Arora C, 'Logged, listened, and legally ignored: the mirage of privacy in AI therapy' (august 2025) EJPLT online News.
Available at: https://universitypress.unisob.na.it/ojs/index.php/ejplt/announcement

EJPLT
ISSN 2704-8012



Logged, listened, and legally ignored: the mirage of privacy in AI therapy

Registrato, ascoltato e legalmente ignorato: il miraggio della privacy nella terapia dell'IA

CHIRANSHU ARORA
Qualified Legal Professional, LLM in Intellectual Property

Abstract

At 2 a.m., when conventional support systems become inactive, users increasingly rely on AI mental health chatbots like Replika, Wysa, and Woebot. These platforms offer anonymity, empathy, and accessibility; but, beneath their reassuring interfaces exists a largely unexamined risk: membership inference attacks. This article explores the probing of machine learning models utilised in mental health applications to statistically determine if an individual's data was included in the training set, without accessing any actual content. The article critically assesses the deficiencies in existing privacy safeguards and the ethical vulnerabilities of AI-driven therapy, utilising worldwide legal frameworks, empirical data, and significant case studies. It advocates for the immediate legal recognition of inferred data, the implementation of inference-aware privacy standards, and enhanced transparency measures. This study argues that as mental health care becomes more digitised, protecting users' dignity necessitates not only technological advancement but also ethical design and regulatory foresight that acknowledge the complexities of the human mind.

© The author(s) 2025, published by Suor Orsola Benincasa Università Editrice.
This contribution is licensed under a Creative Common Attribution 4.0 International Licence CC-BY-NC-ND, all the details on the license are available at:
https://creativecommons.org/licenses/by-nc-nd/4.0/

Abstract
Alle 2 di notte, quando i sistemi di supporto convenzionali diventano inattivi, gli utenti si affidano sempre più spesso a chatbot di intelligenza artificiale per la salute mentale come Replika, Wysa e Woebot. Queste piattaforme offrono anonimato, empatia e accessibilità, ma sotto le loro interfacce rassicuranti esiste un rischio in gran parte non esaminato: gli attacchi di inferenza associativa. Questo articolo esplora il sondaggio dei modelli di apprendimento automatico utilizzati nelle applicazioni per la salute mentale per determinare statisticamente se i dati di un individuo sono stati inclusi nel set di addestramento, senza accedere a nessun contenuto effettivo. L'articolo valuta criticamente le carenze delle attuali tutele della privacy e le vulnerabilità etiche della terapia guidata dall'IA, utilizzando quadri giuridici mondiali, dati empirici e casi di studio significativi. Si chiede l'immediato riconoscimento legale dei dati dedotti, l'implementazione di standard di privacy consapevoli dell'inferenza e il potenziamento delle misure di trasparenza. Questo studio sostiene che, con la crescente digitalizzazione dell'assistenza per la salute mentale, la protezione della dignità degli utenti richiede non solo un progresso tecnologico, ma anche una progettazione etica e una previsione normativa che riconoscano le complessità della mente umana.

Keywords: Mental Health; Data Privacy; Artificial Intelligence (AI); AI Mental Health Applications; Membership Inference Attacks; Sensitive Data

Summary: 1. Introduction. – 2. The illusion of safe spaces: the AI mental health apps. – 3. Legal grey areas: when algorithms outspace law. – 4. When privacy fails in silence. – 5. The psychological costs behind ethics. – 6. Are technical solutions enough? – 7. Recommendations. – 8. Conclusion: we are all test subjects now.

1. Introduction.

Mental health has evolved from a murmur in the corner to a headline act, and AI has seized the stage with an unexpected grace. Millions of individuals have found solace in chatbots such as Replika, Wysa, and Woebot**,** which have become their companions at 2 a.m. when human therapists are unavailable.[1] These digital tools are designed to be always-on emotional first responders, and they are expected to provide convenience, compassion, and confidentiality. However, a subtle technical vulnerability resides beneath their tranquil interface: it is feasible for an individual to statistically determine whether you have ever engaged with one without examining a single word of your interaction. Welcome to the unsettling realm of membership inference attacks.

