# Readiness Evaluation for AI Deployment and Implementation for Mental Health: A Review and Framework

**Authors: Elizabeth C. Stade, Johannes Eichstaedt, Jane P. Kim, Shannon Wiltsey Stirman**

**Author Note**

Address correspondence to Elizabeth C. Stade (betsystade@stanford.edu) or Shannon Wiltsey Stirman (sws1@stanford.edu). A preprint of this manuscript is available at:

**Structured Abstract (300 words)**

**Introduction**

While generative artificial intelligence (AI) may lead to technological advances in the mental health field, it poses safety and ethical risks for mental health service consumers, clinicians, and healthcare systems. To ensure the responsible deployment of AI mental health systems and to support decision-making regarding use of AI, a principled method for evaluating and reporting on generative AI applications is needed.

**Method**

We conducted a narrative review and provide a summary and analysis of the most relevant existing evaluation frameworks and criteria from mental health and healthcare fields to identify guidance for evaluation. We considered proposed criteria, ethical considerations, and unique characteristics of Generative AI technology to determine key criteria and considerations for deployment.

**Observations and Discussion**

Our findings suggest that current frameworks are insufficiently tailored to unique considerations for AI and mental health. We introduce a framework for a readiness for deployment evaluation of AI mental health applications, established based on foundational principles of transparency, consumer autonomy, maximizing benefits and minimizing harm. It comprises considerations of Safety, Privacy/confidentiality, Equity, Effectiveness, Engagement, and Implementation. The framework can be used to evaluate whether AI mental health applications are ready for clinical deployment, and could form the basis for continuous evaluation of these applications.

## Introduction

Generative artificial intelligence (AI), with its ability to understand context, summarize, ask questions and to a lesser extent, reason, holds tremendous potential to expand, augment, or improve mental health care, research, and training/supervision.[1–3] Yet it also holds potential for safety and ethical concerns, including LLM systems providing inappropriate or harmful interventions, producing discriminatory diagnosis or intervention, failing to understand relevant context, and lacking safeguards for patient safety.[1,4,5] This balance of potential and risk raises the question of how to determine whether AI mental health applications are ready for deployment into clinical care. This question is especially urgent as researchers and industry race to develop AI mental health applications.[6,7] Rigorous evaluation and transparency in reporting will support consumer autonomy and organizational decisions about the deployment and use of AI-based technologies. Panel 1 summarizes further justifications for a framework of this nature.

## Method and Selection Criteria

We conducted a narrative review of frameworks and criteria for evaluation of Generative AI-based tools for mental health related assessment and interventions. We searched Medline, PsychINFO, and Google Scholar databases to identify full-text, peer-reviewed, data-based studies and reviews, editorials and opinion pieces. In addition, conducted and Internet Search using Google to identify relevant products by professional or government organizations. Articles were included that we judged to represent proposed criteria for responsible or ethical deployment of generative AI-based mental health or health-related assessment or intervention tools. We searched terms covering Generative AI including conversational agents, Large Language Models, LLM, GPT, conversational AI, mental health-related terms such as psychiatr*, psycholog* or mental* well-being, wellness, or psychiat*, or psychological disorder, and terms relating to specific disorders. We checked reference lists of included papers and of reviews on Generative AI in mental health and healthccare and included articles known by the authors to be relevant to evaluation or deployment of generative AI deployment or implementation, or implementation of digital mental health tools. We identified frameworks that were broadly relevant and/or could provide guidance regarding criteria to consider for responsible and ethical deployment of generative AI in clinical care and summarized the criteria proposed by the most representative and relevant frameworks.

## Discussion and Observations

**Existing Frameworks Relevant to AI Mental Health** Several existing frameworks, criteria, and codes are relevant to the question of how AI mental health applications should be considered for their suitability for evaluation of this nature. Table 1 depicts the principles and considerations set forth in representative frameworks from each area described above, while Figure 1 highlights common themes, areas of consistency, and considerations relevant to the AI mental health space across these representative frameworks. A review of these considerations and criteria led to the selection of essential factors that should be evaluated and reported to support decision-making about the deployment and use of AI-based technology for mental health. These include frameworks from medical and psychological ethics (e.g., principles set forth in the Belmont Report; American Psychological Association's Ethical Principles of Psychologists),[8] and AI governance (e.g., the White House Blueprint for an AI Bill of Rights).[9] Additional frameworks relevant to responsible deployment include those from implementation science (e.g., APEASE Criteria for Designing and Evaluating Interventions),[10] digital mental health (e.g., American Psychiatric Association Mental Health App Evaluation Framework),[11] health equity (e.g., ConNECT framework for health equity in behavioral medicine),[12] and bioethics (Intervention Ensemble for Clinical Machine Learning Systems for evaluating AI healthcare tools).[13] However, as documented in Table 1, none of these frameworks are sufficient for evaluating AI mental health applications. Psychological ethics, implementation science, digital mental health, and health equity frameworks tend to be focused on their

particular domain and tend to insufficiently address the unique considerations of AI and LLMs,[14] while AI governance frameworks focus broadly on applications of AI and thus insufficiently address the particular needs related to healthcare.

