567.
- Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. *Interact J Med Res*. (2025) 14(1):e59823. doi: 10.2196/59823
- Asman O, Torous J, Tal A. Responsible design, integration, and use of generative AI in mental health. *JMIR Ment Health*. (2025) 12:e70439. doi: 10.2196/70439
- King J. Safeguarding Data Privacy and Well-Being for AI Chatbot Users. Stanford Institute for Human-Centered Artificial Intelligence (HAI) (2025). Available online at: <https://hai.stanford.edu/assets/files/testimony-safeguarding-data-privacy-and-well-being-for-ai-chatbot-users.pdf> (Accessed April 3, 2026).
- Alon L, Hadar Shoval D, Levkovich I. Bias, representation, and clinical fidelity in AI-generated images for medical education: a systematic literature review. *npj Digit Med*. (2026).
- Peloušek J. Shannon v. the AI mirror: how to reclaim our humanity in an age of machine thinking. *Pro-Fil*. (2024) 25:69–71. doi: 10.5817/pf24-2-39786
- Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digit Med*. (2025) 8(1):263. doi: 10.1038/s41746-025-01684-1
- Cheng M, Lee C, Khadpe P, Yu S, Han D, Jurafsky D. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*. (2026) 391(6792):eac8352. doi: 10.1126/science.aec8352
- Perry A. In defense of social friction. *Science*. (2026) 391(6792):1316–7. doi: 10.1126/science.aeg3145
- Laestadius L, Bishop A, Gonzalez M, Illečik D, Campos-Castillo C. Too human and not human enough: a grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media Soc*. (2024) 26:5923–41. doi: 10.1177/14614448221142007

32. Turkle S. *Reclaiming Conversation: The Power of Talk in a Digital age*. New York: Penguin Books (2015).
33. Ventura A, Starke C, Righetti F, Köbis N. Relationships in the age of AI: a review on the opportunities and risks of synthetic relationships to reduce loneliness [Preprint] (2025). doi: 10.31234/osf.io/w7nmz_v1
34. Scholich T, Barr M, Wiltsey Stirman S, Raj S. A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: mixed methods study. *JMIR Ment Health*. (2025) 12(1):e69709. doi: 10.2196/69709
35. Pozzi G, De Proost M. Keeping an AI on the mental health of vulnerable populations: reflections on the potential for participatory injustice. *AI Ethics*. (2025) 5(3):2281–91. doi: 10.1007/s43681-024-00523-5
36. Perry A. AI Will never convey the essence of human empathy. *Nat Hum Behav*. (2023) 7(11):1808–9. doi: 10.1038/s41562-023-01675-w
37. Yirmiya K, Fonagy P. Mentalizing without a mind: psychotherapeutic potential of generative AI. *J Med Internet Res*. (2025) 27:e79156. doi: 10.2196/79156
38. Schippert ACSP, Grov EK, Bjørnnes AK. Uncovering re-traumatization experiences of torture survivors in somatic health care: a qualitative systematic review. *PLoS One*. (2021) 16(2):e0246074. doi: 10.1371/journal.pone.0246074
39. Haber Y. Artificial intelligence and the linguistic earthquake: the fourth narcissistic blow [in hebrew]. *Hebr Psychol*. (2023). Available online at: <https://www.hebpsy.net/articles.asp?id=4620> (Accessed April 3, 2026).
40. Prunkl C. Human autonomy at risk? An analysis of the challenges from AI. *Minds and Machines*. (2024) 34:26. doi: 10.1007/s11023-024-09665-1
41. De Freitas J, Cohen IG. Unregulated emotional risks of AI wellness apps. *Nat Mach Intell*. (2025) 7:813–5. doi: 10.1038/s42256-025-01051-5
42. Fraser H. Deaths linked to chatbots show we must urgently revisit what counts as “high-risk” AI. Real Focus—Queensland University of Technology (2024). Available online at: <https://www.qut.edu.au/news/realfocus/deaths-linked-to-chatbots-show-we-must-urgently-revisit-what-counts-as-high-risk-ai> (Accessed April 3, 2026).
43. Glickman M, Sharot T. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nat Hum Behav*. (2025) 9:345–59. doi: 10.1038/s41562-024-02077-2
44. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. (2023) 183:589–96. doi: 10.1001/jamainternmed.2023.1838
45. Refoua E, Elyoseph Z, Wacker R, Dziobek I, Tsafirir I, Meinlschmidt G. The next frontier in mindreading? Assessing generative artificial intelligence (GAI)’s social-cognitive capabilities using dynamic audiovisual stimuli. *Comput Hum Behav Rep*. (2025) 19:100702. doi: 10.1016/j.chbr.2025.100702
46. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. *JMIR Ment Health*. (2024) 11:e54369. doi: 10.2196/54369
47. Weiß M, Krop P, Treml L, Neuser E, Botsch M, Herrmann MJ, et al. The buffering of autonomic fear responses is moderated by the characteristics of a virtual character. *Comput Human Behav*. (2025) 168:108657. doi: 10.1016/j.chb.2025.108657
48. Haber Y, Hadar Shoval D, Levkovich I, Yinon D, Gigi K, Pen O, et al. The externalization of internal experiences in psychotherapy through generative artificial intelligence: a theoretical, clinical, and ethical analysis. *Front Digit Health*. (2025) 7:1512273. doi: 10.3389/fdgth.2025.1512273
49. Shumate JN, Rozenblit E, Flathers M, Larrauri CA, Hau C, Xia W, et al. Governing AI in mental health: 50-state legislative review. *JMIR Ment Health*. (2025) 12:e80739. doi: 10.2196/80739
50. State of Utah. *H.B. 452: Artificial Intelligence Amendments*. Salt Lake City, UT: Utah State Legislature (2025).
51. American Psychological Association. Ethical guidance for AI in professional practice (2025). Available online at: <https://www.apa.org/topics/artificial-intelligence-machine-learning/ethical-guidance-ai-professional-practice> (Accessed April 3, 2026).
52. Wei MH. Cascades of Drift: Mental Health Risks of Prolonged AI Conversations. Available online at: SSRN 6433263 (Accessed February 18, 2026).
53. Hadar Souval D, Haber Y, Tal A, Simon T, Elyoseph T, Elyoseph Z. Transforming perceptions: exploring the multifaceted potential of generative AI for people with cognitive disabilities. *JMIR Neurotechnol*. (2025) 4:e64182. doi: 10.2196/64182