Imagine that you are in the midst of a difficult night and you open an application. You disclose that you are grappling with anxiety, and there is a

[1] K K Fitzgerald, A Darcy and M Vierhile, 'The Delivery of Cognitive Behaviour Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomised Controlled Trial' (2017) 4(2) *JMIR Mental Health* e19 https://mental.jmir.org/2017/2/e19/ accessed 28 May 2025 doi: 10.2196/mental.7785.

possibility that you are contemplating suicide. A data-savvy employer or malignant hacker deduces that you interacted with that chatbot weeks later, despite never having accessed the content. In what manner? By conducting a probe of the AI model to determine whether your data was included in its training set.

This is a membership inference attack, a highly technological and emergent method of leaking identity via machine learning models. This is akin to an individual determining that you have undergone therapy by examining the imprints outside the clinic; however, this time, the footprints are algorithmic. *Shokri et al. (2017)* were among the first to empirically demonstrate the vulnerability of even well-known models to this attack.[2]

2. The illusion of safe spaces: the AI mental health apps.

It is claimed by Wysa to be your anonymous companion for well-being. Replika aspires to serve as your emotional support assistant. Woebot asserts that it can assist in stress management through conversations that are founded on cognitive behavioural therapy (CBT). They all convey a warm, fuzzy image of ethical AI and data protection. However, how many of them genuinely employ differential privacy? How many organisations engage in red-teaming procedures or undertake external audits?

Replika experienced a significant increase in usage during the pandemic, particularly due to its emotionally intelligent conversations. It was temporarily banned in Italy in 2023 for allegedly permitting sexually explicit content directed at minors.[3] Italy then imposed a €5 million fine on Replika in 2025 for GDPR breaches, underscoring the ongoing concerns regarding the privacy dangers associated with AI.[4] That is not merely an algorithm that has gone rogue; it is a failure of regulatory foresight. Similarly, *BetterHelp,* a mental health platform providing online counselling and therapy services, came under fire in 2023 after it was revealed that users' private health information was shared with third companies including Facebook and Snapchat. This raised severe concerns about privacy issues and lack of informed consent. In a subsequent settlement with the FTC, the business consented to pay $7.8 million to resolve allegations related to these data sharing practices.[5]

In 2022, *Talkspace*, a well-known virtual counselling platform, experienced a data breach that compromised the confidential correspondence, session plans,

---

[2] R Shokri, M Stronati, C Song and V Shmatikov, 'Membership Inference Attacks Against Machine Learning Models' (2017) *IEEE Symposium on Security and Privacy (SP)* 3, doi: 10.1109/SP.2017.41.

[3] Garante per la Protezione dei Dati Personali, 'Provision of 2 February 2023 – Replika' (English version) (2 February 2023) Doc. web no. 9852214 https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9852214#english accessed 27 May 2025.

[4] European Data Protection Board, 'AI: the Italian Supervisory Authority fines company behind chatbot "Replika"' (21 May 2025) https://www.edpb.europa.eu/news/national-news/2025/ai-italian-supervisory-authority-fines-company-behind-chatbot-replika_en accessed 27 May 2025.

[5] Federal Trade Commission, 'FTC Bans BetterHelp From Revealing Consumers' Data, Including Sensitive Mental Health Information, to Facebook' (FTC, 16 March 2023) https://web.archive.org/web/20231025073635/https://www.ftc.gov/news-events/news/press-releases/2023/03/ftc-ban-betterhelp-revealing-consumers-data-including-sensitive-mental-health-information-facebook accessed 03 June 2025.

and payment records of users[6] - highlighting the potential risks that exist even when human therapists are involved. That same year, the Mozilla Foundation identified *Youper*, another AI mental health app, as having one of the worst privacy policies.[7] The app collected a lot of personal data, including location, data from wearable sensors, and even information from other apps. It also had provisions for sharing this data with affiliates and third parties for marketing purposes.[8] Nevertheless, by 2023, Youper had accomplished great strides in minimising data sharing and enhancing password requirements. It was hailed as one of the most improved apps, but it continues to face criticism for prompting users to answer sensitive questions before they had the chance to review its privacy policy.[9]