There have been recent efforts to systematically address AI in medicine, including bioethical proposals for evaluating AI tools in healthcare,[13] the development of the Consolidated Standards of Reporting Trials–Artificial Intelligence (CONSORT-AI),[15] and the National Academy of Medicine's proposed Artificial Intelligence Code of Conduct project.[16] However, special considerations are needed for the mental health domain. Applications of AI in mental health are poised to take patient language as input and produce patient-facing language as output. Such direct interaction with patients contrasts with the use of AI in medicine as it relates to optimizing imaging or lab-based assessment or improving pharmacological or device-based treatment. Other unique considerations include the sensitive nature of psychotherapy content, the potential for risks of harm to self or others among individuals in acute distress, concerns about stigma and/or potential social and occupational implications related to seeking mental health treatment, and the importance of culturally responsive conceptualization and treatment.

Given the unique demands of mental health assessment and treatment, we argue that an AI mental health-specific framework is needed. The proposed framework is depicted in Figure 2, and is intended to uphold key principles outlined in the Belmont Report to ensure maximum benefit and minimize potential harms, promote transparency and support individual autonomy,[17] while also allowing organizations and systems to make informed decisions about the appropriateness and potential for successful implementation and use of specific AI-based technologies.

**Elements of the AI Mental Health Evaluation and Reporting Framework: The READI criteria**

To address this need, here we introduce the Safety, Privacy/confidentiality, Equity, Effectiveness, Engagement, and Implementation for the evaluation of AI mental health applications. To position the READI framework in the broader literature, in Table 2 we highlight crosswalks to the aforementioned existing criteria and codes that are relevant to the intersection of AI and mental health. Foundational to this framework are principles of transparency and consumer autonomy, which guided the selection of these criteria for evaluation and reporting. Individuals and Organizations should be made aware of how exactly AI is used in a given technology, how extensively AI is used across the application (i.e., circumscribed versus broad scope of AI), and the limits of its abilities (e.g., the product's inability to diagnose, initiate consults, contact emergency services, or admit individuals to a higher level of care if required). Additionally, individuals and organizations should be informed of the most recent evaluations based on the criteria below to support decisions about deployment. Below, we briefly define each component of the READI framework and highlight how the component can be evaluated in an objective manner. These criteria include characteristics or features that may change (e.g., privacy/confidentiality) or be context-dependent (e.g., equity, engagement, implementation considerations). We recommend that new Generative AI technologies be evaluated before large-scale deployment, and that they be subject to ongoing evaluation and consideration, given the rapidly changing technology and the dynamic contexts into which these tools may be deployed. In Table 3, we define each component's aspirational goal/value, outline criteria for evaluating each component and a set of questions to be used for evaluation, and define requirements for information that should be disclosed or reported with regard to each component.

***Safety***

The first component within the proposed framework for evaluating AI mental health applications is safety. Safety encompasses multiple dimensions, beginning with the assurance that AI monitors against – and does not actively promote – dangerous or unhealthy human behaviors, such as self-harm, suicide, abuse, harmful substance use, or other dangerous behaviors.[18] Notably, contextual nuances, such as knowledge of the individual or population being treated, play an important role in determining whether or not a behavior is unsafe. For example, a chatbot designed to help individuals with eating disorders was

found to be providing diet and weight loss advice;[1] this AI behavior was unsafe in this context but might not be deemed harmful in alternative contexts.

Beyond preventing explicit harm, safety encompasses ensuring that the AI application itself is "healthy." Given that LLMs are trained on internet data, they may have inflammatory and extreme traits which could interfere with the development of a healthy therapeutic experience for the patient. Therefore safety will entail the application not exhibiting such traits, states, or behaviors.[19,20] While it will be important for all AI applications to be free from "psychopathology," this is especially important for psychiatric applications given the needs of the population the applications will serve. Relatedly, AI applications must not exhibit behaviors that could exacerbate or maintain the patient's presenting problem, such as promoting problematic thinking patterns like all-or-nothing thinking.[21] In light of safety considerations of this nature, substantial involvement of individuals with content expertise during development and testing is essential.

Lastly, safety considerations extend to the monitoring and reporting of adverse events, including instances in which the AI behaves unexpectedly or fails to detect high-risk situations. This necessitates a defined protocol for reporting such events, akin to the FDA Adverse Event Reporting System.[22] To reach the highest safety standards, AI mental health applications should scope out predefined plans for how to withdraw or cease the AI intervention and how to switch or escalate to human intervention.

