
CONTINUING EDUCATION

Cybersecurity, Privacy, and Informed Consent in Generative Artificial Intelligence Mental Health Applications: A Continuing Education Manual for Psychologists

Reading Manual for Licensed Psychologists
and Graduate-Level Clinicians

FRANKLIN CARVAJAL

August 2026

Contents

Abstract	4
Introduction	5
Historical and Theoretical Foundations	6
Conceptual Models and Mechanisms	8
Integrated Review of the Evidence	10
Deployment Patterns	11
Risk Landscape	11
Methodological Quality of the Evidence Base	12
Safeguards and Regulatory Alignment	13
Synthesis Table of Integrated Findings	13
Professional Interpretation and Critical Analysis	14
Psychological Pathways and Stress Responses	17
Mental Health Outcomes and Severity Spectrum	19
System-Level and Contextual Psychological Effects	21
Clinical Implications for Psychologists	23
Assessment and Formulation	23
Treatment Planning	23
Documentation	24
Referral and Coordination	24
Professional Development	24
Ethical Considerations in Practice	24
Competence	25
Informed Consent	25
Confidentiality	26

Multiple Relationships and Boundaries	26
Cultural Responsiveness	26
Scope of Practice	27
Practice Recommendations	28
Well-Supported Recommendations	28
Tentative Recommendations	29
Synthesis Table of Practice Recommendations	30
Future Directions and Research Gaps	31
Conclusion	31
References	32

Abstract

The rapid integration of generative artificial intelligence (AI) systems into mental health service delivery has outpaced the development of governance frameworks capable of protecting patient welfare, privacy, and autonomy. Large language models (LLMs) are now widely deployed in psychoeducation, emotional support, screening, crisis-adjacent interactions, and therapeutic dialogue, often through consumer-facing platforms operating outside traditional clinical oversight. This continuing education manual synthesizes contemporary evidence on the cybersecurity, privacy, and ethical risks associated with generative AI mental health chatbots and examines the parallel crisis in informed consent governing the use of patient data to train these systems. Drawing on a PRISMA-guided systematic review of 33 primary studies (Sawesi et al., 2026) and a conceptual analysis of consent frameworks in digital mental health (Larrauri et al., 2026), the manuscript characterizes multi-layered threat surfaces at the model, interface, data, and infrastructure levels; interprets the collective evidence for practitioners; explores psychological pathways by which unsafe outputs, privacy breaches, and inadequate consent processes affect vulnerable users; and proposes an integrated framework linking risk types to technical controls and regulatory standards including the U.S. Food and Drug Administration Software as a Medical Device (SaMD) guidance, the Health Insurance Portability and Accountability Act (HIPAA), International Organization for Standardization (ISO) 14971, and the National Institute of Standards and Technology (NIST) AI Risk Management Framework. Clinical implications, ethical considerations grounded in principles of competence, informed consent, confidentiality, and cultural responsiveness, and concrete practice recommendations are developed throughout. Case scenarios illustrate application of the material to routine clinical decision-making. The manuscript is intended for licensed psychologists and graduate-level clinicians seeking to develop informed, defensible, and ethically grounded positions on the use, recommendation, and oversight of generative AI tools in their practices.

Introduction

Few developments in the history of behavioral health technology have moved as rapidly, or with as little clinical scaffolding, as the deployment of large language model-based conversational agents in mental health contexts. Within a very short window, systems originally designed for general-purpose text generation have become *de facto* sources of emotional support, coping guidance, and even crisis-adjacent responses for millions of people. Some of this activity occurs within regulated clinical settings; a substantial and rapidly growing share occurs in consumer-facing environments where users interact with chatbots without any clinician involvement, without formal informed consent processes, and often without a clear understanding of how their disclosures are stored, analyzed, or reused (Sawesi et al., 2026). For licensed psychologists, this shift is not incidental to practice. It reshapes what patients bring into the consulting room, what they expect from digital adjuncts to care, and what obligations clinicians assume when they recommend, tolerate, or ignore the use of AI-based tools by the people they treat.

The clinical stakes are unusually high. Mental health interactions involve disclosures of suicidal ideation, trauma, psychiatric history, substance use, interpersonal violence, and other information that carries profound implications for safety, employment, insurance, custody, and personal identity. When such disclosures are made to a human clinician bound by professional ethics, statutory confidentiality protections, and licensure oversight, they are protected by a mature system of safeguards. When the same disclosures are made to a generative chatbot governed by a terms-of-service agreement, the protective architecture is dramatically thinner and, in many cases, structurally incompatible with the confidentiality expectations users bring to the interaction (Larrauri et al., 2026; Sawesi et al., 2026). This asymmetry between user expectation and platform reality lies at the heart of the challenges considered in this manual.

Two complementary bodies of scholarship inform the synthesis presented here. The first, a PRISMA-guided systematic review by Sawesi et al. (2026), mapped 33 primary studies examining cybersecurity, privacy, and safety risks associated with generative AI systems used in mental health. Their analysis identified a multi-layered risk landscape spanning model, interface, data, and infrastructure levels, and it documented substantial gaps between existing governance frameworks (including FDA SaMD guidance, HIPAA, ISO 14971, and the NIST AI Risk Management Framework) and the actual operational characteristics of probabilistic, interaction-driven conversational systems. The second, a conceptual commentary by

Larrauri, Martinez-Martin, and Torous (2026), analyzed the erosion of informed consent as companies increasingly train mental health AI models on therapy transcripts under contractual assent regimes that do not meet the ethical or clinical standards of true informed consent. Read together, these works articulate an emerging picture: the technical, regulatory, and ethical infrastructure required for safe integration of generative AI into mental health services has not kept pace with deployment, and the burden of navigating this gap falls, in significant measure, on individual clinicians and patients.

This manual addresses that reality. It is designed to help licensed psychologists develop a technically informed, ethically grounded, and clinically usable understanding of the risks and safeguards associated with generative AI mental health tools. The material is not a technical manual for engineers, nor is it a regulatory primer for policy specialists; rather, it is a synthesis for practicing clinicians who must decide how to talk with patients about AI use, how to document consent when digital tools are involved in care, how to recognize when a patient's engagement with a chatbot is affecting clinical presentation, and how to advocate for the governance changes their patients require. Throughout, case scenarios illustrate how the evidence translates into decisions that clinicians actually make.

