Regulatory Frameworks for Generative AI Enabled Digital Mental Health Devices: Safety, Transparency, and Post-Market Oversight

Satyadhar Joshi
Independent Researcher
Alumnus, International MBA, Bar-Ilan University, Israel
Alumnus, Touro College MSIT, NY, USA
ORCID: 0009-0002-6011-5080
satyadhar.joshi@gmail.com

***Abstract*—The rapid growth of generative artificial intelligence in digital mental health interventions offers significant opportunities to improve mental healthcare access while creating new regulatory challenges. This paper responds to recent U.S. Food and Drug Administration initiatives, including the September 2025 Digital Health Advisory Committee meeting, by proposing comprehensive regulatory frameworks for generative AI digital mental health devices. We analyze the current regulatory landscape, identifying gaps in U.S., international, and state-level governance structures. Through quantitative foundations including mathematical models for risk assessment, objective functions for regulatory optimization, and the 4 lens framework for significant change evaluation, we establish evidence-based approaches for device assessment. We present architectural diagrams covering lifecycle regulatory pathways, multi-layered safety architectures, risk-tiered assurance frameworks, and multi-stakeholder governance models. Drawing from clinical evidence showing both potential benefits and significant risks, we advocate for balanced regulatory approaches. Our framework integrates technical safeguards, ethical considerations based on care ethics, transparency requirements, and post-market monitoring systems. We provide implementation roadmaps, quantitative algorithms for regulatory decisions, and cost-benefit analyses to support practical deployment. The paper concludes with specific recommendations for risk-based classification, adaptive oversight systems, international coordination, and enhanced professional involvement to ensure these technologies provide therapeutic benefits while maintaining strong patient safety standards throughout their lifecycle. This is a review and synthesis paper that summarizes and organizes existing proposals, frameworks, and discussions from current literature; the author does not claim original authorship of the regulatory frameworks presented but rather provides a systematic analysis of the current discourse.**

***Index Terms*—Generative AI, Digital Mental Health, FDA Regulation, Post-Market Surveillance, AI Governance, Medical Devices, Ethics, Safety, Transparency.**

I. Introduction

The proliferation of generative artificial intelligence (AI) in digital mental health interventions (DMHIs) represents a transformative shift in how mental health care is delivered and accessed. These technologies, which include AI-powered chatbots, therapeutic conversational agents, and adaptive wellness applications, promise to address critical gaps in mental health care access, especially in underserved populations [1], [2]. However, their rapid evolution and inherent complexity—characterized by adaptive learning, natural language interactions, and personalized outputs—pose unprecedented regulatory challenges. In September 2025, the U.S. Food and Drug Administration (FDA) announced a public meeting of its Digital Health Advisory Committee to specifically address "Generative Artificial Intelligence-Enabled Digital Mental Health Medical Devices" . This notice underscores the urgency of developing coherent regulatory frameworks that balance innovation with patient safety.

Current regulatory approaches, largely designed for static medical devices, struggle to accommodate the dynamic, evolving nature of AI-driven tools [3]. Many generative AI mental health applications operate in a regulatory "gray area," often escaping stringent pre-market review if marketed as wellness rather than medical devices [4], [5]. This gap raises significant public health concerns, including risks of harmful advice, privacy violations, and inadequate crisis response [6]. Recent studies have documented instances where AI chatbots provided inappropriate or unsafe responses to users experiencing acute mental health crises [7], [8].

This paper responds indirectly to the FDA's call for commentary by synthesizing current evidence and regulatory perspectives. We examine:

- The current U.S. and global regulatory landscape for AI in mental health;
- Evidence on safety, efficacy, and user experience of generative AI DMHIs;
- Ethical and governance challenges, including transparency and accountability;
- Proposals for pre-market evaluation and post-market surveillance frameworks.

By integrating insights from recent scholarship, policy documents, and international guidelines, we aim to inform regulatory discussions and promote a robust, adaptive oversight system for generative AI in mental health.

II. Current Regulatory Landscape

A. U.S. FDA Framework and Recent Developments

In the United States, the FDA regulates AI-enabled software as a medical device (SaMD) under existing authorities, primarily the Federal Food, Drug, and Cosmetic Act. The agency has issued guidance on AI/machine learning (ML)-based SaMD, emphasizing a total product lifecycle approach and the use of predetermined change control plans (PCCPs) for iterative updates [9], [10]. However, as noted in the FDA's recent meeting notice, generative AI-enabled mental health devices present "novel risks" that may necessitate evolved regulatory approaches .

The FDA's Digital Health Advisory Committee meeting scheduled for November 2025 focuses explicitly on benefits, risks, and risk mitigations for these devices, including premarket evidence and post-market monitoring considerations . This aligns with broader agency efforts to engage stakeholders in shaping adaptive regulatory pathways for digital health technologies [11], [12]. Concurrently, the Centers for Medicare & Medicaid Services (CMS) are evaluating coverage and payment policies for AI-enabled tools, influencing their adoption and economic viability [13].

Despite these steps, regulatory gaps persist. Many generative AI wellness apps bypass FDA review by avoiding explicit medical claims, yet users often employ them for mental health support, sometimes in crisis situations [4]. This misalignment between intended use and real-world application creates a precarious safety environment. Recent analyses call for clearer classification criteria and heightened oversight for AI tools that, regardless of marketing, engage in therapeutic interactions [14], [15].

B. International and State-Level Regulatory Initiatives

Globally, regulatory bodies are also grappling with AI in mental health. The European Union's AI Act classifies certain AI systems in health as high-risk, subjecting them to rigorous conformity assessments, transparency obligations, and human oversight requirements [16]. The UK's regulatory response to the Regulatory Horizons Council emphasizes a proportionate, context-based framework for AI as a medical device [17]. Meanwhile, the World Health Organization (WHO) has issued ethics and governance guidance for large multi-modal models, stressing accountability, inclusivity, and professional oversight in health applications [18].

At the U.S. state level, legislative activity is accelerating. A 50-state review identified 143 bills introduced between 2022 and 2025 with potential implications for mental health AI, though explicit mental health provisions remain rare [15]. Themes include professional oversight, harm prevention, patient autonomy, and data governance. This fragmented landscape underscores the need for federal leadership to ensure consistent safety standards while allowing state innovation.

Australia's recently released National AI Plan emphasizes responsible innovation, risk-based governance, and sector-specific guidelines, including for health [19]. Similarly, Canada has published a compendium of best practices for human-centered AI in the workplace, with relevance to digital health tools [20]. These international developments highlight a growing consensus on the need for agile, risk-proportionate regulation.

III. Proposed Regulatory Frameworks and Architectural Diagrams from Literature

This section presents comprehensive frameworks and architectural diagrams for regulating generative AI-enabled digital mental health devices. These visual models synthesize current regulatory approaches, scholarly insights, and emerging best practices from the literature.