***Privacy/confidentiality***

The next components of the framework are patient rights of privacy and confidentiality. In the US, a major challenge to the privacy pillar as it relates to mental health applications is a lack of legal protection. Due to a loophole, the Health Insurance Portability and Accountability Act's (HIPAA) Privacy Rule, which sets standards for the use and disclosure of individuals' health information, does not protect most health information collected by direct-to-consumer AI mental health applications.[23] Greater consumer privacy protections are needed. Until then, it is imperative that patient information be safeguarded at a level consistent with the HIPAA Privacy Rule. AI mental health applications should not disclose individuals' health information without their authorization and should offer the ability for individuals to examine their health information. It is also essential that consumers understand the safeguards in place to prevent data breaches, steps that will be taken in the event of a breach, and that they are notified about the nature and scale of any violations.

To ensure transparency and promote individual autonomy in deciding whether and how to engage with AI-based applications, terms of service and design decisions can be designed to promote true, informed patient choice with regard to privacy and data collection/storage. This means making users aware if and how the information will be used to refine and improve the application, and whether the data will be used for other purposes (e.g., research, sale to third parties). Additionally, an emerging new class of privacy risk stems from "data leakage" due to consumer data serving as training data for AI models that they could be disclosing ("leaking") in another context. If consumer data is used for any of these applications, providing simple ways to opt in or out of specific uses of their data while retaining access to the application (e.g., not making use of the application contingent on allowing third-party access) provides the greatest level of individual autonomy and access (but is often not without difficulty with regard to user experience and demands on users).

***Equity***

The next component of our proposed framework is equity. Care should be taken to evaluate and disclose potential reification of biases or any de-biasing methods used during both the development and evaluation stages. If the application was fine-tuned or tailored using human samples or training materials (e.g., a manual for culturally competent cognitive behavioral therapy), reporting the demographics of individuals whose data have been used in research or evaluation (including age, race, and gender) and/or the population for which the intervention was designed can facilitate informed decision-making about their

applicability and limitations for specific use-cases or end-users. Rather than solely being evaluated by developers or subject matter experts, applications should be tested by representative end-users, with information provided about the characteristics of the end-users who participated in efforts to test and refine the applications. Furthermore, engagement, effectiveness, and satisfaction data (see below) should be reported across demographic groups, such that it would become apparent if these metrics differed by demographic group (e.g., the intervention is less effective for Black individuals than White individuals).[24] Beyond mitigating against biases, developers should seek to integrate culturally competent and responsive practices into their applications,[25] especially given the opportunities for such afforded by generative AI (e.g., changing the language used to describe the intervention, including culturally-specific content). This is an area that is receiving significant attention in the literature, and it is likely that additional sources of bias will be identified and that methodologies will continue to evolve. Use of best practices, ongoing evaluation, and transparency about evaluation and actions being taken to reduce bias and promote culturally responsive interventions will be of great importance.

***Engagement***

The next component of the READI framework is engagement. Engagement is an important component to evaluate because as opposed to traditional mental health treatments, AI mental health applications may be used in a self-directed manner and will likely be constantly available to the patient. Ideal engagement will consist of a degree of application usage sufficient to produce therapeutic benefit for the patient, but not so extreme as to cause problems such as dependence or overuse for the patient.

On one hand, if no lower bound standard for application engagement is set, or even if standards seek to approximate average levels of engagement with existing digital mental health interventions (which have 1-week dropout rates of 90%),[26] patients may be unlikely to benefit from the intervention. On the other hand, if no upper bound standard for engagement is set, patients could develop patterns of using the application that are counter-therapeutic. Examples include using the application as a safety behavior, forgoing meaningful life activities in favor of application usage, or developing a very strong (pseudo) social connection to the application, perhaps to the detriment of other social connections. Key engagement metrics thus include the number of days the application was used since the onboarding period, the number of consecutive days of application use, and time spent using the application per day/week/month.

Individual needs may play a role in determining the appropriate level of engagement for each patient: Some individuals may engage rarely with the application while consistently integrating the skills and tools they are learning into their daily lives, while other individuals may require daily interactions with the application for optimal outcomes. AI mental health applications will maximally benefit patients' health and wellness by being in service of their non-AI related experiences and relationships (e.g., helping a patient engage in a valued activity despite low motivation; helping a patient address issues in their romantic relationship). Therefore, the current framework advocates for properly dosed engagement which prioritizes the patient's interactions with their external environment.

***Effectiveness***

The next component within the READI framework is effectiveness. As is the standard for psychological treatments in general, AI mental health treatment applications should have evidence of effectiveness in clinically representative settings.[27] The set of key effectiveness outcomes for an AI mental health application includes decreases in symptoms (of the disorder or clinical issue being targeted) and functional impairment, as well as increases in well-being and quality of life.[27] For this criterion to be properly assessed, outcomes will need to be reported clearly. We recommend the disclosure and reporting of the details of the intervention, sample size, demographic details of the sample (age, gender, geographical location, race), whether or not there was a comparison group/condition, details of the comparison group/condition, outcome construct(s) assessed, measure(s) used to assess

construct(s), and details of the measures' construct validity and reliability. This information, in plain language, can support informed decision-making about the applicability and appropriateness of the use (for a single user) or deployment (for a healthcare system, clinic, or other setting) of the application.