Historical and Theoretical Foundations

To understand the current moment in AI mental health governance, it is useful to situate it within a longer arc of medical innovation built on data and tissue acquired from patients whose autonomy was, at best, incompletely respected. Larrauri et al. (2026) draw an explicit parallel between the appropriation of Henrietta Lacks's cells in 1951 and the contemporary appropriation of patient digital data for AI model training. In both instances, biological or informational material acquired in the course of receiving care became the substrate for scientific and commercial value creation on a scale, and across contexts, that neither the individuals nor their clinicians could have anticipated. The parallel is not merely rhetorical. It captures a structural feature of medical innovation: the material extracted from patient encounters tends to acquire a form of durability, or in the language of Larrauri et al. (2026), an "immortality," that extends the reach of a single act of disclosure across time, institutions, and purposes. In the case of HeLa cells, the durability was biological; in the case of patient conversational data used to train language models, the durability is informational and, in some respects, more diffuse and less traceable.

The historical foundation of informed consent in psychology and medicine emerged in response to exactly this sort of asymmetry between what patients disclosed and what was done with those disclosures. Contemporary informed consent doctrine rests on three foundational commitments: disclosure of relevant information, voluntariness of the decision, and comprehension by the person consenting (Larrauri et al., 2026). Each of these commitments is stressed in ways that are historically novel by generative AI systems. Disclosure is stressed because the downstream uses of conversational data (secondary uses, inference generation, group profiling, model updates) cannot always be fully specified at the time of consent. Voluntariness is stressed because many patients access AI-mediated services through employer-sponsored platforms where refusal may entail forgoing care altogether (Larrauri et al., 2026). Comprehension is stressed because terms of service documents are typically written at reading levels and lengths that empirical evidence suggests few users can meaningfully process (Larrauri et al., 2026).

The theoretical foundation for evaluating cybersecurity and privacy risks in medical devices has evolved primarily around deterministic systems. Traditional medical software validation frameworks, including ISO 14971's structured approach to hazard identification and risk control (Sawesi et al., 2026), assume that a system's behavior can be characterized, tested against defined specifications, and validated with reasonable confidence that its outputs will remain within predictable bounds across the anticipated range of inputs. Generative AI systems violate this assumption. Their outputs are probabilistic, context-dependent, and sensitive to user input in ways that make exhaustive validation impossible in principle. The Software as a Medical Device (SaMD) framework advanced by the International Medical Device Regulators Forum and adopted by the FDA (Sawesi et al., 2026) was developed with deterministic software in mind, and although its principles remain relevant, its application to generative conversational systems requires substantial extension.

Similar strain is evident in privacy governance. The HIPAA Security Rule establishes administrative, physical, and technical safeguards for electronic protected health information within covered entities (Sawesi et al., 2026), but many consumer-facing mental health chatbot platforms do not qualify as covered entities. In the European Union, the General Data Protection Regulation (GDPR) offers broader baseline protections, yet even GDPR's "right to erasure" faces practical limits when personal data has been incorporated into trained models where selective removal may be technically infeasible (Larrauri et al., 2026). These regulatory frameworks were designed for a world in which health data existed largely in structured records within

institutional custody. They map imperfectly onto a world in which patients disclose deeply sensitive information through free-text conversations with commercial platforms whose data-handling practices are opaque even to sophisticated users.

The theoretical challenge, then, is not simply to update existing frameworks with new provisions for AI. It is to recognize that the operational logic of generative conversational systems differs from that of the technologies for which current safeguards were designed, and to construct governance approaches capable of tracking that logic. Sawesi et al. (2026) argue that this requires a multi-layered approach in which risks are analyzed at the model level, the interface level, the data level, and the infrastructure level, with recognition that vulnerabilities in one layer propagate to others. Larrauri et al. (2026) argue in parallel that consent must be reconceptualized as an ongoing, dynamic process rather than a one-time transactional event. Both arguments locate the theoretical center of the problem in the interaction between system properties and human vulnerability, rather than in either dimension alone.

Conceptual Models and Mechanisms

A useful conceptual entry point is what Sawesi et al. (2026) characterize as the "structural tension" inherent in generative AI mental health systems: the same conversational interface that enables scalable and accessible support also functions as the primary attack surface for cybersecurity threats, privacy breaches, and safety failures. This framing is important because it locates risk not in any single failed component but in the essential architecture of the systems. A conversational interface must accept open-ended user input; that same openness allows prompt injection attacks. A generative model must produce fluent, contextually adaptive responses; that same fluency enables convincing hallucinations. A useful digital mental health platform must collect sensitive user disclosures; that same collection creates high-value data assets that attract secondary use, monetization, and breach. Risk is not a defect to be patched; it is a structural consequence of the design.

Sawesi et al. (2026) organize this structural analysis around four interacting system layers. The **model layer** encompasses the trained language model itself and its behavioral properties, including susceptibility to data poisoning during training, unintended memorization of training data, hallucination, bias, and stigmatizing outputs. The **interface layer** encompasses the conversational surface through which users interact with the system, including susceptibility to prompt injection, adversarial manipulation, and failures of crisis detection or escalation. The **data**

layer encompasses the storage, processing, and secondary use of conversational data, including sensitive disclosures, re-identification risk, third-party sharing, and retention practices. The **infrastructure layer** encompasses the underlying computing environment, including access controls, API security, authentication mechanisms, and integration with other systems. Each layer generates distinctive threats, and vulnerabilities at one layer typically compound risks at others. Inadequate input sanitization at the interface layer, for example, can enable exfiltration of data from the data layer; weak infrastructure controls can enable unauthorized model queries that in turn expose model-layer memorization.

Complementing this technical model, Larrauri et al. (2026) develop a conceptual account of how consent frameworks map (or fail to map) onto AI-based mental health services. They distinguish three consent frameworks with progressively weaker protections. The most rigorous is **informed consent for clinical research**, which emphasizes explicit disclosure, documented voluntariness, and demonstrated comprehension. The intermediate is **authorization for secondary use of data**, which typically operates through de-identification, institutional review board waivers, or legal grounds that dispense with individual consent for defined purposes. The weakest, and the one on which digital mental health platforms most heavily rely, is **contractual assent embedded in terms of service**, in which users click "I agree" to lengthy documents they typically neither read nor understand. Larrauri et al. (2026) argue that reliance on the third framework is fundamentally inadequate for mental health data given its sensitivity, the vulnerability of the populations disclosing it, and the largely unforeseeable downstream uses enabled by AI systems.