In 2024, the critical necessity of rigorous security measures for data in AI mental health platforms was underscored by the exposure of over 5.3 TB of confidential patient data by *Confidant Health*, an AI-powered mental health service, as a result of a compromised server.[10] The potential consequences of employing such a paradigm in therapy, even hypothetically, could be irreparable. The industry appears to be more reactive than proactive, consistently falling two steps behind the issue.[11]

3. Legal grey areas: when algorithms outpace laws.

In 2024, the European Union's AI Act was formally adopted, establishing the world's first comprehensive legal framework on artificial intelligence. It classifies mental health chatbots as "high-risk AI systems" if they assert to offer psychological or therapeutic services. This results in them being subjected to more stringent responsibilities, including transparency, human supervision, and risk management. Nevertheless, membership inference attacks are still not explicitly addressed, particularly in the context of inferred sensitive data from anonymised or synthetic datasets. The AI Act does not explicitly mandate differential privacy or prohibit data re-identification practices. It does, however, include provisions for post-market monitoring, which may enable regulators to respond in the event that inference damages are discovered. In general, the European Data Protection Board (EDPB) has acknowledged inference hazards, but the legal interpretation in the context of LLMs and chatbots is still in the process of evolving.

HIPAA is applicable exclusively to "covered entities" in the United States,

[6] K Hill, 'Talkspace, a Therapy App, Faces Questions About Client Privacy' The New York Times (New York, 7 August 2020) https://www.nytimes.com/2020/08/07/technology/talkspace.html accessed 02 June 2025.
[7] Mozilla Foundation, *Youper* (Privacy Not Included, 2023) https://www.mozillafoundation.org/en/privacynotincluded/youper/ accessed 04 June 2025.
[8] Ibid.
[9] Ibid.
[10] A Bîzgă, 'Confidant Health Data Breach: Thousands of Patient Therapy Sessions and Patient's PII Exposed in an Unsecured Database Online' (Bitdefender, 9 September 2024) https://www.bitdefender.com/en-us/blog/hotforsecurity/confidant-health-data-breach-thousands-of-patient-therapy-sessions-and-patients-pii-exposed-in-an-unsecured-database-online accessed 29 May 2025.
[11] L Gatt, IA Caggiano, 'Consumers and Digital Environments as a Structural Vulnerability Relationship' (2022) 2 European Journal of Privacy Law & Technologies 8–16 https://universitypress.unisob.na.it/ojs/index.php/ejplt/article/view/1724 accessed 22 July 2025.

such as insurers or institutions, and does not apply to independent mental health applications. This is the reason why Flo Health was able to share period data with Facebook and Google without any criminal breach—only a minor reprimand from the FTC in 2021. In response to incidents in which chatbots offered detrimental advice, including the encouragement of self-harm, California introduced legislation in 2025 that sought to prohibit AI systems from masquerading as licensed health professionals.

As of 2025, the Privacy Act 1988 of Australia does not contain explicit provisions for AI model inference risks, despite the fact that it does provide protections for health information. The ACCC's Digital Platform Services Inquiry (2025) in Australia did raise concerns regarding opaque AI profiling; however, probabilistic privacy breaches, such as membership inference, have yet to be addressed. Nevertheless, the Office of the Australian Information Commissioner (OAIC) issued guidance in October 2024, which underscored the necessity for organisations to guarantee the accuracy and impartiality of AI products that manage personal information. Furthermore, effective June 2025, a statutory tort for severe invasions of privacy will enable individuals to file lawsuits for substantial privacy breaches, such as unauthorised surveillance.

The Artificial Intelligence and Data Act (AIDA) of Canada, which is a component of the more extensive Bill C-27, is presently undergoing parliamentary review and is anticipated to be passed in late 2025. AIDA is designed to regulate high-impact AI systems, including mental health chatbots. Conversely, there are no existing legal obligations regarding model memorisation, membership inference, or the reverse engineering of sensitive disclosures, and the language surrounding inferred data is restricted.

According to India's Digital Personal Data Protection Act 2023, mental health data is considered sensitive personal data. As of 2025, the Act is being progressively enforced, allowing users to access information regarding the use of their data, such as summaries of processing activities and third-party disclosures. Even so, the terminology is ambiguous when it comes to AI model outputs or inferences, and its implementation remains inconsistent.