### *Implementation Considerations*

The final component of the framework addresses the need for AI mental health applications to be developed with implementation in mind. AI mental health applications will need to be integrated into routine care settings to have the greatest impact. This will require addressing the needs and practical challenges associated with their incorporation into clinical practice, existing technologies (e.g., electronic medical records), and workflows. It should be recognized that any benefit derived from AI tools is contingent upon a larger set of parameters dictating how and under what conditions the tool is implemented.[13] Therefore, we recommend collecting data on the feasibility, acceptability, compatibility and perceptions of the AI mental health application at multiple levels within healthcare systems (e.g., patient, clinician, healthcare system administrator, etc.), providing information about how it functions in these settings, compatibility with existing policies (e.g., HIPAA) and systems such as electronic medical records, costs, and other information about its potential for broad implementation.

## Conclusions

AI mental health applications can dramatically enhance the quality and scalability of evidence-based psychotherapies. However, they also have the potential to cause harm.[1] The area of mental health is an uncommonly high-stakes domain, given the vulnerability of the patient population and the high-risk topics addressed in mental health settings (e.g., suicide, violence, abuse, self-harm). Given that many AI mental health applications will be developed by the private sector, in the absence of a standardized set of evaluation criteria, companies may feel pressure to optimize towards specific business objectives without sufficiently attending to concerns regarding clinical effectiveness or patient rights. At the same time, applications developed by researchers or healthcare systems may overlook important considerations like usability, engagement, effectiveness, and applicability for populations beyond the system for which they were developed. For these reasons, a framework for the evaluation and transparent reporting of AI mental health applications is needed that can span the academic and private domains.

The framework proposed here lays out a set of criteria that could be used to evaluate any generative AI mental health application. Our goal in developing this proposed framework was not to exhaustively list all different frameworks relevant to AI, but rather to ensure that we were addressing the critical factors. Thus, the proposed framework seeks to balance clinical effectiveness, concerns for human rights (privacy/confidentiality, transparency and autonomy, equity), the desire to mitigate potentially harmful qualities of AI (safety, engagement), and the desire for these applications to be maximally useful in clinical settings (implementation). While we do not explicitly specify the party responsible for performing the evaluation of an AI mental health application, this framework provides needed guidance on factors that can be transparently reported and updated over time as technology and healthcare solutions evolve. It identifies considerations that can be reviewed when AI-based technologies are being considered for individual use or large-scale deployment.

Ideally, application developers, perhaps in partnerships with researchers and/or end-users (e.g., organizations), can collect and provide the information in plain language. However, simply responding to the criteria set forth in the framework and disclosing the requested information does not determine whether the application is appropriate for clinical deployment. Ultimately, these judgments are likely best made by end-users in consultation with their providers, and for broad implementation, by groups of individuals with mental health domain expertise, lived experience, and visibility into existing systems, including consumers, clinicians and healthcare administrators. Therefore, following similar frameworks in the AI medicine space,[28] we propose that the evaluation of AI mental health applications should be a joint responsibility between application developers, consumers, policymakers, and clinicians/administrators. As

solutions in the mental health space are often marketed to payers (such as insurance or Medicare) or larger healthcare systems, it is conceivable that evaluation frameworks such as READI become part of their diligence and evaluation process. Evaluation and reporting as set forth in the READI framework can increase confidence and trust in AI innovations and empower consumers and practitioners to make informed choices about whether and how AI is used to support mental health.

**Contributors**
ECS and SWS were responsible for project conception; ECS drafted the manuscript; all authors were responsible for reviewing the manuscript.

**Declaration of Interests**
ECS reports receiving consulting fees from Jimini Health and Sonar Inc.; JCE reports receiving consulting fees and equity in Jimini Health and equity in Sonar Inc.; SWS reports equity in Hello Therapeutics.

**Acknowledgements**
We gratefully acknowledge Zoe Tait for her help in preparing figures for this manuscript. JCE receives funding from the Institute for Human-Centered Artificial Intelligence and Center for Artificial Intelligence in Medicine and Imaging at Stanford University and from the National Institute for Mental Health (grant number RF1-MH125702); JPK receives funding from the National Center for Advancing Translational Sciences (grant number R01-TR003505); SWS receives funding from the Center for Artificial Intelligence in Medicine and Imaging at Stanford and the National Institute of Mental Health (grant number RF1-MH128785). The funding sources had no involvement in study design, data collection, analysis, or interpretation, manuscript writing, or the decision to submit this paper for publication.