A further conceptual mechanism worth emphasizing is what might be called the **inference-generating property** of AI systems. Traditional privacy analysis focuses on the confidentiality of specific data elements: a name, a diagnosis, a medication. AI systems, however, generate inferences that may not exist as explicit elements in any data record. From patterns in language use, timing of interactions, topics of discussion, and other conversational features, AI systems can infer psychological attributes, mental health status, risk profiles, and personal characteristics that users never explicitly disclosed (Larrauri et al., 2026; Sawesi et al., 2026). This property has two important implications. First, it means that de-identification, even when rigorously performed, does not eliminate privacy risk, because inferences can be linked to individuals through other data. Second, it means that the informational content of a conversation with a chatbot is not limited to what the user consciously shared; it includes what a sufficiently sophisticated model can derive from the aggregate pattern of the interaction. For mental health disclosures, where users may

inadvertently reveal information about employment, relationships, substance use, or trauma, the inference-generating property of AI systems creates a category of exposure that traditional privacy frameworks do not adequately address.

Finally, the conceptual model must account for **user expectations as a determinant of harm**. Sawesi et al. (2026) and Larrauri et al. (2026) both emphasize that users of mental health chatbots frequently bring to those interactions expectations of confidentiality, therapeutic competence, and clinical safeguards that are largely inherited from experience with human clinicians and from cultural representations of mental health care. When those expectations diverge sharply from platform reality, harm can occur even when platforms operate within their disclosed policies. A user who discloses suicidal ideation to a chatbot believing that the interaction is confidential and therapeutically competent has been harmed by the mismatch between expectation and reality, regardless of whether the platform's terms of service technically permitted the disclosure to be logged, analyzed, and used for model training. Recognizing user expectation as an independent driver of harm is essential for developing governance frameworks capable of protecting welfare rather than merely liability.

Integrated Review of the Evidence

The systematic review conducted by Sawesi et al. (2026) provides the most structured account currently available of how generative AI is being deployed in mental health, what risks these deployments generate, and how existing safeguards align with those risks. Their PRISMA-guided synthesis identified 33 primary studies meeting eligibility criteria from an initial pool of 444 records identified across PubMed, APA PsycNet, and Google Scholar. The included studies were heterogeneous in design (cross-sectional evaluations, qualitative and mixed-methods studies, benchmarking studies, surveys, randomized trials, and platform development studies) and geographically concentrated in high-income countries, with the United States accounting for 42.4% of studies. Only one included study was conducted in a lower-resource setting.

Deployment Patterns

Deployment patterns across the reviewed literature reveal several consistent features. Generative AI mental health chatbots are most frequently used for therapy or emotional support (13 of 33 studies, 39.4%), followed by safety evaluation or benchmarking (9 of 33, 27.3%) and psychoeducation or advice (6 of 33, 18.2%). At the domain level, suicide prevention or crisis detection is the most common application area (10 of 33, 30.3%), a pattern with particular significance given that crisis interactions represent the highest-stakes context in which conversational systems currently operate. Most reviewed systems (21 of 33, 63.6%) relied on general-purpose LLMs rather than purpose-built mental health models, and nearly half (16 of 33, 48.5%) were deployed through consumer-facing platforms rather than clinical research environments or supervised clinical settings.

This deployment profile is clinically consequential. When Sawesi et al. (2026) classified the reviewed applications by clinical risk, they found that 78.8% were classified as high risk and an additional 6.1% as very high risk, with only a small fraction categorized as moderate or lower risk. In other words, the modal use case for generative AI in mental health involves a general-purpose language model, deployed to consumers without clinician supervision, in a domain (crisis, therapy, emotional support) where errors carry serious clinical consequences.

Risk Landscape

The risk landscape identified across the reviewed studies falls into three overlapping categories: cybersecurity risks, privacy risks, and safety risks associated with chatbot outputs.

Cybersecurity risks were the least frequently evaluated across the reviewed literature, with formal adversarial testing or structured threat modeling appearing in only a small minority of studies. Where cybersecurity concerns were reported, they included unauthorized access to chatbot platforms or user data, data leakage or breach, prompt injection and API abuse, adversarial manipulation of model outputs, and data poisoning during training (Sawesi et al., 2026). The relative sparseness of formal cybersecurity evaluation across the literature is itself a critical finding: it means that the empirical evidence base for characterizing actual attack behavior against mental health chatbots is thin, and that risk assessment in this domain is currently more conceptual than empirical.

Privacy risks were more extensively documented and reflect the inherently sensitive nature of mental health interactions. Sensitive disclosure of personal information

was the most frequently identified concern (18.2% of studies), followed by re-identification risk (12.1%), and governance and confidentiality gaps (15.2%). Studies examining user perspectives consistently found that individuals frequently misunderstand the level of privacy protection provided by chatbot systems, often assuming healthcare-grade confidentiality where such protections do not exist (Sawesi et al., 2026). This finding is directly relevant to informed consent, because it suggests that the assent obtained through terms-of-service agreements is regularly premised on a mistaken understanding of what will happen to disclosed information.

Safety risks related to chatbot outputs were the most extensively documented category and, in many respects, the most clinically pressing. Harmful or unsafe advice was reported in 45.5% of studies, and failures in crisis response or escalation appeared in 27.3%. Additional documented safety concerns included bias or stigmatizing responses (24.2%), hallucination or misinformation (24.2%), overreliance or dependency risks (12.1%), and misclassification or diagnostic errors (3.0%) (Sawesi et al., 2026). The prevalence of documented safety failures across such a large share of the reviewed literature suggests that these are not isolated incidents in poorly designed systems but persistent features of current-generation generative AI deployed in mental health contexts.

Methodological Quality of the Evidence Base

An important dimension of the reviewed evidence is its own methodological limitations. Sawesi et al. (2026) classified 24 of 33 studies (72.7%) as having moderate risk of bias, 6 (18.2%) as high risk, and only 3 (9.1%) as low risk. Critically, in the specific domain of cybersecurity and safety evaluation rigor, 78.8% of studies (26 of 33) were classified as high risk, reflecting limited use of formal adversarial testing, structured threat modeling, or systematic security evaluation methods. The implication is that the evidence base characterizing generative AI mental health risks is itself constrained by the same limitations that characterize the systems it evaluates: descriptive and exploratory rather than rigorous and adversarial, and heavily reliant on benchmarking and simulated prompts rather than sustained real-world deployment data.

Safeguards and Regulatory Alignment

The final integrative finding from Sawesi et al. (2026) concerns the state of safeguards and regulatory alignment. Residual risk was discussed in all 33 included studies, with 57.6% reporting moderate residual risk and 36.4% reporting high residual risk. Only 6.1% reported low residual risk, indicating that even where safeguards are described, they are commonly insufficient to fully mitigate identified threats. Human-in-the-loop oversight was the most frequently reported safeguard (60.6%), followed by ethics approval (69.7%) and security or performance evaluation (63.6%). More structured practices were substantially less common: threat modeling was reported in only 30.3% of studies, explicit references to regulatory or legal compliance appeared in only 24.2%, and red-teaming or adversarial testing was reported in just 6.1%.