4. When privacy fails in silence.

Imagine a lady who talks about postnatal depression with a chatbot. Months later, during a custody dispute, her ex's lawyer runs a model trained on chatbot data and proves that she was involved. There is no leaked interaction, simply proof that she used the bot at a certain time. No rules were breached. But what about her privacy? Completely violated.

This isn't simply speculation. Researchers from the University of Texas at Austin were able to run membership inference attacks on text classification models that had been trained on Reddit mental health forums in 2020.[12] In the real world, that may mean producing a list of individuals who are likely to have PTSD, depression, or substance misuse only by looking at model outputs, without having to break into any databases.

[12] Y Yang, P Gohari, U Topcu, *'On the Vulnerability of Recurrent Neural Networks to Membership Inference Attacks'* (University of Texas at Austin, 2021) https://arxiv.org/pdf/2110.03054v1 accessed 05 June 2025.

The U.S. startup *Cerebral* got a lot of bad press after a data breach in 2023 revealed that it had been using invisible pixel trackers from Google, Meta (Facebook), TikTok, and other third parties on its online services since October 2019, compromising user privacy. A $7 million settlement was struck between Cerebral and the Federation Trade Commission in April 2024.[13]

It's not just revealing your identity that is a quiet threat; it's also subtle exclusion. Think about what it would be like if your insurance company knew you were using a chatbot for depression but never told you why your premiums went up. This hidden cost is algorithmic discrimination by omission, which is a type of harm that inference attacks can do by quietly shifting power without letting people know or giving them a way to fight it.

For users who are already on the fringes, the stakes are considerably higher.[14] A leaked model encounter might put a closeted kid in a conservative space or a domestic abuse survivor's safety at risk. Inference threats don't simply put data at risk; they put lives at risk as well, making people more vulnerable in ways that are frequently not seen until harm ensues.

5. The psychological cost beyond ethics.

Let's be real. Mental health care is about being open and honest. If individuals start to doubt their digital therapists, we might witness a chilling effect, where people don't ask for treatment because they're scared. That's a steady, undetectable loss of trust that has severe effects on mental health. And it's already happening.

In 2022, a Reddit thread about Woebot had ratings of individuals saying they didn't like the app's ambiguous terms of privacy. One person said, “I felt safer writing in Google Docs than talking to Woebot.”[15] That single statement should scare developers.

Also, consent isn't real if consumers aren't educated about these hazards in plain terms. No one clicks "I agree" thinking that their usage will be able to be statistically extracted later. If people don't know what they're agreeing to, it's just legal theatre.

In 2024, JMIR Mental Health released a study that asked people in Australia and mental health experts about using AI in mental health care. The study indicated that nearly half of the community members (46.7%) and more than half of the mental health professionals (51.4%) who used AI faced risks or

---

[13] Federal Trade Commission, *Proposed FTC Order Will Prohibit Telehealth Firm Cerebral from Using or Disclosing Sensitive Data* (25 April 2024) https://www.ftc.gov/news-events/news/press-releases/2024/04/proposed-ftc-order-will-prohibit-telehealth-firm-cerebral-using-or-disclosing-sensitive-data accessed 06 June 2025.

[14] L Gatt, ‘Legal anthropocentrism between nature and technology: the new vulnerability of human beings’ (2022) 1 European Journal of Privacy Law & Technologies 15–26 https://universitypress.unisob.na.it/ojs/index.php/ejplt/article/view/1636 accessed 22 July 2025.

[15] Reddit user, *‘Mental health apps have terrible privacy’ (Reddit, 2022)* https://www.reddit.com/r/technology/comments/uhi1ic/mental_health_apps_have_terrible_privacy/ accessed 05 June 2025.

downsides, such as concerns with data privacy, ethical use, possible misdiagnosis, unhelpful advice, and less personal connection.[16]

A survey of 1,000 college students who used Replika in 2024 revealed that students are especially prone to loneliness and less inclined to ask for help from other people. This shows that they are becoming more reliant on AI companions.[17]

6. Are technical solution enough?

Yes, in theory. Differential privacy adds noise to model outputs to stop someone from being able to perfectly recreate the data. Federated learning lets you train without having to collect data in one place. Model unlearning might let consumers ask for data to be deleted after training.[18]

In reality, these are costly, complicated, and not often put to use. As an instance, Replika doesn't make public their privacy audits or model design. When venture capital is flowing and user growth is king, data safety is usually not a priority for startups. Many AI health apps include business models that incentivise people to save data. User interactions turn into training fuel, a way to get ahead of the competition, and ultimately revenue. Rather than being an asset, forgetting becomes a burden in these types of economies.