**Panel 1: Why is an AI mental health framework needed?**

- There exist different incentive structures (and blindspots) across academia, healthcare systems, and industry, all of which play a role in developing AI mental health applications.
- In academia, where incentives often revolve around the impact and uniqueness of research, implementation and distribution concerns and engaging design of the application may be overlooked.
- Healthcare systems, driven by concerns about reputation, privacy, and cost-effectiveness, may underpromote innovation.
- In industry, the pursuit of growth, revenue and high customer lifetime value may incentivize a focus on user engagement and retention at the expense of focusing on the intervention's short-term effectiveness or ethical considerations like user privacy.
- The creation of a patient-centered, ethically-driven, and market-independent framework for evaluation, which sits outside of academia, healthcare systems, and industry, would help to address the blindspots originating in the different incentive structures.
- Such a framework could form the basis for both the initial and the ongoing evaluation of AI mental health applications.
- Framework components may be particularly well-suited to constitute key performance indicators or objectives and key results in the case of industry.
- Furthermore, rather than tasking the developer with the dissemination of the requisite information for application evaluation (which may not be done transparently and completely), and rather than tasking the end-user (i.e., patient) with the evaluation of the application, an alternative approach involves the establishment of a unified framework (i.e., set of standards) to ensure the systematic provision and evaluation of such critical information.

**References**


1. De Choudhury M, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health. Published online November 7, 2023. Accessed December 23, 2023. http://arxiv.org/abs/2311.14693
2. Kjell ONE, Kjell K, Schwartz HA. Beyond rating scales: With targeted evaluation, language models are poised for psychological assessment. *Psychiatry Res*. Published online December 2023:115667. doi:10.1016/j.psychres.2023.115667
3. Stade EC, Stirman SW, Ungar LH, et al. *Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation*. PsyArXiv; 2023. doi:10.31234/osf.io/cuzvr
4. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digit Health*. 2023;9:1-11. doi:10.1177/20552076231183542
5. Khawaja Z, Bélisle-Pipon JC. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Front Digit Health*. 2023;5:1278186. doi:10.3389/fdgth.2023.1278186
6. Lai T, Shi Y, Du Z, et al. Psy-LLM: Scaling up global mental health psychological services with AI-based large language models. Published online September 1, 2023. Accessed January 17, 2024. http://arxiv.org/abs/2307.11991
7. Mehta A, Niles AN, Vargas JH, Marafon T, Couto DD, Gross JJ. Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (Youper): Longitudinal observational study. *J Med Internet Res*. 2021;23(6):e26771. doi:10.2196/26771
8. American Psychological Association. Ethical principles of psychologists and code of conduct. Published online 2016. https://www.apa.org/ethics/code/ethics-code-2017.pdf
9. White House Office of Science and Technology Policy. Blueprint for an AI bill of rights. Published online 2022. https://www.whitehouse.gov/ostp/ai-bill-of-rights/
10. Michie S, Atkins L, West R. *The Behaviour Change Wheel: A Guide to Designing Interventions*. Silverback Publishing; 2014.
11. Lagan S, Emerson MR, King D, et al. Mental health app evaluation: Updating the American Psychiatric Association's framework through a stakeholder-engaged workshop. *Psychiatr Serv*. 2021;72(9):1095-1098. doi:10.1176/appi.ps.202000663
12. Alcaraz KI, Sly J, Ashing K, et al. The ConNECT Framework: A model for advancing behavioral medicine science and practice to foster health equity. *J Behav Med*. 2017;40(1):23-38. doi:10.1007/s10865-016-9780-4
13. McCradden MD, Joshi S, Anderson JA, London AJ. A normative framework for artificial intelligence as a sociotechnical system in healthcare. *Patterns*. 2023;4(11):100864. doi:10.1016/j.patter.2023.100864
14. Diaz-Asper C, Hauglid MK, Chandler C, Cohen AS, Foltz PW, Elvevåg B. A framework for language technologies in behavioral research and clinical applications: Ethical challenges, implications, and solutions. *Am Psychol*. 2024;79(1):79-91. doi:10.1037/amp0001195
15. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health*. 2020;2(10):e537-e548. doi:10.1016/S2589-7500(20)30218-1
16. National Academy of Medicine. Health care artificial intelligence code of conduct: Toward a code of conduct for artificial intelligence in health, health care, and biomedical science. Published online 2023. https://nam.edu/wp-content/uploads/2023/12/AICC-Flier_FINAL-12.5-upload.pdf
17. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*.; 1979. https://www.hhs.gov/ohrp/sites/default/files/the-