Synthesis Table of Integrated Findings

The following table synthesizes the collective findings from Sawesi et al. (2026) across the four system layers identified in their framework, integrating risk categories, prevalence in the literature, and severity implications for clinical decision-making.

System Layer	Primary Risks Identified	Prevalence in Reviewed Literature	Clinical Severity
Model	Hallucination, bias, stigmatizing responses, data poisoning	Hallucination/misinformation reported in 24.2% of studies; bias in 24.2%	High – direct impact on advice quality
Interface	Prompt injection, crisis-response failure, overreliance	Crisis-response failures in 27.3%; harmful advice in 45.5%	Very High – primary attack surface and direct user contact
Data	Sensitive disclosure, re-identification, retention, third-party sharing	Sensitive disclosure in 18.2%; governance gaps in 15.2%	High – long-lasting confidentiality implications
Infrastructure	Unauthorized access, API abuse, weak governance	Unauthorized access in 12.1%; formal audit in 6.1%	High – enables large-scale compromise

Note. Synthesized from Sawesi et al. (2026). Percentages reflect the proportion of the 33 included studies reporting each risk category; studies frequently reported

multiple risks, so categories are not mutually exclusive.

Professional Interpretation and Critical Analysis

Interpreting the collective evidence requires more than restating its findings. It requires evaluating what the pattern of findings, and the pattern of gaps in those findings, implies for the field. Three interpretive observations follow.

First, the empirical evidence base for safety and cybersecurity in generative AI mental health is systematically weaker than the deployment scale would warrant. Sawesi et al. (2026) document that 78.8% of studies were rated high risk in cybersecurity evaluation rigor, that formal adversarial testing appeared in only 6.1% of studies, and that structured threat modeling appeared in only 30.3%. This distribution is not a reflection of low risk; it is a reflection of a mismatch between the maturity of the technology's clinical role and the maturity of the evaluation infrastructure surrounding it. Systems now used by millions of people for mental health support are being evaluated primarily through benchmarking studies, simulated prompts, and descriptive analyses of user experience, rather than through the sort of adversarial, real-world, longitudinal evaluation that would ordinarily precede large-scale deployment of a technology capable of influencing high-stakes clinical outcomes. For psychologists, the practical implication is that claims made by chatbot vendors about safety, security, or clinical validity should be treated with substantial skepticism unless supported by independent, adversarial evaluation.

Second, the location of risk in the interface layer has particular significance for clinical practice. Sawesi et al. (2026) emphasize that the conversational interface functions as both the source of therapeutic benefit and the primary attack surface. This dual role means that single-layer evaluation, assessing content safety or software validation in isolation, is fundamentally inadequate. It also means that clinicians who evaluate the safety of a chatbot solely by assessing the appropriateness of its outputs are examining only one dimension of a multi-dimensional risk landscape. A chatbot that produces empathic, appropriate responses in typical interactions may still be vulnerable to prompt injection attacks that manipulate its behavior in atypical interactions, to data exfiltration that compromises the confidentiality of everything it has been told, and to infrastructure-level compromise that exposes user identity and history. Clinical evaluation of AI tools must consider risks across all four layers, not just the layer visible to end users.

Third, the gap between user expectation and platform reality is not merely an incidental problem to be fixed with better user interfaces; it is a structural feature of the current market for consumer mental health AI.

Larrauri et al. (2026) demonstrate that this gap is sustained by contractual assent regimes that reduce complex data governance to a binary accept-or-refuse choice, by market dynamics that reward rapid data accumulation, and by regulatory frameworks that impose weaker requirements on non-covered entities than on traditional healthcare providers. The gap cannot be closed by exhorting users to read terms of service more carefully; the empirical evidence that Larrauri et al. (2026) cite suggests that few users can meaningfully process such documents even when they attempt to do so. Closing the gap requires structural change: separate and explicit opt-in consent for training uses, patient-led governance of AI systems, and dynamic consent models that treat consent as an ongoing process rather than a one-time transaction.

Where the two source works converge, they do so on a common structural diagnosis: existing governance frameworks were designed for technologies that operate differently from generative AI systems, and the resulting mismatch produces both underprotection of patients and underspecification of clinician obligations. Where they diverge, the divergence is largely one of emphasis rather than substance. Sawesi et al. (2026) foreground technical risk categories and their alignment with medical device regulation; Larrauri et al. (2026) foreground consent, autonomy, and community governance. These emphases are complementary rather than competing. A comprehensive account of AI mental health governance must integrate the technical risk framework of Sawesi et al. (2026) with the consent and governance framework of Larrauri et al. (2026), because neither dimension is adequate on its own.

Case Scenario 1: Evaluating a Patient's Use of a Consumer Mental Health Chatbot

A licensed psychologist working in an outpatient practice meets with a new patient, a 34-year-old woman referred by her primary care physician for evaluation of persistent depressive symptoms. During the initial intake, the patient mentions that for the past three months she has been "talking to" a widely marketed AI chatbot on her phone "almost every night," describing it as "the only thing that gets me through." She reports having disclosed detailed histories of childhood trauma, current relationship distress, suicidal thoughts (which she describes as fleeting and without intent or plan), and use of prescription medications she obtained from a friend rather than a prescriber. She has not previously discussed the chatbot with any clinician, and she assumes that the interactions are "just like therapy" in terms of confidentiality.

Analysis. This scenario illustrates several intersecting themes from the reviewed evidence. The patient is engaging with a general-purpose LLM deployed through a consumer-facing platform, consistent with the modal use case identified by Sawesi et al. (2026). She has disclosed information across multiple high-sensitivity categories, including suicidal ideation, trauma, and unauthorized medication use, in a context where her expectations of confidentiality (Larrauri et al., 2026) almost certainly diverge from the actual data-handling practices of the platform. Her fleeting suicidal ideation raises the further question of how the chatbot responded, given the documented prevalence of crisis-response failures across the reviewed literature (Sawesi et al., 2026).

The clinical decisions this scenario requires are substantial. The psychologist must decide how to conduct informed consent for the ongoing therapeutic relationship in a way that addresses the patient's existing AI use, whether and how to discuss the privacy implications of her chatbot disclosures (recognizing that this discussion may itself be distressing), whether her reliance on the chatbot reflects an overreliance pattern that may be delaying appropriate care, and how to document these considerations in a manner that protects both the patient's welfare and the clinician's own defensibility. The evidence does not prescribe a single correct approach, but it does suggest that ignoring the AI dimension of the patient's presentation would be clinically incomplete. Consistent with the ethical principle of respect for autonomy, the psychologist's role is not to instruct the patient to stop using the chatbot but to ensure that her continued use is informed by an accurate understanding of what the technology can and cannot provide and of what happens to the information she discloses.