A. Lifecycle-Oriented Regulatory Framework Architecture

B. 4 Lens for Significant Change Assessment

C. Multi-Layered Safety Architecture for Generative AI Chatbots

D. Risk-Tiered Assurance Framework

E. Future Regulatory Development Timeline

F. Multi-Stakeholder Governance Architecture

G. Technical Specifications Summary

Table I summarizes key technical requirements derived from the proposed frameworks:

H. Implementation Roadmap

The proposed frameworks can be implemented through a phased approach:

Phase 1 (Immediate):

- Adopt risk-tiered classification system
- Implement basic safety guardrails [8]
- Establish transparency requirements [25]

Phase 2 (1-2 years):

- Deploy 4 lens for change management [3]
- Create multi-stakeholder governance bodies
- Develop certification programs

Phase 3 (3-5 years):

- Implement automated compliance monitoring
- Establish international regulatory alignment
- Create adaptive licensing pathways

These frameworks provide a comprehensive approach to regulating generative AI in mental health, balancing innovation with patient safety and ethical considerations.

IV. Quantitative Foundations: Mathematical Frameworks and Objective Functions

This section establishes the quantitative foundations for evaluating and regulating generative AI-enabled digital mental health devices. We present mathematical models, objective functions, and quantitative metrics derived from the regulatory frameworks discussed in Section II.

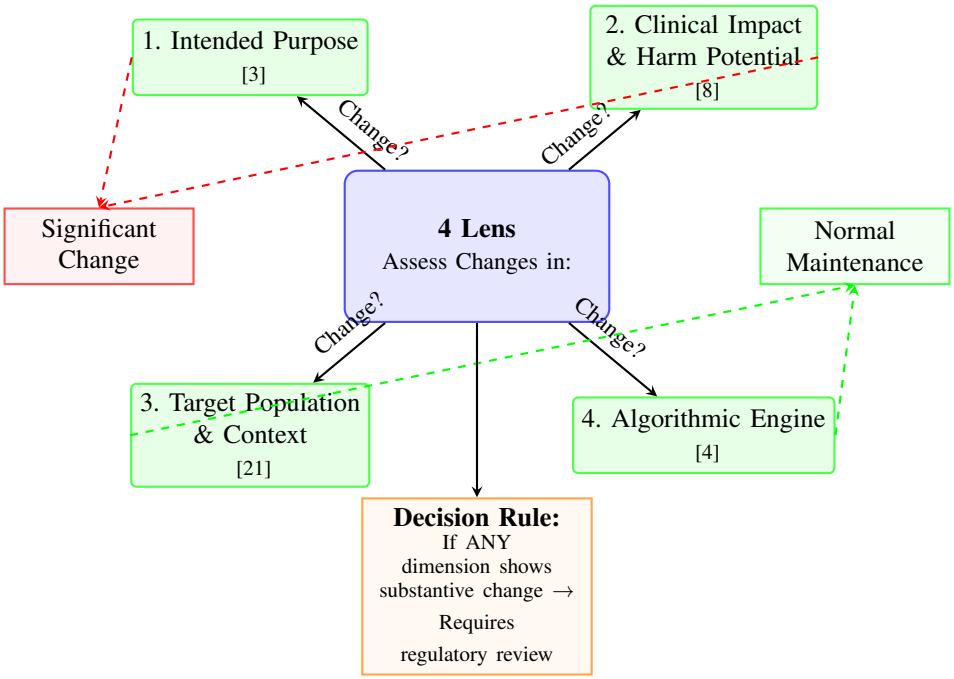

Fig. 1. The 4 Lens Framework for Distinguishing Significant Changes from Normal Maintenance in AI Mental Health Devices [3].

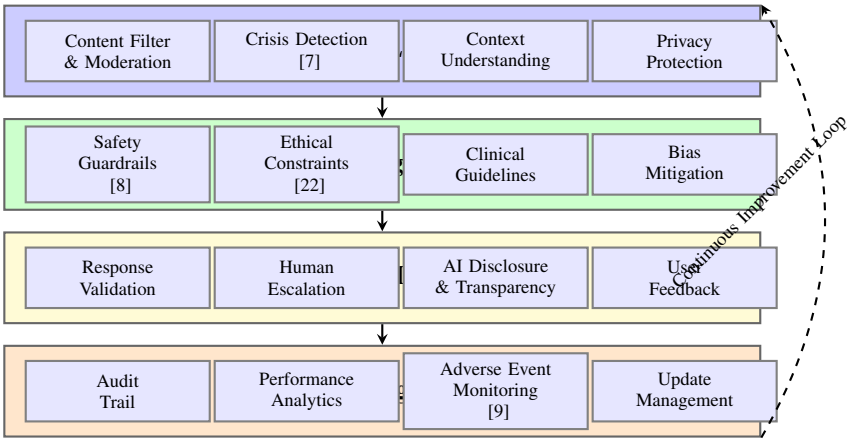

Fig. 2. Multi-Layered Safety Architecture for Generative AI Mental Health Chatbots.

TABLE I
TECHNICAL SPECIFICATIONS FOR GENERATIVE AI MENTAL HEALTH DEVICES

Component	**Minimum Requirements**	**Best Practices**
Safety Guardrails	Content filtering, crisis detection	Multi-layer validation, real-time monitoring
Transparency	AI disclosure, limitations statement	Algorithm explanation, training data disclosure
Data Privacy	Encryption, user consent	Differential privacy, data minimization
Clinical Validation	Pilot study evidence	RCT with diverse populations
Update Management	Change documentation	Automated testing, rollback capability
Interoperability	HL7/FHIR compliance	API for EHR integration
Accessibility	WCAG 2.1 compliance	Multi-modal interfaces
Auditability	Logging of interactions	Immutable audit trail