OpenAI even got in trouble for keeping user prompts, and only added an opt-out option months after the outrage.[19] And this company has what seems to be the most resources and prominence. When even the most elite fail, what hope do we have for the hundreds of AI mental health apps that appear in the App Store monthly?

7. Reccomendations.

1. *Explicit laws on inference attacks:*

Legislation that specifically acknowledges membership inference as a unique type of exploitation of personal data is desperately needed, especially when it comes to sensitive health information. When it comes to health or mental well-being, current rules such as the GDPR[20] or Australia's Privacy Act[21] need to change to designate AI model involvement as sensitive data

[16] S Cross and others, 'Use of AI in Mental Health Care: Community and Mental Health Professionals Survey' (2024) 11 *JMIR Mental Health* e60589 https://mental.jmir.org/2024/1/e60589 accessed 08 June 2025.
[17] B Maples, M Cerit, A Vishwanath and others, 'Loneliness and Suicide Mitigation for Students Using GPT3-Enabled Chatbots' (2024) 3 *npj Mental Health Research* 4 https://doi.org/10.1038/s44184-023-00047-6
[18] Lucilla Gatt et al, 'BCI Devices and Their Legal Compliance: A Prototype Tool for Its Evaluation and Measurement' (2022) 2(1) European Journal of Privacy Law & Technologies https://universitypress.unisob.na.it/ojs/index.php/ejplt/article/view/1640 accessed 22 July 2025.
[19] OpenAI, *New Ways to Manage Your Data in ChatGPT* (OpenAI, 2023) https://openai.com/index/new-ways-to-manage-your-data-in-chatgpt/ accessed 07 June 2025.
[20] General Data Protection Regulation (GDPR), Regulation (EU) 2016/679.
[21] *Privacy Act 1988 (Cth), Australian Government.*

processing. Using precedents like Schrems II[22] and FTC v. Flo Health[23], which highlighted the significance of data misuse through inference and indirect collecting, courts could expand their interpretations.

In order to prevent AI systems trained in one jurisdiction from disclosing private information in another, lawmakers ought to create cross-border inference safeguards and mandate inference-specific breach notification rules. To enhance oversight and enforcement, regulatory agencies should establish specialised AI data protection departments.

2. *Inference-Ready Definitions of Sensitive Data:*

In addition to officially collected data, regulatory frameworks should incorporate inferred data as well. Research on membership inference attacks[24] has shown that models can reveal whether a user was a part of the training data, which is a delicate topic in general but particularly so when discussing mental health issues.

Although the US Office for Civil Rights' position on health inference[25], India's Digital Personal Data Protection Act, 2023[26], and the EU's EDPB guidelines[27] all make inroads towards acknowledging the risks of inference, more precise safeguards are required. To avoid gaps where assumed identities or situations are not regulated, the idea of "derivable personal data" should be formalised.

3. *Transparent User Logs and Consent Mechanisms:*

There is currently no mental health chatbot that gives users the keys to see how their data changes the product, but the blueprint does exist. Railsware provides internal dashboards that connect user feedback with feature development.[28] Think about what would happen if the same clarity worked in the other direction: a transparency lens that allowed users to see how their anonymous comments and emotions shape the tools that were supposed to help them. This concept translates complicated backend activities into understandable user insights, making it compatible with user-centric solutions like Spotify Wrapped and Apple's Privacy Report. Detailed consent logs, data destruction confirmations, and transparency reports ought to be standard practice in the sector.

---

[22] Court of Justice of the European Union, *Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems (Schrems II)* (Case C-311/18) [2020] ECLI:EU:C:2020:559 https://curia.europa.eu/juris/document/document.jsf?docid=228677 accessed 08 June 2025.

[23] Federal Trade Commission, 'Agreement Containing Consent Order: In the Matter of Flo Health, Inc.' (13 January 2021) https://www.ftc.gov/system/files/documents/cases/flo_health_order.pdf accessed 08 June 2025.