Psychological Pathways and Stress Responses

The clinical significance of generative AI mental health tools depends heavily on the psychological pathways through which their use affects users, particularly users experiencing psychiatric distress. Several such pathways emerge from the evidence.

The first pathway involves **disclosure and intimacy dynamics**. Sawesi et al. (2026) note that users may disclose to chatbots at levels comparable to human therapy sessions, and that intimacy and privacy expectations shift during extended chatbot interactions in ways that widen the gap between user assumptions and actual data governance. This disclosure pattern reflects well-documented psychological features of computer-mediated communication, including reduced perceived judgment and increased sense of anonymity. In the mental health context, these features may facilitate initial engagement (which can be therapeutic) but also amplify the exposure created by any privacy breach or secondary data use. From a stress-response perspective, users who develop a sense of confidential intimacy with a chatbot and subsequently learn that their disclosures have been used for model training, sold to third parties, or breached may experience acute distress that meaningfully affects their willingness to seek any future mental health support.

The second pathway involves **overreliance and help-seeking displacement**. Sawesi et al. (2026) report that overreliance was documented as a safety risk in 12.1% of reviewed studies and that users may develop emotional dependence on chatbot systems or substitute them for professional care. The mechanism is not mysterious: chatbots are continuously available, non-judgmental, and free (in most consumer implementations), and they respond in ways that many users experience as validating. For users experiencing depressive symptoms, social anxiety, or barriers to accessing professional care, these features can support a pattern of engagement that meets short-term emotional needs while displacing help-seeking behaviors that would offer more comprehensive or clinically appropriate support. The psychological pathway here is one in which the accessibility that makes AI tools attractive also enables patterns of use that delay engagement with human clinicians, sometimes with significant clinical cost.

The third pathway involves **crisis dynamics**. The evidence reviewed by Sawesi et al. (2026) documents failures in crisis response as one of the most prevalent safety risks (27.3% of studies). The clinical implication is that users in acute distress who turn to a chatbot may receive responses that fail to recognize the severity of the disclosure, that fail to provide appropriate escalation to emergency services or crisis lines, that provide generic reassurance rather than specific safety guidance, or (in some

documented cases) that provide content potentially reinforcing the crisis rather than mitigating it. From a stress-response perspective, an inadequate response during a crisis may deepen the sense of isolation and hopelessness that characterizes the crisis itself. A user who discloses suicidal ideation and receives a response perceived as dismissive, formulaic, or misattuned may be at greater risk than if they had not disclosed at all.

The fourth pathway involves **inferential exposure and its psychological consequences**. Because AI systems can infer psychological attributes and mental health status from patterns in language use (Larrauri et al., 2026), users may be identified as members of stigmatized categories on the basis of information they never explicitly disclosed. If this inferential capability is used, whether by employers, insurers, or advertisers, to make decisions affecting the user, the psychological consequences may include not only the material harms of adverse decisions but also the distinctive distress of being categorized on the basis of information one did not consciously share. This exposure creates a form of privacy loss that traditional confidentiality frameworks do not adequately address and that clinicians should be prepared to discuss with patients who use AI tools.

Case Scenario 2: Crisis Response and Chatbot Interaction

A psychologist receives an urgent call from the parents of a 17-year-old patient with a history of major depressive disorder and prior suicidal ideation. They have discovered messages on their son's phone showing that he has spent the past several hours in an extended conversation with a general-purpose AI chatbot, disclosing detailed suicidal ideation with plan. The chatbot's responses, as reviewed by the parents, included generic supportive statements, some references to breathing exercises, and a single mention of a crisis hotline embedded in a longer response. The chatbot did not initiate any escalation, did not repeatedly emphasize the crisis line, and did not respond to specific statements about method with any specific safety intervention. The patient is now safe with his parents but reports feeling that the chatbot "understood him" and is resistant to going to an emergency department.

Analysis. This scenario aligns closely with the crisis-response failures documented across the systematic review by Sawesi et al. (2026). The chatbot's response, as described, illustrates the pattern of inconsistent escalation, absence of appropriate referrals, or inadequate responses to suicidal ideation that appears in a substantial share of the reviewed literature. The clinical decisions here are urgent and multi-layered. Immediate safety planning takes priority, including assessment for potential emergency-department evaluation, removal of access to means, and mobilization of family support. Simultaneously, the psychologist must recognize the psychological significance of the patient's attachment to the chatbot: his statement that the chatbot "understood him" reflects the intimacy dynamic that Sawesi et al. (2026) and Larrauri et al. (2026) both describe, and treating that attachment as merely delusional or misguided is likely to increase resistance to appropriate care. The therapeutic task is to acknowledge the real experience of feeling understood while carefully articulating why the chatbot cannot substitute for the sustained, competent, and accountable response that his current situation requires. Documentation should capture both the crisis response and the AI-mediated dimensions of the presentation, because these considerations will bear on ongoing safety planning and family communication.

Mental Health Outcomes and Severity Spectrum

The mental health outcomes associated with generative AI use span a broad severity spectrum, from mild and often positive effects in low-stakes psychoeducational contexts to potentially catastrophic outcomes in high-stakes crisis contexts. Understanding this spectrum is important for clinicians because it informs how they should respond to different patterns of patient AI use.

At the lower-severity end, generative AI tools may support **psychoeducation, symptom monitoring, and skill practice** in ways that are broadly consistent with the goals of evidence-based care. For patients with mild-to-moderate symptoms who have access to appropriate clinical care and who use AI tools as adjuncts rather than substitutes, chatbots may provide accessible, non-judgmental engagement between sessions, facilitate cognitive restructuring practice, and support behavioral activation planning. Sawesi et al. (2026) note that psychoeducation and advice constituted a smaller but meaningful share of use cases (18.2%). In these lower-severity contexts, the risks documented across the reviewed literature (hallucination, bias, privacy concerns) remain relevant but their clinical consequences are typically less immediate.

At the middle of the severity spectrum, generative AI tools intersect with **moderate to severe symptoms in ways that create meaningful but variable clinical risk**. Patients with major depressive disorder, generalized anxiety disorder, post-traumatic stress disorder, or other conditions of moderate severity may use chatbots for emotional support in ways that provide subjective benefit while introducing risks of overreliance, exposure to biased or stigmatizing content, or reinforcement of maladaptive patterns. The clinical significance of chatbot use at this severity level often depends on whether the use is integrated with or substitutive of professional care.