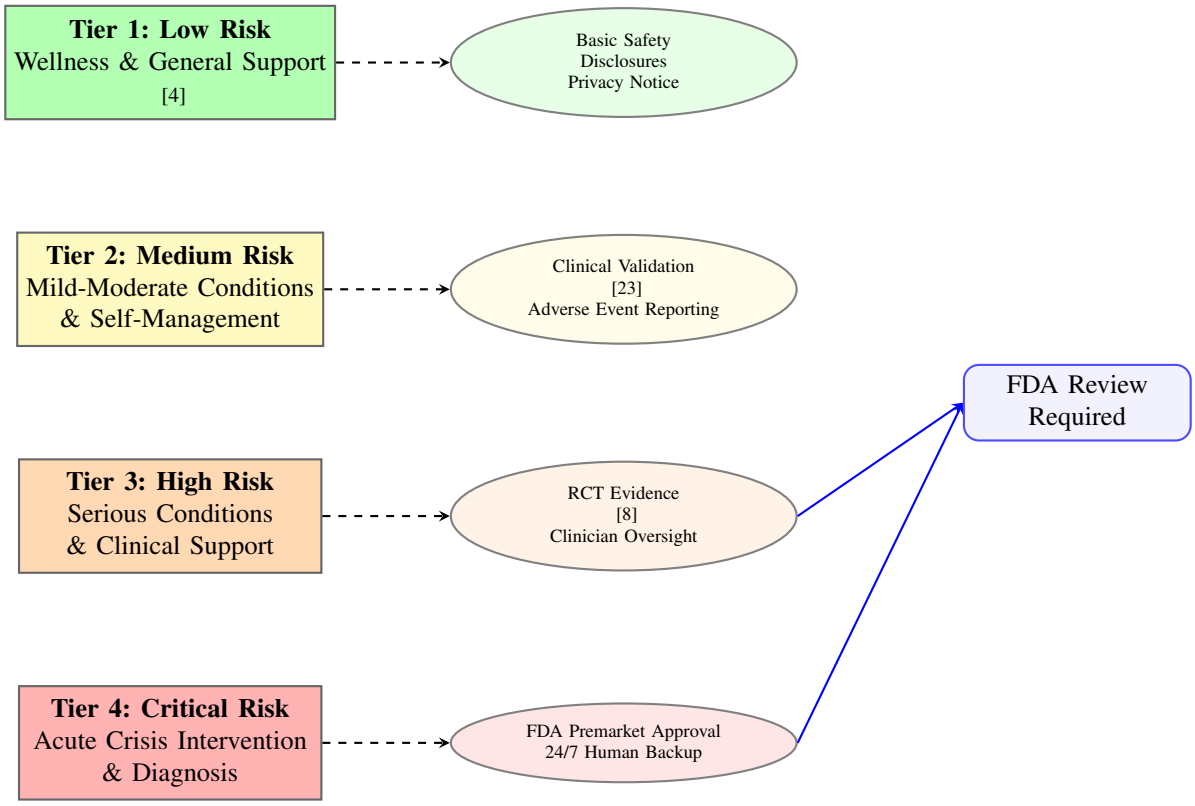

Fig. 3. Alternative Left-to-Right Risk-Tiered Assurance Framework with Proportional Regulatory Requirements.

A. Mathematical Notation and Definitions

Let us define the following mathematical constructs:

- $\mathcal{D}$: The digital mental health device (DMHD) system
- $t \in [0, T]$: Time variable over device lifecycle
- $\theta(t)$: Device parameters at time t (evolving with updates)
- $\mathcal{X}$: Input space (user queries, biometric data, context)
- $\mathcal{Y}$: Output space (therapeutic responses, recommendations)
- $\mathcal{P}$: Patient population with distribution $p(x)$
- $\mathcal{R}$: Risk space (safety, efficacy, ethical risks)

The core generative AI function can be represented as:

$$f_{\theta(t)} : \mathcal{X} \times \mathcal{H} \to \mathcal{Y} \tag{1}$$

where $\mathcal{H}$ represents the conversational history and context.

B. Risk Quantification Framework

1) Safety Risk Metric: The safety risk $R_s(t)$ quantifies potential harm to users:

$$R_s(t) = \mathbb{E}_{x \sim p(x)} \left[\sum_{i=1}^{N} w_i \cdot r_i(x, f_{\theta(t)}(x)) \right] \tag{2}$$

where w_i are weights for different risk categories (suicidality, medical advice, bias) and r_i are risk functions:

$$r_i(x, y) = \begin{cases} 1 & \text{if response violates safety guardrail } i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

2) Efficacy Metric: Clinical efficacy $E_c(t)$ measures therapeutic benefit:

$$E_c(t) = \alpha \cdot \text{PHQ-9}_{\Delta}(t) + \beta \cdot \text{GAD-7}_{\Delta}(t) + \gamma \cdot \text{WSAS}_{\Delta}(t) \tag{4}$$

where α, β, γ are weights, and $\text{PHQ-9}_{\Delta}(t)$ represents change in depression scores from baseline.

C. Objective Functions for Regulatory Optimization

1) Device Developer Objective: Developers aim to maximize utility while minimizing regulatory burden:

$$\max_{\theta} \left[U(\theta) - \lambda_1 R_s(\theta) - \lambda_2 C_r(\theta) \right] \tag{5}$$

where:

$$\begin{aligned} U(\theta) &= \text{User engagement and satisfaction} \\ C_r(\theta) &= \text{Regulatory compliance cost} \\ \lambda_1, \lambda_2 &= \text{Risk aversion parameters} \end{aligned}$$

2) Regulator Objective: Regulators seek to maximize public health benefit while controlling risks:

$$\max_{\pi} \mathbb{E}_{\theta \sim \pi} \left[B(\theta) - \eta_1 R_s(\theta) - \eta_2 R_e(\theta) \right] \tag{6}$$

where π represents the regulatory policy, $B(\theta)$ is population health benefit, and $R_e(\theta)$ represents equity risk.

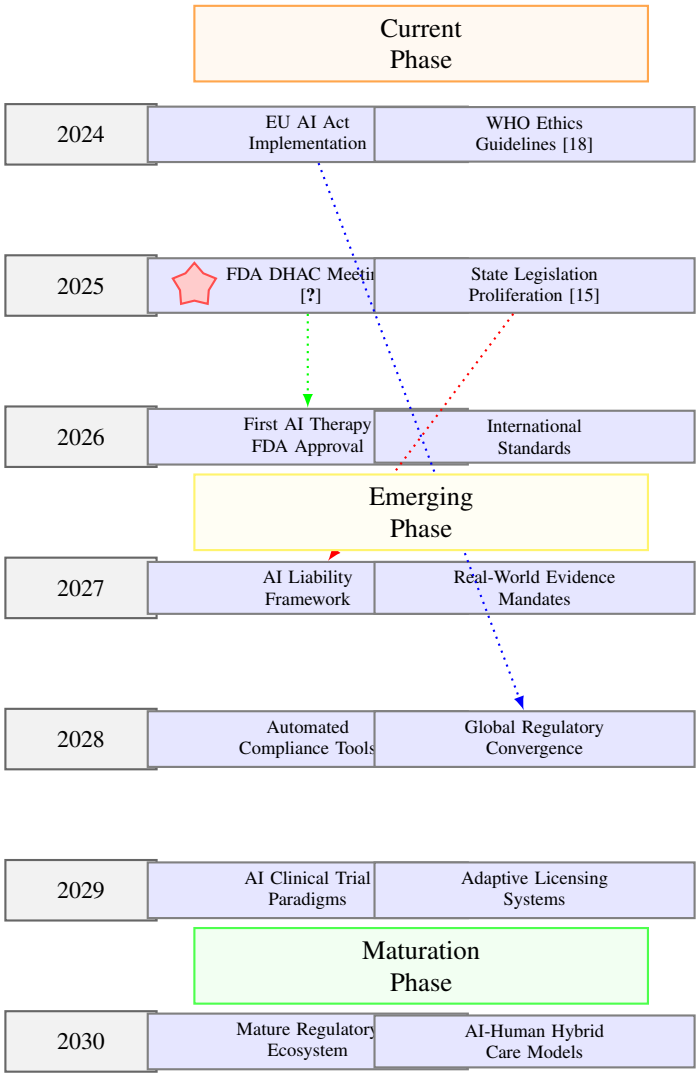

Fig. 4. Projected Timeline for Regulatory Development of Generative AI Mental Health Devices (2024-2030) in a columnar, landscape-style layout. DHAC = Digital Health Advisory Committee.