[24] Shokri et al (n 2).

[25] U.S. Department of Health and Human Services, 'Use of Online Tracking Technologies by HIPAA Covered Entities and Business Associates' (HHS.gov, June 2024) https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/hipaa-online-tracking/index.html accessed 09 June 2025.

[26] Government of India. *Digital Personal Data Protection Act 2023*. https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf

[27] European Data Protection Board, 'Opinion 28/2024 on Certain Data Protection Aspects of the Training of AI Models' (Adopted 15 May 2024) para 43 https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en accessed 10 June 2025.

[28] Railsware, 'Product Development Dashboard: What It Is, Benefits, and How to Use It' (Railsware, 16 April 2021) https://railsware.com/blog/product-development-dashboard/ accessed 10 June 2025.

It is also crucial to remember that consent must not be viewed as one-off event. The best mental health platforms should be able to create digital records of users' permissions, authenticated by cryptography, that show exactly when and for what purposes those permissions were granted. These receipts would allow both users and regulators to confirm the full consent lifecycle by utilising blockchain or other immutable ledger technology, especially in cases when queries are raised months later over the use of inferred data or judgements made by opaque models. This makes consent both traceable and contestable, which is a crucial safety measure for AI used in such fragile, risky settings.

4. *Real-time Explainability Tools (with Predictive Alerts integration):*

AI apps need to inform users of the data that is being used and saved in real time. In 2023, Vikranth3140's open-source Mental Health Support Chatbot added session summaries that gave users an overview of the emotional insights, coping mechanisms, and pertinent chatbot triggers that were triggered throughout each conversation. [29] This type of user-centric data framing improved explainability and informed consent in mental health AI by enabling users to see in retrospect how the chatbot perceived their disclosures.

Such solutions, which are based on the research of Explainable AI (XAI)[30], enable people to comprehend and question AI judgements, guaranteeing informed consent and equalising the developer-user dynamic. Explainability, however, shouldn't come too late. In addition, next-gen tools should have inference risk predictive alarms, which are internal scoring algorithms that highlight encounters when user input is exceedingly personal, emotional, or intense. These notifications have the potential to turn explainability into preventative safety by triggering privacy warnings or opt-out prompts prior to the making of sensitive inferences.

5. *Ethical AI-by-design frameworks:*

When developing AI, developers must consider misuse. This involves independent audits, regular red-teaming, and ethics board supervision. The Generative Adversarial Red Teaming (GARD) program of the Defence Advanced Research Projects Agency (DARPA) has established a standard for cooperative AI security by making available the tools and results of its continuous adversarial testing projects created in collaboration with top academic and business partners.[31] This idea is reinforced by frameworks like NIST's AI Risk

[29] Vikranth3140, *Mental Health Support Chatbot* (GitHub, 2023) https://github.com/Vikranth3140/Mental-Health-Support-Chatbot accessed 11 June 2025.

[30] M Muthukrishnan, S Jangoan, K K Sharma and G Krishnamoorthy, 'Demystifying Explainable AI: Understanding, Transparency, and Trust' (2024) 6(2) International Journal for Multidisciplinary Research 1 https://www.ijfmr.com/papers/2024/2/14597.pdf accessed 11June 2025.

[31] Defense Advanced Research Projects Agency, 'DARPA Open Sources Resources to Aid Evaluation of Adversarial AI Defenses' (3 February 2021) https://www.darpa.mil/news/2021/adversarial-ai-defenses accessed 11 June 2025.

Management Framework[32] and the OECD's AI Principles[33], which advocate for ethical embedding from design to deployment.

In addition to institutional supervision, ethical competence ought to be ingrained at individual level. According to the same licensing requirements employed in clinical professions, development teams working on mental health tools should be obliged to complete certified training in AI ethics, mental health risk, and data protection. Such qualifications could possibly be verified via a public register or badge system, guaranteeing that ethical competence is a minimum requirement for individuals designing sensitive technologies.

It's also necessary to set up explicit Inference Incident Disclosure Protocols that require timely and open reporting of any unwanted or damaging inferences. These protocols build trust in AI applications for mental health by promoting accountability, alerting users and stakeholders to possible hazards, and directing prompt redressal.