belmont-report-508c_FINAL.pdf
18. Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from Language Models. Published online December 8, 2021. doi:10.48550/arXiv.2112.04359
19. Behzadan V, Munir A, Yampolskiy RV. A psychopathological approach to safety engineering in AI and AGI. Published online May 22, 2018. Accessed July 12, 2023. http://arxiv.org/abs/1805.08915
20. Lin B, Bouneffouf D, Cecchi G, Varshney KR. Towards healthy AI: Large language models need therapists too. Published online April 1, 2023. Accessed June 29, 2023. http://arxiv.org/abs/2304.00416
21. Beck AT. Cognitive models of depression. *J Cogn Psychother Int Q*. 1987;1:5-37.
22. Sonawane KB, Cheng N, Hansen RA. Serious adverse drug events reported to the FDA: Analysis of the FDA adverse event reporting system 2006-2014 database. *J Manag Care Spec Pharm*. 2018;24(7):682-690. doi:10.18553/jmcp.2018.24.7.682
23. Gerke S, Rezaeikhonakdar D. Privacy aspects of direct-to-consumer artificial intelligence/machine learning health apps. *Intell-Based Med*. 2022;6:100061. doi:10.1016/j.ibmed.2022.100061
24. Lalor J, Yang Y, Smith K, Forsgren N, Abbasi A. Benchmarking intersectional biases in nlp. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics; 2022:3598-3609. doi:10.18653/v1/2022.naacl-main.263
25. Sue S, Zane N, Nagayama Hall GC, Berger LK. The case for cultural competency in psychotherapeutic interventions. *Annu Rev Psychol*. 2009;60(1):525-548. doi:10.1146/annurev.psych.60.110707.163651
26. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: Systematic search and panel-based usage analysis. *J Med Internet Res*. 2019;21(9). doi:10.2196/14567
27. Tolin DF, McKay D, Forman EM, Klonsky ED, Thombs BD. Empirically supported treatment: Recommendations for a new model. *Clin Psychol Sci Pract*. 2015;22(4):317-338. doi:10.1111/cpsp.12122
28. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health*. 2022;4(5):e384-e397. doi:10.1016/S2589-7500(22)00003-6

Table 1. Representative Frameworks in Different Areas: Artificial Intelligence, Psychological Ethics, Implementation, Digital Mental Health, Health Equity, and Bioethics

| | **White House Blueprint for an AI Bill of Rights** | **Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI)** | **American Psychological Association Ethical Principles of Psychologists and Code of Conduct** | **APEASE Criteria for Designing and Evaluating Interventions** | **American Psychiatric Association Mental Health App Evaluation Framework** | **ConNECT Framework** | **Intervention Ensemble for Clinical Machine Learning Systems** |
|---|---|---|---|---|---|---|---|
| Category | Artificial intelligence (general) | Artificial intelligence (medicine) | Psychological ethics | Implementation | Digital mental health | Health equity | Bioethics |
| Target audience | Individuals or entities building or deploying automated systems involving AI | Researchers reporting clinical trials; readers evaluating clinical trials | Psychologists | Researchers and practitioners developing behavior change interventions | Mental health care providers | Healthcare practitioners and researchers | Stakeholders interested in integrating AI tools into clinical practice |
| Purpose | Set of principles and associated practices to guide the design and deployment of | Checklist of items to address AI-specific content not covered by the | Set of principles and ethical standards for the professional practice of psychology. | Set of dimensions to aid in selecting and developing interventions that can be implemented successfully. | Hierarchical rating system to guide the process of evaluating, and decision-making about the clinical use of, mental | Set of principles to achieve health equity in behavioral medicine | Set of elements with which to evaluate AI healthcare tools (alongside other aspects that comprise the |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | artificial intelligence applications with regard to the rights of the American public. | Consolidated Standards of Reporting Trials | | | health applications | | medical intervention) |
| Key citation | White House Office of Science and Technology Policy (2022) | Liu et al. (2020) | American Psychological Association (2002) | Michie et al. (2014) | Lagan et al. (2021) | Alcaraz et al., (2017) | McCradden et al., (2023) |
| Principles/ dimensions/ standards | Safe and Effective Systems<br>Algorithmic Discrimination Protections<br>Data Privacy<br>Notice and Explanation<br>Human Alternatives, Consideration, and Fallback | Elaborations/extensions to trial reporting standards in the following domains:<br>Title and abstract<br>Background /objective<br>Participant<br>Interventions<br>Harms<br>Funding | Beneficence and Nonmaleficence<br>Fidelity and Responsibility<br>Integrity<br>Justice<br>Respect for People's Rights and Dignity | Affordability;<br>Practicability<br>Effectiveness and cost-effectiveness<br>Acceptability<br>Side-effects/safety<br>Equity | Accessibility and background<br>Privacy and security<br>Clinical foundation<br>Engagement style<br>Therapeutic goal | Integrating Context<br>Fostering a Norm of Inclusion<br>Ensuring Equitable Diffusion of Innovations<br>Harnessing Communication Technology<br>Prioritizing Specialized Training | Use case<br>Task and outcome<br>Performance threshold setting<br>Performance across subpopulations<br>Use parameters and limitations<br>Monitoring |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Missing | Not tailored to mental health | Not tailored to mental health<br>No overarching ethical values/principles | No consideration of AI | No consideration of AI | No consideration of AI | No consideration of AI<br>Not tailored to mental health | Not tailored to mental health |