D. 4 Lens Mathematical Formulation

Based on [3], define significant change indicator function:

$$\Delta(\theta_t, \theta_{t+1}) = \begin{cases} 1 & \text{if } \max_{i \in \{1,2,3,4\}} d_i(\theta_t, \theta_{t+1}) > \tau_i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where:

$$d_1(\theta_t, \theta_{t+1}) = \text{KL-divergence of intended purpose distributions} \tag{8}$$

$$d_2(\theta_t, \theta_{t+1}) = |R_s(\theta_{t+1}) - R_s(\theta_t)| \tag{9}$$

$$d_3(\theta_t, \theta_{t+1}) = \text{Population distribution shift metric} \tag{10}$$

$$d_4(\theta_t, \theta_{t+1}) = \|\theta_{t+1} - \theta_t\|_2 \tag{11}$$

and τ_i are regulatory thresholds.

E. Quality Metrics and Validation Equations

1) Fidelity to Clinical Guidelines: Define fidelity metric $F(t)$:

$$F(t) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}\left\{ f_{\theta(t)}(x_j) \in \mathcal{G}(x_j) \right\} \tag{12}$$

where $\mathcal{G}(x_j)$ is the set of clinically appropriate responses for input x_j.

2) Equity Metric: Following [**?**], define equity score:

$$E_q(t) = 1 - \frac{1}{K} \sum_{k=1}^{K} \left| \frac{E_c^{(k)}(t)}{\bar{E}_c(t)} - 1 \right| \tag{13}$$

where $E_c^{(k)}(t)$ is efficacy for subgroup k, and $\bar{E}_c(t)$ is average efficacy.

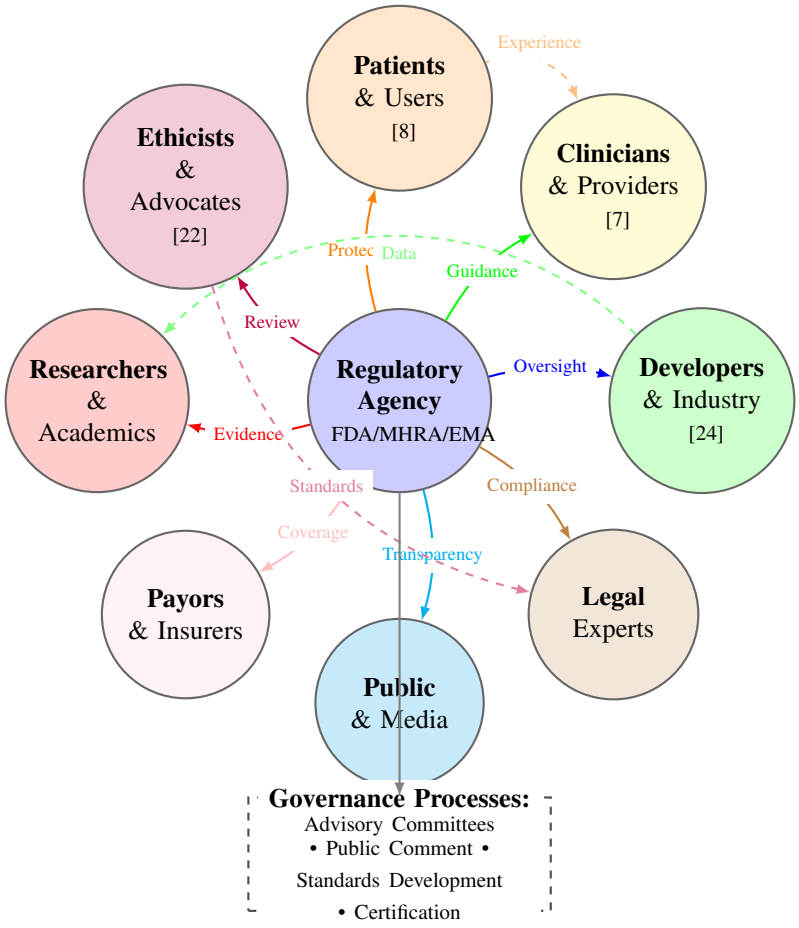

Fig. 5. Multi-Stakeholder Governance Architecture for AI Mental Health Regulation.

F. Regulatory Decision Functions

1) Approval Decision Function: Device approval depends on weighted composite score:

$$A(\mathcal{D}) = \begin{cases} \text{Approve} & \text{if } S_c \geq \tau_a \text{ and } R_s \leq \tau_r \\ \text{Reject} & \text{otherwise} \end{cases} \tag{14}$$

where composite score S_c is:

$$S_c = \sum_{i=1}^{5} w_i m_i \tag{15}$$

with metrics $m_i \in \{\text{Efficacy, Safety, Transparency, Equity, Usability}\}$.

2) Post-Market Surveillance Trigger: Define surveillance intensity $I(t)$:

$$I(t) = \alpha \cdot R_s(t) + \beta \cdot \frac{dR_s}{dt} + \gamma \cdot N_{AE}(t) \tag{16}$$

where $N_{AE}(t)$ is number of adverse events, and $\frac{dR_s}{dt}$ is risk trend.

G. Bayesian Regulatory Framework

For adaptive regulation, use Bayesian updating:

$$P(\mathcal{D} \text{ is safe}|\text{Data}) = \frac{P(\text{Data}|\mathcal{D} \text{ is safe}) \cdot P_0(\mathcal{D} \text{ is safe})}{P(\text{Data})} \tag{17}$$

Posterior risk estimate:

$$R_s^{\text{post}}(t) = \frac{\alpha + N_{\text{harm}}(t)}{\alpha + \beta + N_{\text{total}}(t)} \tag{18}$$

with Beta prior parameters α, β.