At the higher-severity end, the outcomes associated with generative AI use become substantially more concerning. Sawesi et al. (2026) document that 78.8% of chatbot applications in the reviewed literature were classified as high risk, and that crisis and therapeutic engagement contexts dominated this classification. For patients experiencing suicidal ideation, acute psychotic symptoms, severe trauma responses, or other high-severity states, the documented prevalence of crisis-response failures, harmful advice, and hallucination creates a landscape in which chatbot interactions may produce clinically significant harm. The severity of potential harm here reflects both the acuity of the patient's condition and the documented limitations of current-generation systems.

An important complicating factor across the severity spectrum is **user vulnerability**. Larrauri et al. (2026) emphasize that mental health data is often generated during episodes of vulnerability and distress and is highly stigmatized, and that these features justify heightened consent standards. The same features shape outcomes: patients experiencing acute distress are typically less able to critically evaluate chatbot responses, more susceptible to overreliance dynamics, and more affected by any subsequent privacy breach. Vulnerability, in this sense, is not just a risk factor but a moderator of the relationship between AI use and outcomes.

Case Scenario 3: The Adolescent Patient and Boundaries of Chatbot Use

A psychologist provides care for a 15-year-old patient with generalized anxiety disorder and mild depressive symptoms. Over several sessions, the patient describes using an AI chatbot marketed to adolescents for "wellness support." She describes using it to practice social scripts before challenging conversations at school and to talk through her worries at bedtime. She reports that the chatbot's responses are usually helpful, though she notes that "sometimes it gives kind of weird advice" and that on one occasion the chatbot suggested she cut back on eating when she mentioned feeling anxious about her body. Her parents are aware of the chatbot but have not read its terms of service and are uncertain what information the chatbot collects.

Analysis. This scenario illustrates a middle-severity use pattern with genuine potential benefits (social skills rehearsal and cognitive processing) alongside identifiable risks (the documented pattern of bias and potentially harmful advice reported in the reviewed literature; the privacy concerns amplified by the patient's minor status; the parents' lack of awareness of data practices). The clinical response should recognize the therapeutic value the patient reports without minimizing the concerning content she has described. The chatbot's suggestion regarding eating is precisely the kind of harmful advice that Sawesi et al. (2026) documented as prevalent across the reviewed literature, and it is particularly concerning given the developmental context and the specific vulnerability of adolescents to disordered eating content. From an ethical standpoint, the psychologist has an obligation to ensure that the patient and her parents make informed decisions about ongoing chatbot use, including discussion of the platform's data practices where these can be determined and consideration of whether continued use is appropriate given the specific content of concern.

System-Level and Contextual Psychological Effects

Beyond individual clinical encounters, the widespread deployment of generative AI in mental health contexts produces system-level effects that shape the environment in which clinicians and patients operate. Several such effects merit consideration.

The first is the **normalization of consumer-grade mental health disclosure**. As Sawesi et al. (2026) document, nearly half of the reviewed literature examined consumer-facing deployments. As AI-mediated mental health engagement becomes more common, patients enter clinical encounters having already established patterns of disclosure to non-clinical platforms. This changes what patients bring to sessions, what they expect from clinicians, and how they conceptualize the boundaries of mental health support. It also creates a category of prior disclosures that exist outside the clinical record but that may bear on the clinical presentation. Clinicians

who ignore this dimension may miss important information about their patients' current functioning and prior help-seeking behavior.

The second is the **shift in expectations regarding accessibility and availability**. Chatbots are continuously available; human clinicians are not. As patients become accustomed to 24-hour engagement with AI-mediated support, expectations regarding clinician availability, response time, and between-session accessibility may shift in ways that affect treatment frames, boundary management, and clinician workload. This shift is not intrinsically problematic, but it requires that clinicians be explicit about the frame of their own availability and about the differences between the sort of support they can provide and the sort of engagement patients may have come to expect from AI tools.

The third is the **regulatory and governance environment itself as a psychological factor**. Larrauri et al. (2026) note that on January 6, 2026, the FDA issued guidance documents suggesting a willingness to classify more AI applications as "wellness" devices or software outside the scope of formal regulation. This regulatory posture, along with the broader pattern of regulatory gaps documented by Sawesi et al. (2026), creates an environment in which patients cannot reasonably rely on external oversight to protect them from harmful or exploitative practices. The psychological consequence, for both patients and clinicians, is that the burden of navigating the risks associated with AI use falls disproportionately on individual decision-making, without the structural safeguards that ordinarily protect patients in other clinical contexts.

The fourth is the **workforce and equity dimension**. Sawesi et al. (2026) note that only one of 33 reviewed studies was conducted in a lower-resource setting. The populations most likely to face barriers to accessing human clinical care, including those in low-resource settings, uninsured populations, adolescents, and members of stigmatized groups, are also those most likely to turn to consumer AI tools as a primary source of mental health support. This creates a stratified access environment in which the highest-quality, most accountable mental health care remains available primarily to those with resources, while consumer AI tools, with all their documented limitations, disproportionately serve those without. Larrauri et al. (2026) note the risk of repeating historical patterns in which marginalized populations contribute to medical progress without sharing in its benefits or protections. From a public health perspective, this is a significant equity concern that shapes the clinical and ethical context in which individual clinicians make decisions.

Clinical Implications for Psychologists

The implications of the reviewed evidence for clinical practice are substantial and span multiple domains of routine work.

Assessment and Formulation

Psychologists should routinely inquire about patients' use of AI-mediated mental health tools as part of intake assessment and ongoing case formulation. The dominance of consumer-facing deployment documented by Sawesi et al. (2026) means that a substantial share of patients presenting for care will have current or recent AI engagement that may bear on their presentation. Relevant inquiry includes the nature of the tools used, the frequency and duration of engagement, the content typically discussed, the patient's subjective experience of the interactions, any concerning responses received, and the patient's understanding of the platform's data practices. This inquiry serves multiple purposes: it captures information relevant to symptom presentation, it identifies potential safety concerns (such as inadequate crisis responses), it opens a therapeutic conversation about the role of AI tools in the patient's overall care, and it provides an opportunity to address misunderstandings about privacy and confidentiality.

Treatment Planning

Where patients continue to use AI tools alongside clinical care, treatment planning should explicitly address the integration or coordination between these modalities. In some cases, chatbot use may support treatment goals; in others, it may undermine them by displacing help-seeking, reinforcing maladaptive patterns, or introducing content inconsistent with the treatment approach. The clinical decision is not typically whether to permit or prohibit chatbot use (a decision that in most cases lies outside the clinician's authority) but how to work with the patient's existing use in a way that supports treatment goals. This may include discussion of what topics are appropriate for chatbot engagement versus what topics warrant clinical attention, guidance about recognizing crisis-relevant content and responding appropriately, and, in some cases, recommendations to reduce chatbot use where its clinical effects appear net negative.