Table 2. Crosswalk from Representative Frameworks to the READI Framework for the Evaluation of AI Mental Health Applications

| **Framework Component** | **Whitehouse Proposal for an AI Bill of Rights** | **Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI)** | **American Psychological Association Ethical Principles of Psychologists and Code of Conduct** | **APEASE Criteria for Designing and Evaluating Interventions** | **American Psychiatric Association Mental Health App Evaluation Framework** | **ConNECT Framework** | **Intervention Ensemble for Clinical Machine Learning Systems** |
|---|---|---|---|---|---|---|---|
| Safety | Safe and Effective Systems<br>Human Alternatives, Consideration, and Fallback | | Beneficence and Nonmaleficence | Side-effects/safety | Privacy and security | | Monitoring |
| Privacy/Confidentiality | Data Privacy<br>Human Alternatives, Consideration, and Fallback | | Respect for People's Rights and Dignity | | Privacy and security | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Effectiveness | Safe and Effective Systems | | Beneficence and Nonmaleficence | Effectiveness and cost-effectiveness | Clinical Foundation | | Performance threshold setting |
| Equity | Algorithmic Discrimination Protections | | Justice | Equity | Access and Background | Integrating Context<br>Fostering a Norm of Inclusion<br>Ensuring Equitable Diffusion of Innovations<br>Harnessing Communication Technology<br>Prioritizing Specialized Training | Performance across subpopulations |
| Engagement | | | | | | | |
| Implementation | | | | Practicability<br>Affordability<br>Acceptability | Usability<br>Data Integration towards Therapeutic Goal | | The task and outcome<br>Use parameters and limitations |

Figure 1. Areas Contributing to an AI Mental Health Framework

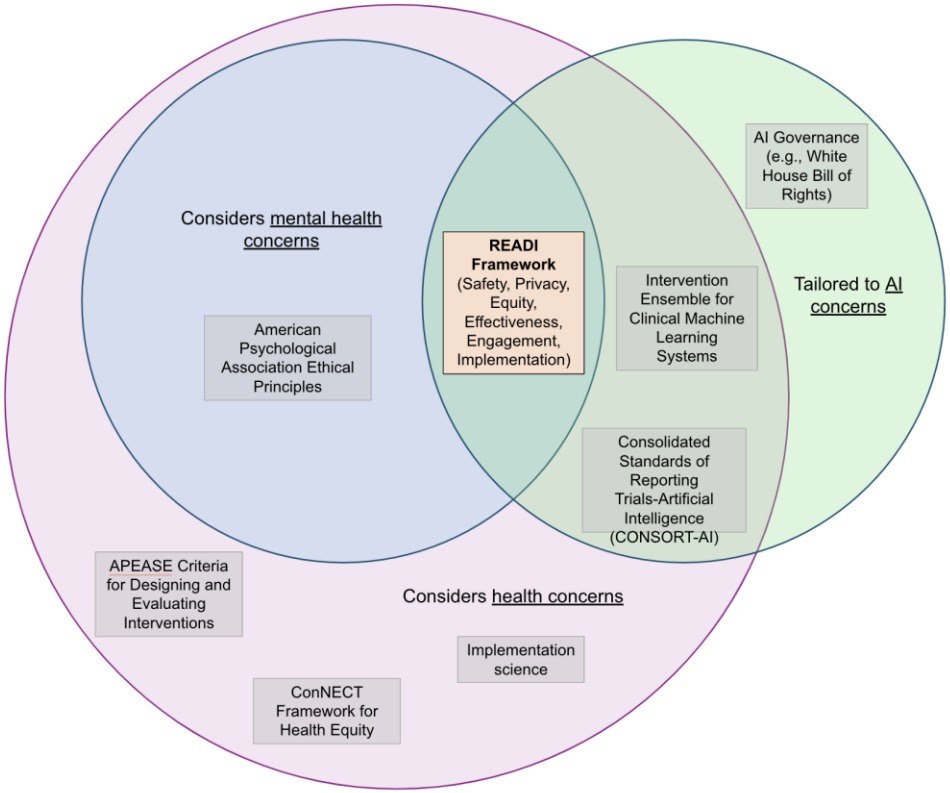

Figure 2. READI Framework Components


Safety
Application prevents dangerous human behaviors and is "healthy" itself
Implementation
Application integrates well into real-world clinical settings
Privacy/ confidentiality
Patient information is kept private and confidential
READI Framework
Foundational Principles: Transparency, Autonomy, Minimize Harm & Maximize Benefit
Engagement
Application is appropriately engaging (not too much or too little)
Equity
Application is entirely unbiased in its communication, effectiveness, and engagement; application is equally acceptable and useable across all demographic groups; application is culturally responsive
Effectiveness
Application integrates clinical science principles and is clinically effective