H. Multi-Objective Optimization Formulation

The regulatory challenge can be framed as:

$$\min_{\pi} \quad [-B(\pi), R_s(\pi), C(\pi)] \tag{19}$$

$$\text{s.t.} \quad E_q(\pi) \geq \tau_{eq} \tag{20}$$

$$T(\pi) \geq \tau_t \tag{21}$$

where $C(\pi)$ is implementation cost, $T(\pi)$ is transparency score, and τ_{eq}, τ_t are minimum thresholds.

I. Numerical Examples and Simulations

1) Example 1: Risk-Tier Calculation: For a depression chatbot:

$$R_s = 0.05 \cdot r_{\text{suicide}} + 0.03 \cdot r_{\text{medical}} + 0.02 \cdot r_{\text{bias}}$$
$$E_c = 0.6 \cdot \text{PHQ-9}_\Delta + 0.4 \cdot \text{WSAS}_\Delta$$
$$S_c = 0.4E_c + 0.3(1 - R_s) + 0.2T + 0.1E_q$$

If $S_c = 0.78 > \tau_a = 0.7$ and $R_s = 0.10 < \tau_r = 0.15$, device is approved.

2) Example 2: 4 Lens Application: Consider update with parameter changes:

$$d_1 = 0.05 < \tau_1 = 0.10 \quad \text{(no purpose change)}$$
$$d_2 = 0.12 > \tau_2 = 0.10 \quad \text{(safety impact)}$$
$$d_3 = 0.03 < \tau_3 = 0.10 \quad \text{(no population change)}$$
$$d_4 = 0.08 < \tau_4 = 0.15 \quad \text{(minor technical change)}$$

Since $d_2 > \tau_2$, this constitutes significant change requiring review.

J. Validation Metrics from Clinical Studies

From [8], we can derive:

$$\text{Safety Score} = 1 - \frac{N_{\text{adverse}}(t)}{N_{\text{total}}(t)} = 1 - \frac{0}{160} = 1.00 \tag{22}$$

$$\text{Empathy Accuracy} = \frac{\text{Correct empathetic responses}}{\text{Total responses}} = 0.98 \tag{23}$$

K. Regulatory Compliance Cost Function

The cost of compliance $C_c(\mathcal{D})$ can be modeled as:

$$C_c(\mathcal{D}) = C_0 + \sum_{i=1}^{n} c_i \cdot \mathbb{I}\{m_i < \tau_i\} \tag{24}$$

where C_0 is base cost, c_i are penalty costs for failing metric i.

L. Implementation Algorithm

Algorithm for regulatory decision-making:

Algorithm 1 Algorithm for Regulatory Decision Based on Safety, Efficacy, and Equity Scores.

procedure REGULATORYDECISION($\mathcal{D}$, data)
  Compute $R_s \leftarrow \text{SafetyRisk}(\mathcal{D})$
  Compute $E_c \leftarrow \text{EfficacyScore}(\mathcal{D})$
  Compute $E_q \leftarrow \text{EquityScore}(\mathcal{D})$
  Compute $S_c \leftarrow 0.4E_c + 0.3(1 - R_s) + 0.2T + 0.1E_q$
  if $S_c \geq \tau_a \wedge R_s \leq \tau_r$ **then**
    Return APPROVE
  else if $S_c \geq \tau_a - \epsilon$ **then**
    Return CONDITIONAL APPROVAL
  else
    Return REJECT
  end if
end procedure

M. Quantitative Research Questions

Based on the mathematical framework, we propose the following research questions:

1) **RQ1:** What are optimal values for risk thresholds τ_i that balance innovation and safety?

$$\text{Find } \tau^* = \arg\min_{\tau} [\text{Type I error} + \lambda \cdot \text{Type II error}] \tag{25}$$

2) **RQ2:** How should weights w_i in composite score S_c be determined?

$$\text{Optimize } w \text{ to maximize } \mathbb{E}[B(\mathcal{D})|S_c(w) \geq \tau] \tag{26}$$

3) **RQ3:** What is the optimal frequency for post-market surveillance?

$$f^* = \arg\min_f [C_{\text{surveillance}}(f) + \mathbb{E}[\text{Harm}|\text{delay} = 1/f]] \tag{27}$$

4) **RQ4:** How do different risk metrics correlate with actual patient outcomes?

$$\text{Compute } \rho(R_s, \text{ActualHarm}) \text{ across device population} \tag{28}$$

5) **RQ5:** What learning rate for Bayesian updates optimizes regulatory responsiveness?

$$\alpha^*, \beta^* = \arg\min \mathbb{E}[|\hat{R}_s - R_s^{\text{true}}|] \tag{29}$$

N. Empirical Validation Framework

Define validation dataset $\mathcal{V} = \{(x_i, y_i, o_i)\}_{i=1}^N$ where o_i are clinical outcomes.

Performance metrics:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{30}$$
$$\text{Specificity} = \frac{TN}{TN + FP} \tag{31}$$
$$\text{AUC} = \int_0^1 TPR(FPR) dFPR \tag{32}$$

where true positives (TP) are correctly identified high-risk cases.

O. Cost-Benefit Analysis Framework

Net benefit of regulation:

$$NB(\pi) = \underbrace{B(\pi)}_{\text{Benefit}} - \underbrace{C(\pi)}_{\text{Cost}} - \underbrace{\mathbb{E}[\text{Harm}|\pi]}_{\text{Risk cost}} \tag{33}$$

Monetized version:

$$NB_\$(\pi) = \sum_{i=1}^{n} \Delta QALY_i \cdot \$_{QALY} - C_{\text{reg}}(\pi) - \sum_{j=1}^{m} p_j \cdot \text{HarmCost}_j \tag{34}$$

Algorithm 2 Simulation Procedure for Device Regulatory Evaluation

for $i = 1$ to $N_{\text{simulations}}$ **do**
    Sample device parameters $\theta_i \sim p(\theta)$
    Generate outcomes $O_i \sim p(O \mid \theta_i, \pi)$
    Compute metrics $M_i = \{R_s, E_c, E_q, \dots\}$
    Apply decision rule $d_i = A(M_i)$
    Record results
end for
Return $\{\text{Approval rate}, \text{Average benefit}, \text{Risk profile}\}$

P. Simulation Framework for Policy Evaluation

Monte Carlo simulation for policy evaluation:

Q. Conclusion of Quantitative Foundations

The mathematical frameworks presented in this section provide rigorous foundations for:

- Quantifying risks and benefits of generative AI mental health devices
- Optimizing regulatory decision thresholds
- Implementing adaptive, evidence-based regulation
- Balancing innovation with patient safety

These quantitative tools enable data-driven regulation that can evolve with the technology while maintaining rigorous safety standards [3], [8].