6. *Mandating differential privacy, federated learning and zero-knowledge proofs:*

Federated learning [34] and differential privacy [35] should no longer be considered frills but rather mandatory necessities. FedMentalCare, a mental health platform, used federated learning in 2023 together with novel techniques like Low-Rank Adaptation (LoRA) to secure sensitive user data on the device.[36] This approach complied with strict laws like HIPAA and GDPR by enabling the fine-tuning of large language models for mental health applications without compromising privacy. Particularly important in high-risk fields like mental health, these protections significantly lessen the likelihood of inference attacks and central data breaches.

Additionally, platforms should implement zero-knowledge proofs (ZKPs), a cryptographic technique that enables businesses to demonstrate adherence to data protection regulations without disclosing any relevant user information. With the use of ZKPs, we can mathematically audit privacy claims and ensure that models don't memorise sensitive inputs or violate training constraints. ZKPs add provability to privacy-by-design frameworks, which sets them apart from conventional privacy audits that rely on vendor disclosure or trust.

7. *Incentivise audits and red-teaming:*

By providing grants, tax breaks, or expedited permits to businesses that disclose their training data regulations and submit to third-party red-teaming, governments can encourage transparency. Proposals for a 25% federal tax

[32] National Institute of Standards and Technology (NIST), *AI Risk Management Framework* (2023) https://www.nist.gov/itl/ai-risk-management-framework accessed 09 June 2025.

[33] OECD, *Principles on Artificial Intelligence* (2019) https://www.oecd.org/en/topics/sub-issues/ai-principles.html accessed 12 June 2025.

[34] P Kairouz and others, 'Advances and Open Problems in Federated Learning' (2019) https://doi.org/10.48550/arXiv.1912.04977 accessed 11 June 2025.

[35] A Roth, C Dwork, *The Algorithmic Foundations of Differential Privacy* (Foundations and Trends in Theoretical Computer Science, 2014) https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf accessed 12 June 2025.

[36] Y Jiang, W Jin and L Zhao, 'FedMentalCare: Privacy-Preserving Fine-Tuning of Large Language Models for Mental Health Analysis via Federated Low-Rank Adaptation' (arXiv, 9 March 2025) https://arxiv.org/pdf/2503.05786v2 accessed 12 June 2025.

credit for businesses funding AI security research have come from advocacy organisations like Americans for Responsible Innovation. The organisation has suggested for participating businesses to be required to make public all of their studies, including those addressing the transparency of training data and the results of red-teaming.[37]

By redefining transparency as a competitive advantage, these incentives motivate both large corporations and startups to go above and beyond the call of duty.

8. Conclusions: we are all test subjects now.

AI mental health chatbots are more than just convenient; they represent a significant shift in how we seek help, share our issues, and handle vulnerability. However, with that intimacy comes an urgent demand: control, clarity, and the inalienable right to be forgotten.

Membership inference is a real-world vulnerability, not a fringe problem or theoretical flaw. In a digital economy based on behavioural breadcrumbs, the capacity to extract your mental health disclosures from opaque models is not only disturbing; it is also exploitative.

This is not a rejection of AI in mental health; in fact, just the opposite. These tools are here to stay and are changing access in significant ways. But we owe it to all users to create systems that prioritise dignity. That entails ethical design that assumes failure, legal frameworks that regard inferred data with the same reverence as explicit data, and architectures that can forget—on purpose.

Because if talking into a chatbot seems like shouting in a mob, it is not therapy. It's an expose.

The future of digital care requires that we do better—not by rejecting innovation, but by expanding its responsibilities. Let us approach transparency not as a regulatory checkbox, but as a design issue worth addressing. Let us reward companies who go above and beyond minimal compliance. And let's envision technologies that respect rather than simply reply.

If the mind is the most private area we have, then the technology that affects it must meet greater standards. This is just the beginning, and the conversation is far from over.

[37] Americans for Responsible Innovation, *AI Security Tax Incentives Memo* (2025) https://ari.us/wp-content/uploads/2025/03/AI-Security-Tax-Incentives-Memo-ARI.pdf accessed 13 June 2025.