Figure 3. The READI Framework and Associated Evaluation Components

**Foundational Principles: Transparency, Autonomy, Minimize Harm & Maximize Benefit**

| READI Component | Criteria | Evaluation Questions | Suggested Reporting Information |
|---|---|---|---|
| **Safety**: Application prevents dangerous human behaviors and is "healthy" itself | • AI monitors for and does not promote or reinforce dangerous or risky patient/user behaviors<br>• AI does not exhibit psychopathological traits, states, or behaviors such as narcissism, depression, or gaslighting.<br>• Adverse events are consistently monitored for and reported<br>• There is a plan in place for withdrawing or ceasing the AI intervention and/or or escalating to human intervention | • Are dangerous or risky human (patient/user) behaviors and/or intent to harm monitored for? (Which ones? How?)<br>• Are unhealthy AI behaviors monitored for? Which ones? How are they assessed?<br>• If yes to any of the above, are there quality assurance mechanisms (e.g., audits)? Is there a reporting mechanism for monitoring failures? | • Adverse events<br>• Steps taken to ensure safety (e.g., hard coding of a suicide warning message)<br>• Guardrails in place |
| **Privacy**: Patient information is kept private and confidential | • Patient information is safeguarded at a level consistent with the HIPAA Privacy Rule.<br>• Limited and use of health information, transparent accessible information on how it is used<br>• Broader use and/or sale of health information optional, explicitly opted into.<br>• Application usage is not contingent upon allowing third-party access to health information | • To whom or to what entities is patient information disclosed, and under what conditions?<br>• How and when are patients presented with a choice about data collection and use? What is the default decision?<br>• What pieces of PHI/data collection is application usage contingent upon?<br>• For patients who have opted into data collection: which data, how will it be used, by whom, and for how long? | • Privacy breaches<br>• Steps taken to ensure privacy (e.g., data storage) |
| **Equity**: Application is entirely unbiased in its communication, effectiveness, and engagement; application is equally acceptable and useable across all demographic groups; application is culturally responsive | • Demographic-based biases, including but not limited to racism, sexism, homophobia and ableism are monitored for.<br>• Debiasing methods are in place to mitigate against biases relevant to the clinical population.<br>• The application is tested by representative end-users for bias and cultural sensitivity<br>• Continuous improvement and use of best practices to promote equity | • Are demographic-based biases monitored for? Which ones? How are they assessed?<br>• Are debiasing methods integrated into the application? Which ones? How do they work?<br>• Were human samples used to fine-tune or tailor model(s)? If yes, what is the sample size, and the age, race, ethnicity and gender of the sample?<br>• Was the application tested by representative end-users and assessed for bias and cultural sensitivity? | • Steps taken during development to ensure equity and cultural appropriateness<br>• Demographic information, including age, race, and gender, of human samples used to fine-tune or tailor model(s).<br>• Engagement/outcome data by demographic group |
| **Effectiveness**: Application integrates clinical science principles and is clinically effective | • Key effectiveness outcomes are reported, including a measure of disorder or clinical issue being targeted and a well-being or quality-of-life measure.<br>• Metrics reported for the whole sample tested, as well as by demographic group (e.g., age, race, ethnicity and gender).<br>• Details of the effectiveness study reported, including intervention and control condition (if applicable) details, sample size and sample descriptive statistics. | • How was effectiveness measured? Was the sample representative of the population?<br>• Which outcome construct(s) were assessed? Were established and reliable measure(s) used to assess the construct(s)?<br>• For the effectiveness data reported: How was the intervention administered? for how long? What was the comparison group/condition?<br>• What were the sample size and demographic details of the sample (age, gender, geographical location, race)? | • Effectiveness metrics<br>• Changes in: Symptoms, Quality of Life, Well-being, Functioning, Population and moderators of effectiveness |
| **Engagement**: Application is appropriately engaging (not too much or too little) | • Key engagement metrics are reported including:<br>  • number of days the application was used since onboarding<br>  • number of consecutive days of application use<br>  • time spent using the application per day/week/month<br>  • Satisfaction, "Alliance" | • What is the average number of days the application was used?<br>• On average, how many consecutive days was the application used?<br>• On average, how many minutes was the application used per day/week/month?<br>• How do users rate the application on "engaging" and "user-friendliness" metrics?<br>• What do users report as their reasons for ongoing use and/or discontinuation? | • Amount of time and frequency of use and proportion of people who complete the program, by demographic group (age, race, ethnicity and gender)?<br>• Digital Alliance and Satisfaction |
| **Implementation**: Application integrates well into real-world clinical settings | • Feasibility and acceptability data of the application are reported for each relevant stakeholder<br>• Rates of referrals/reach/adoption of the product.<br>• Compatibility with existing workflows and technology<br>• Compliance with institutional requirements and policies. Adaptation and fidelity, where applicable (e.g., where LLMs are based on specific interventions) if/as the model is refined<br>• Cost/affordability | • How feasible and acceptable did stakeholders involved in decision-making about use and deployment find the application to be?<br>• Can/does the application integrate into existing systems, processes, and workflows?<br>• How widely is it being used within the system? Is it being used correctly/retaining fidelity as the model is refined over time?<br>• What are the costs to the system and individual? Can use of the application be reimbursed? | • Potential for integration into existing technologies and workflows<br>• Findings on Feasibility, Acceptability, Reach , Fidelity, Cost |