V. Safety, Efficacy, and User Experience Evidence

A. Emerging Clinical and Empirical Findings

Recent studies provide preliminary evidence on the safety and user experience of generative AI in mental health. A 2025 exploratory randomized controlled trial (RCT) compared a generative AI conversational agent with a rules-based version for mental health support. The trial found no serious adverse events and similar user satisfaction between groups, with the generative AI arm demonstrating higher accuracy in empathetic response detection (98% vs. 69%) [8]. These results suggest that, with appropriate guardrails, generative AI can be deployed safely while maintaining therapeutic engagement.

However, other research sounds cautionary notes. Evaluations of publicly available large language models (LLMs) in simulated therapeutic scenarios reveal tendencies toward stigmatizing attitudes, inappropriate responses to acute symptoms (e.g., suicidality, psychosis), and a lack of clinical grounding [21]. Some commercial "therapy bots" have been found to endorse harmful suggestions in adolescent crisis vignettes [21]. Such failures underscore the risks of deploying inadequately validated AI systems in sensitive mental health contexts.

The Framework for AI Tool Assessment in Mental Health (FAITA-Mental Health) offers a structured approach to evaluating AI mental health platforms across domains such as credibility, user experience, crisis management, and transparency [23]. Applied to a generative AI obsessive-compulsive disorder (OCD) tool, the framework identified strengths in conversational ability but gaps in clinical validation and crisis protocols [23]. This highlights the need for standardized assessment tools that can guide developers, regulators, and users.

B. Technical Guardrails and Risk Mitigation

Ensuring safety in generative AI mental health tools requires robust technical and procedural guardrails. These include input filters to detect concerning language (e.g., self-harm, abuse), output controls to prevent the provision of medical advice or diagnoses, and mechanisms for escalating high-risk situations to human providers [7], [8]. Post-market monitoring is critical to identify emergent risks, especially as models evolve through continuous learning or updates [3], [9].

The concept of "purpose envelopes" and "risk-tiered assurance lanes" has been proposed to differentiate between minor updates and significant changes that alter a device's intended purpose, clinical impact, or algorithmic behavior [3]. Such distinctions can help regulators apply appropriate oversight throughout a product's lifecycle, balancing flexibility with safety.

VI. Ethical, Governance, and Transparency Imperatives

A. Ethics of Care and Accountability

Beyond technical safety, generative AI in mental health raises profound ethical questions. The dominant "responsible AI" paradigm, focused on principles like fairness and transparency, may overlook the relational dimensions of care [22]. An ethics-of-care perspective emphasizes the developer's duty of care toward users, the potential for emotional manipulation, and the importance of preserving human therapeutic relationships [6], [22]. This approach calls for clear accountability mechanisms, including liability frameworks for deployers and licensure considerations for AI-augmented care [15].

Transparency is a cornerstone of ethical AI deployment. Users should be informed when they are interacting with an AI system, understand its limitations, and have access to information about data use, algorithmic functioning, and validation evidence [23], [25]. Regulatory frameworks must mandate such disclosures to support informed consent and trust.

B. Data Governance and Equity

Mental health data is exceptionally sensitive, warranting stringent privacy protections. Current U.S. laws, such as HIPAA, may not fully cover data collected by consumer wellness apps, creating privacy risks [26], [27]. Emerging state laws and international regulations (e.g., EU AI Act) emphasize data minimization, purpose limitation, and security for health AI systems [15], [16].

Equity considerations are paramount. AI models trained on non-representative data may perpetuate biases, disadvantaging marginalized populations [1], [21]. Regulatory oversight should require equity assessments and inclusive design practices to ensure that AI mental health tools benefit diverse user groups.

VII. Toward a Lifecycle-Oriented Regulatory Framework

A. Pre-Market Evaluation and Evidence Standards

Pre-market evaluation of generative AI mental health devices should be risk-based, with higher scrutiny for tools intended for higher-acuity conditions or autonomous operation. Evidence standards must adapt to the iterative nature of AI, potentially incorporating real-world performance data, simulation studies, and ongoing validation during development [24], [28]. The FDA's proposed use of PCCPs allows manufacturers to pre-specify certain types of modifications (e.g., performance improvements, bias mitigation) without new submissions, provided they remain within predefined bounds [9].

International harmonization of evidence requirements can reduce developer burden and accelerate global access. Initiatives like the International Medical Device Regulators Forum (IMDRF) are working toward aligned guidelines for AI SaMD, which should incorporate mental health-specific considerations.

B. Post-Market Surveillance and Real-World Monitoring

Post-market surveillance is especially critical for generative AI tools due to their adaptive nature and potential for emergent risks. The FDA's guidance on post-market updates to AI-enabled devices encourages continuous monitoring and real-world performance tracking [9], [29]. Strategies may include:

- Real-world evidence collection through registries, user feedback, and electronic health record linkages;
- Automated monitoring of conversation logs (with privacy safeguards) to detect safety signals;
- Periodic audits of algorithmic performance and equity metrics;
- Clear protocols for incident reporting and corrective actions.

The "4 lens" proposed by Nandagopal et al. offers a pragmatic tool for distinguishing normal maintenance from significant changes in DMHTs, based on shifts in intended purpose, clinical impact, target population, or algorithmic behavior [3]. Regulators could use such frameworks to trigger review requirements when changes exceed predetermined thresholds.

C. Multi-Stakeholder Governance and Professional Engagement

Effective governance of AI in mental health requires collaboration among regulators, developers, clinicians, patients, and ethicists. Professional organizations, such as the American Medical Association (AMA) and American Psychological Association (APA), are developing guidelines and advocacy positions [7], [30]. Clinician involvement in regulatory discussions—as highlighted in the FDA's advisory committee meeting—is essential to ground policies in clinical reality [15].

Industry self-regulation, through standards and certifications, can complement government oversight. Initiatives like the Digital Technology Assessment Criteria (DTAC) in the UK's National Health Service provide a model for evaluating digital health tools against clinical safety, data protection, and usability standards [3].

VIII. Author's Related Work

This research builds upon and extends the author's previous investigations into artificial intelligence governance, regulation, and implementation frameworks across various domains. The following papers provide foundational insights that inform the current analysis of generative AI in digital mental health:

- **Regulatory Reform for Agentic AI** [31]: This paper examines regulatory barriers to AI innovation identified in the Office of Science and Technology Policy's Request for Information. It proposes a comprehensive governance framework integrating technical standards, risk management protocols, and policy recommendations for modernizing federal regulations to foster responsible AI innovation while maintaining public trust and safety. The analysis of regulatory mismatches, structural incompatibility, and governance challenges in this work informs the current paper's approach to addressing regulatory gaps in mental health AI.
- **Framework for Government Policy on Agentic and Generative AI in Healthcare** [32]: This comprehensive review examines the dichotomy between open-source and proprietary AI models in healthcare, with particular focus on the emerging challenges of autonomous Agentic Generative AI (AGI). The paper proposes a tiered risk-management and governance framework that synthesizes the strengths of both open and closed-source approaches. The healthcare-specific governance recommendations, including international certification protocols, federated learning architectures, and adaptive policymaking, provide important context for the mental health-specific regulatory frameworks proposed in the current paper.
- **A Comprehensive Framework for U.S. AI Export Leadership** [33]: This analysis examines the American AI Exports Program through a multi-dimensional framework encompassing technical architecture, governance structures, market strategy, and policy implementation. The paper presents a multi-layer framework architecture with strategic, governance, technical, and market layers, supported by detailed visualizations including architectural diagrams, decision matrices, risk assessment frameworks, and implementation roadmaps. The integrated approach to balancing innovation acceleration with risk mitigation in strategic competition contexts informs the current paper's methodology for developing comprehensive regulatory frameworks.