Documentation

Documentation practices should reflect the increased complexity that AI tools introduce. Where relevant, clinical documentation should capture patients' AI use, the clinician's discussion of AI-related risks and benefits, informed consent discussions covering the boundaries of the therapeutic frame in the context of AI adjuncts, and any safety planning that addresses AI-mediated crisis-response failures. Documentation should also reflect that patients' prior AI disclosures constitute a category of information relevant to care but outside the clinician's direct knowledge, which may bear on the completeness of the clinical record.

Referral and Coordination

Psychologists should be prepared to distinguish between AI tools that may be reasonably integrated into a broader care plan and those that raise sufficient concerns to warrant explicit clinical guidance about limiting use. This distinction is not always straightforward given the limited empirical evidence base, but the reviewed literature suggests that particular caution is warranted regarding tools used in high-risk contexts (crisis, suicide risk, severe symptoms), tools with documented histories of harmful responses, and tools with unclear or problematic data-handling practices.

Professional Development

The rapid evolution of AI tools in mental health creates a continuing obligation for psychologists to maintain reasonable competence regarding the technologies their patients use and the risks these technologies present. This is not an obligation to become a technical expert in machine learning; it is an obligation to maintain sufficient understanding to conduct informed consent discussions, recognize relevant clinical concerns, and provide appropriate guidance. Continuing education, engagement with professional guidance documents, and consultation with colleagues who have developed expertise in this area all contribute to fulfilling this obligation.

Ethical Considerations in Practice

The ethical dimensions of generative AI mental health tools intersect with essentially every major principle of psychological practice. A systematic consideration of these principles illustrates the scope of the ethical landscape.

Competence

The principle of competence requires that psychologists provide services within the boundaries of their training, experience, and continuing professional development. As AI tools become more prevalent in patients' lives, competence in evaluating and discussing these tools becomes part of the general competence required for practice, at least at the level necessary to conduct informed clinical conversations. Clinicians who lack understanding of the risks documented by Sawesi et al. (2026) and the consent limitations analyzed by Larrauri et al. (2026) may find themselves unable to provide the kind of informed guidance their patients need. The scope-of-practice implication is not that psychologists must become AI experts but that they must acquire and maintain reasonable working knowledge of the AI landscape as it affects the populations they serve.

Informed Consent

The principle of informed consent, well-developed as a foundation of psychological practice, extends in the AI context in two directions. First, it applies to the clinician's own use of AI tools in practice: any use of AI-mediated services, AI-assisted documentation, AI-supported assessment, or AI-based clinical decision support requires appropriate informed consent from patients, including discussion of what information will be processed by AI systems, what data will be stored and where, and what limitations apply. Second, and consistent with Larrauri et al. (2026), the informed consent conversation between clinicians and patients should extend to patients' independent use of AI tools, at least to the extent necessary to ensure that patients understand the difference between the confidentiality protections that govern the clinical relationship and those (or their absence) that govern their chatbot use.

Confidentiality

The principle of confidentiality is stressed in fundamental ways by generative AI tools. Sawesi et al. (2026) document that users frequently assume healthcare-grade confidentiality when interacting with consumer platforms that provide no such protection. This gap between expectation and reality has direct clinical implications: patients may disclose to clinicians information they previously disclosed to a chatbot, believing that both disclosures were similarly protected, when in fact the chatbot disclosure may already have been logged, analyzed, retained, or reused in ways the patient does not understand. Clinicians have an ethical obligation to help patients understand this asymmetry, not to shame or blame them for prior disclosures but to support informed decision-making going forward.

Multiple Relationships and Boundaries

Although AI tools do not create multiple relationships in the traditional sense, they raise analogous boundary considerations. When a patient's primary emotional engagement is with a chatbot rather than with the clinician, or when the clinician's practice becomes intertwined with specific AI tools through employer or health system arrangements, the frame of the therapeutic relationship may shift in ways that warrant explicit attention. Clinicians should be particularly attentive to arrangements in which their practice is contractually linked to specific AI platforms, given the potential for conflicts of interest to arise between clinical judgment and platform incentives.

Cultural Responsiveness

The equity dimension of AI mental health deployment has particular ethical significance. The documented concentration of research in high-income settings, the disproportionate reliance on consumer AI tools among populations facing barriers to human clinical care, and the potential for AI systems to reproduce or amplify existing biases all create obligations for culturally responsive practice. Clinicians should be attentive to how AI tools may work differently for different populations, to the possibility that a tool that appears clinically neutral in aggregate may produce biased outcomes for members of specific groups, and to the broader equity implications of a stratified access environment.

Scope of Practice

Finally, the ethical principle of practicing within scope of practice takes on new dimensions when AI tools blur the boundaries between clinical care and consumer support. Clinicians should not represent AI tools as clinical services they are not, should not endorse or recommend tools they lack the basis to evaluate, and should be clear with patients about the distinction between the sustained, competent, accountable relationship they can offer and the interactions available through consumer AI platforms.

Case Scenario 4: The Employer-Sponsored Platform

A licensed psychologist is approached by a mid-size employer seeking to contract for staff mental health services. As part of the arrangement, the employer indicates that all employees will have first-line access to a generative AI chatbot integrated into the employee assistance program, with referrals to human clinicians (including the psychologist) available for cases the chatbot flags. The employer explains that the chatbot vendor uses de-identified transcripts to improve its models, and that continued use of the platform requires acceptance of the vendor's terms of service.

Analysis. This scenario illustrates several of the structural concerns raised by Larrauri et al. (2026), including the compromise of voluntariness when employees must accept platform terms to access care, the reliance on de-identification without acknowledgment of re-identification risk, and the potential for employer or insurer inferences to affect employment decisions. From a professional ethics perspective, the psychologist entering this arrangement takes on obligations that extend beyond the individual therapeutic relationship. These include ensuring that referrals from the chatbot to the clinician preserve appropriate informed consent, that the psychologist's own practice does not inadvertently validate representations made by the chatbot that the psychologist could not independently support, and that the psychologist maintains clinical autonomy despite the contractual link with the platform. In some cases, these obligations may require declining the arrangement or negotiating substantial modifications; in others, they may require careful documentation and ongoing vigilance. The scenario illustrates why patient-led governance, as advocated by Larrauri et al. (2026), matters at a system level: individual clinicians navigating arrangements like this one are more likely to succeed when broader governance structures constrain the range of arrangements employers and vendors can propose.

Practice Recommendations

The following practice recommendations synthesize the reviewed evidence into concrete guidance for clinicians. Recommendations are distinguished by their level of support in the underlying literature: **well-supported** recommendations are directly grounded in the evidence reviewed by Sawesi et al. (2026) and the analysis of Larrauri et al. (2026); **tentative** recommendations extend from that evidence but require further empirical development.