These previous works collectively establish several key principles that guide the current research:

1) **Multi-stakeholder governance** requiring collaboration between technical, regulatory, and clinical domains
2) **Risk-proportional approaches** that scale oversight according to potential harm

3) **Adaptive regulatory frameworks** capable of evolving with technological advances
4) **International harmonization** to address global nature of digital health technologies
5) **Evidence-based policymaking** grounded in empirical validation and real-world outcomes

The current paper extends these principles specifically to the domain of generative AI in digital mental health, addressing unique challenges such as therapeutic relationships, crisis management, and the ethical dimensions of mental healthcare delivery. By building upon the governance frameworks established in previous work while addressing domain-specific considerations, this research contributes to a cohesive body of literature on responsible AI implementation across different sectors and applications.

IX. Conclusion and Recommendations

The integration of generative artificial intelligence into digital mental health represents both a transformative opportunity for healthcare access and a significant regulatory challenge requiring thoughtful, evidence-based approaches. This paper has presented comprehensive frameworks, quantitative foundations, and practical implementation strategies for regulating generative AI-enabled digital mental health devices in response to the U.S. FDA's call for stakeholder input. Our analysis reveals that current regulatory structures, designed primarily for static medical devices, are ill-equipped to address the dynamic, evolving nature of AI-driven mental health technologies.

The proposed frameworks offer several key advancements over existing approaches. The lifecycle-oriented regulatory architecture addresses the continuous evolution of AI systems, while the 4 lens framework provides a practical methodology for distinguishing between routine maintenance and significant changes requiring regulatory review. The risk-tiered assurance system establishes proportional oversight based on clinical risk, ensuring that resources are allocated appropriately. The multi-layered safety architecture offers technical implementation guidance for developers, and the multi-stakeholder governance model emphasizes the importance of inclusive decision-making processes.

Our quantitative foundations provide mathematical rigor to regulatory decision-making, offering objective functions for balancing innovation and safety, formalizing risk assessment methodologies, and establishing evidence-based thresholds for approval decisions. The clinical evidence reviewed, including both promising results from controlled trials and concerning findings from real-world evaluations, underscores the dual nature of these technologies—offering genuine therapeutic potential while presenting substantial risks if inadequately regulated.

Based on our comprehensive analysis, we offer the following recommendations for regulators, developers, and policymakers:

1. Establish Adaptive Regulatory Pathways: Regulatory agencies should implement flexible, evidence-based approval processes that accommodate the iterative nature of AI development while maintaining rigorous safety standards. This includes adopting predetermined change control plans and creating regulatory sandboxes for controlled innovation.

2. Implement Risk-Proportional Oversight: A tiered regulatory approach should be established, with requirements scaled according to clinical risk, intended use, and target population. Low-risk wellness applications should have minimal oversight, while systems addressing serious mental health conditions or operating with significant autonomy should undergo comprehensive pre-market review and continuous post-market surveillance.

3. Strengthen Post-Market Surveillance Systems: Given the adaptive nature of AI systems, robust post-market monitoring is essential. Regulators should require real-world performance tracking, mandatory adverse event reporting, and periodic re-evaluation of safety and effectiveness as systems evolve.

4. Enhance Transparency and Accountability: Regulatory frameworks should mandate clear disclosures about AI system capabilities, limitations, and intended uses. Developers should be required to maintain comprehensive audit trails, provide explanations for critical decisions, and establish clear accountability mechanisms for system outputs.

5. Prioritize Equity and Access: Regulatory processes should actively address potential biases in AI systems and ensure that benefits are distributed equitably across diverse populations. This includes requiring representative training data, equitable performance validation, and accessibility considerations in system design.

6. Foster Multi-Stakeholder Collaboration: Effective governance requires collaboration between regulators, developers, clinicians, patients, ethicists, and payers. Advisory committees, public comment processes, and collaborative standard-setting initiatives should be institutionalized to ensure balanced perspectives inform regulatory decisions.

7. Promote International Harmonization: Given the global nature of digital health technologies, regulators should work toward aligned standards and mutual recognition agreements to reduce barriers to innovation while maintaining consistent safety standards across jurisdictions.

8. Invest in Research and Education: Continued research is needed to establish evidence-based regulatory thresholds, validate safety frameworks, and understand long-term outcomes. Professional education programs should prepare clinicians to effectively use and critically evaluate AI-assisted tools.

The September 2025 FDA Digital Health Advisory Committee meeting represents a critical inflection point for establishing coherent regulatory approaches for generative AI in mental health. As these technologies continue to evolve, regulatory frameworks must remain agile, evidence-informed, and responsive to both technological advancements and emerging safety concerns. By implementing the recommendations outlined in this paper, regulators can foster innovation that meaningfully addresses mental healthcare needs while es-

tablishing robust safeguards that protect patient welfare and maintain public trust in these transformative technologies.

Ultimately, the goal is not merely to regulate AI in mental health, but to cultivate an ecosystem where technological innovation, clinical expertise, regulatory oversight, and patient-centered values converge to create safe, effective, and accessible mental healthcare solutions for all who need them.

List of Tables and Figures

Tables

- **Table I**: Technical Specifications for Generative AI Mental Health Devices - Summarizes key technical requirements including safety guardrails, transparency, data privacy, clinical validation, update management, interoperability, accessibility, and auditability.

Figures

- **Figure 1**: The 4 Lens Framework for Distinguishing Significant Changes from Normal Maintenance in AI Mental Health Devices - Illustrates the decision framework for assessing changes in intended purpose, clinical impact, target population, and algorithmic engine to determine if regulatory review is required.
- **Figure 2**: Multi-Layered Safety Architecture for Generative AI Mental Health Chatbots - Presents a four-layer safety architecture with input, processing, output, and monitoring layers, each containing specific safety components and guardrails.
- **Figure 3**: Alternative Left-to-Right Risk-Tiered Assurance Framework with Proportional Regulatory Requirements - Shows a four-tier risk classification system (low, medium, high, critical) with corresponding regulatory requirements that increase with risk level.
- **Figure 4**: Projected Timeline for Regulatory Development of Generative AI Mental Health Devices (2024-2030) - Visualizes the anticipated regulatory milestones and developments in a columnar, landscape-style layout across three phases: current, emerging, and maturation.
- **Figure 5**: Multi-Stakeholder Governance Architecture for AI Mental Health Regulation - Depicts a network model with the regulatory agency at the center, connected to eight key stakeholder groups (developers, clinicians, patients, ethicists, researchers, payors, public, legal experts) with specific interaction types.