Well-Supported Recommendations

Assess for AI tool use as a routine component of intake and ongoing evaluation. Given the prevalence of consumer-facing AI deployment documented by Sawesi et al. (2026), assessment of AI use should be routine rather than exceptional. Inquiry should address the nature, frequency, and content of use as well as the patient's subjective experience.

Address the confidentiality mismatch explicitly in informed consent conversations. Consistent with Larrauri et al. (2026), patients frequently assume healthcare-grade confidentiality in contexts where none applies. Explicit discussion of this asymmetry supports informed decision-making.

Recognize crisis-response limitations as a specific safety concern. Given the prevalence of crisis-response failures documented by Sawesi et al. (2026), patients at risk for suicidal ideation or other crises should receive specific guidance about not relying on chatbots for crisis support and about the appropriate use of crisis lines, emergency services, and clinician availability.

Document AI-related discussions and considerations. Given the clinical significance of AI use, documentation should reflect the clinician's assessment, discussion, and clinical decisions related to patients' AI engagement.

Maintain ongoing professional development regarding AI tools. Given the pace of evolution in this area, continuing education is required to maintain the working knowledge necessary for competent practice.

Tentative Recommendations

Consider adopting or advocating for dynamic consent models in any AI-integrated practice arrangements. Larrauri et al. (2026) argue that consent should be reconceptualized as an ongoing process. While specific implementation models are still developing, clinicians can support this shift by seeking arrangements that provide patients with ongoing rather than one-time consent options.

Support patient-led governance where opportunities arise. Larrauri et al. (2026) argue for embedding patient representation in AI governance structures. Clinicians can support this by referring patients to advocacy organizations engaged in AI governance work and by advocating within their own institutions for governance structures that include patient voice.

Approach vendor claims about safety, security, and clinical validity with substantial skepticism. Given the methodological limitations documented by Sawesi et al. (2026), particularly the rarity of formal adversarial testing, vendor claims should be evaluated against the standard of independent, adversarial evaluation.

Advocate for regulatory alignment with the multi-layered risk framework. The framework advanced by Sawesi et al. (2026) locates risk across model, interface, data, and infrastructure layers. Clinicians positioned to influence institutional or regulatory policy can support alignment of oversight structures with this multi-layered approach.

Synthesis Table of Practice Recommendations

Domain	Recommendation	Level of Support
Assessment	Routine inquiry about AI tool use	Well-supported
Informed consent	Explicit discussion of confidentiality mismatch	Well-supported
Safety planning	Specific guidance about crisis-response limitations	Well-supported
Documentation	Documentation of AI-related discussions and decisions	Well-supported
Professional development	Ongoing education regarding AI tools	Well-supported
Consent frameworks	Adoption or advocacy for dynamic consent models	Tentative
Governance	Support for patient-led governance structures	Tentative
Evaluation	Skepticism toward vendor safety and validity claims	Tentative
Regulatory engagement	Advocacy for multi-layered risk oversight	Tentative

Note. Synthesized from Sawesi et al. (2026) and Larrauri et al. (2026).

Future Directions and Research Gaps

The reviewed evidence identifies several priority areas for future research and for the development of clinical practice. First, the empirical evaluation of generative AI mental health tools remains substantially underdeveloped, particularly with respect to real-world adversarial testing, longitudinal deployment data, and effects on populations underrepresented in existing research (Sawesi et al., 2026). Second, the development of consent frameworks capable of tracking the actual operational features of AI systems, including dynamic consent, meta consent, and patient-led governance structures, remains largely aspirational despite conceptual maturity (Larrauri et al., 2026). Third, the integration of technical risk frameworks with clinical, ethical, and regulatory approaches requires interdisciplinary work that has been slow to develop.

Specific research gaps include the effects of chatbot use on help-seeking behavior in populations with barriers to human care, the moderating role of user vulnerability in the relationship between AI use and outcomes, the long-term consequences of privacy breaches involving mental health data, the effectiveness of different consent architectures in supporting comprehension and voluntariness, and the effects of AI-mediated care on therapeutic alliance and treatment outcomes in traditional clinical settings. Progress in these areas will require sustained investment in independent research, methodological development capable of accommodating the probabilistic nature of AI systems, and structural support for research that does not depend on vendor cooperation.

Conclusion

Generative artificial intelligence has moved into mental health service delivery faster than the frameworks required to govern it have evolved. The systematic evidence reviewed by Sawesi et al. (2026) documents a multi-layered risk landscape spanning cybersecurity, privacy, and safety concerns across model, interface, data, and infrastructure levels; the conceptual analysis provided by Larrauri et al. (2026) documents a parallel crisis in informed consent, in which the contractual assent regimes governing consumer platforms fall well short of the ethical standards that ordinarily protect patient autonomy and confidentiality. Together, these bodies of work describe a field in which the operational logic of the technology and the operational logic of governance are meaningfully misaligned.

For licensed psychologists, this misalignment translates into concrete obligations. It requires that clinicians develop working knowledge of the AI tools their patients use, that they conduct informed consent conversations extending beyond the traditional therapeutic frame to address AI-related exposures, that they recognize crisis-response failures and other documented risks as specific safety concerns, and that they document their clinical reasoning in ways that reflect the increased complexity AI tools introduce. It also requires that clinicians engage, where opportunities arise, with the broader governance conversations that will ultimately determine whether patients receive the structural protections their autonomy and welfare require.

The parallel between the exploitation that produced HeLa cells and the current appropriation of patient data for AI training, drawn by Larrauri et al. (2026), is not merely rhetorical. It is a reminder that medical innovation built on inadequate protection of patient autonomy is a recurring pattern, and that recognizing the pattern in real time is the precondition for changing it. The safe integration of generative AI into mental health services will depend not only on improving the technology but on developing governance frameworks that reflect the unique risks of conversational systems and on maintaining the ethical commitments that have long distinguished the mental health professions. Clinicians who take up this work will be better positioned to serve their patients well in an environment that is rapidly changing around them.

References

- Larrauri, C. A., Martinez-Martin, N., & Torous, J. (2026). Reclaiming informed consent to train mental health AI with patient data. *npj Digital Medicine*, 9, 423. <https://doi.org/10.1038/s41746-026-02843-8>
- Sawesi, S., Sabbineni, H., Shagamreddy, R. R., & Rashrash, B. (2026). Cybersecurity and privacy risks of generative AI mental-health chatbots: A systematic review and regulatory framework. *Journal of Multidisciplinary Healthcare*, 19, 581251. <https://doi.org/10.2147/JMDH.S581251>