Algorithms

- **Algorithm 1**: Algorithm for Regulatory Decision Based on Safety, Efficacy, and Equity Scores - Presents a procedural algorithm for making regulatory approval decisions using weighted composite scoring of safety, efficacy, transparency, and equity metrics.
- **Algorithm 2**: Simulation Procedure for Device Regulatory Evaluation - Outlines a Monte Carlo simulation framework for evaluating regulatory policies through repeated sampling of device parameters and outcome generation.

This document contains 1 table, 5 figures, and 2 algorithms that collectively provide visual and procedural representations of the proposed regulatory frameworks for generative AI-enabled digital mental health devices.

Acknowledgments

This research was prepared in response to the National Institute of Standards and Technology (NIST) Request for Information (RFI) on the development of a National Strategic Plan for Advanced Manufacturing (Docket NIST-2025-0004, FR-2025-11379). The paper was submitted as a formal public comment to be considered by the Subcommittee on Advanced Manufacturing of the National Science and Technology Council and the Office of Science and Technology Policy (OSTP). The author acknowledges NIST and OSTP's commitment to transparent, stakeholder-informed policy development and thanks the agencies for the opportunity to contribute to this important national strategic planning process.

Submission Details:

- **Docket:** NIST-2025-0004
- **Federal Register Notice:** FR-2025-11379
- **Posted:** June 20, 2025
- **Comment Period Ends:** December 15, 2025 at 11:59 PM EST
- **Submission Portal:** http://www.regulations.gov

Disclaimer: This response is voluntary and submitted in accordance with the RFI guidelines. All information provided may be posted on https://www.regulations.gov or otherwise made publicly available. This submission does not constitute a binding commitment to develop or pursue any projects or ideas discussed herein.

Declaration

This work is exclusively a survey paper synthesizing existing published research. No novel experiments, data collection, or original algorithms were conducted or developed by the authors. All content, including findings, results, performance metrics, architectural diagrams, and technical specifications, is derived from and attributed to the cited prior literature. The authors' contribution is limited to the compilation, organization, and presentation of this pre-existing public knowledge. Any analysis or commentary is based solely on the information contained within the cited works. Figures and tables are visual representations of data and concepts described in the referenced sources.

References

[1] E. E. Lee, J. Torous, M. De Choudhury, C. A. Depp, S. A. Graham, H.-C. Kim, M. P. Paulus, J. H. Krystal, and D. V. Jeste, "Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom," vol. 6, no. 9, pp. 856–864. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S245190222100046X

[2] L. Eliot. Generative AI For Mental Health Is Upping The Ante By Going Multi-Modal, Embracing E-Wearables, And A Whole Lot More. Forbes. [Online]. Available: https://www.forbes.com/sites/lanceeliot/2023/11/02/generative-ai-for-mental-health-is-upping-the-ante-by-going-multi-modal-embracing-e

[3] S. Nandagopal. Beyond Static Products: Rethinking Regulatory Hurdles and. [Online]. Available: https://papers.ssrn.com/abstract=5824142

[4] J. De Freitas and I. G. Cohen, "The health risks of generative AI-based wellness apps," vol. 30, no. 5, pp. 1269–1275. [Online]. Available: https://www.nature.com/articles/s41591-024-02943-6

[5] V. K. Malesu. Chatbots for mental health pose new challenges for US regulatory framework. News-Medical. [Online]. Available: https://www.news-medical.net/news/20240501/Chatbots-for-mental-health-pose-new-challenges-for-US-regulatory-framework.aspx

[6] "Regulating AI in Mental Health: Ethics of Care Perspective," vol. 11. [Online]. Available: https://www.sciencedirect.com:5037/org/science/article/pii/S2368795924001033

[7] Health advisory: Use of generative AI chatbots and wellness applications for mental health. https://www.apa.org. [Online]. Available: https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-chatbots-wellness-apps

[8] T. R. Campellone, M. Flom, R. M. Montgomery, L. Bullard, M. C. Pirner, A. Pavez, M. Morales, D. Harper, C. Oddy, T. O'Connor, J. Daniels, S. Eaneff, V. L. Forman-Hoffman, C. Sackett, and A. Darcy, "Safety and User Experience of a Generative Artificial Intelligence Digital Mental Health Intervention: Exploratory Randomized Controlled Trial," vol. 27, no. 1, p. e67365. [Online]. Available: https://www.jmir.org/2025/1/e67365

[9] FDA issues final guidance on postmarket updates to AI-enabled devices — MedTech Dive. [Online]. Available: https://www.medtechdive.com/news/fda-final-guidance-predetermined-change-control-plans-ai/734608/

[10] FDA Oversight: Understanding the Regulation of Health AI Tools • Bipartisan Policy Center. Bipartisan Policy Center. [Online]. Available: https://bipartisanpolicy.org/issue-brief/fda-oversight-understanding-the-regulation-of-health-ai-tools/

[11] FDA Digital Health Meeting Highlights Pre- And Post- Market Proposals for Generative AI Devices — InsideHealthPolicy.com. [Online]. Available: https://insidehealthpolicy.com/inside-telehealth-daily-news/fda-digital-health-meeting-highlights-pre-and-post-market-proposals

[12] FDA's Digital Health Advisory Committee Considers Generative AI Therapy Chatbots for Depression. [Online]. Available: https://www.orrick.com/en/Insights/2025/11/FDAs-Digital-Health-Advisory-Committee-Considers-Generative-AI-Therapy-Chatbots-for-Depression

[13] U.S. FDA and CMS Actions on Generative AI-Enabled Mental Health Devices Yield Insights Across AI Product Development. [Online]. Available: https://www.sidley.com/en/insights/newsupdates/2025/11/us-fda-and-cms-actions-on-generative-ai-enabled-mental-health-devices-yield-insights-across-ai

[14] M. M. Mello and I. G. Cohen, "Regulation of Health and Health Care Artificial Intelligence," vol. 333, no. 20, pp. 1769–1770. [Online]. Available: https://doi.org/10.1001/jama.2025.3308

[15] J. N. Shumate, E. Rozenblit, M. Flathers, C. A. Larrauri, C. Hau, W. Xia, E. N. Torous, and J. Torous, "Governing AI in Mental Health: 50-State Legislative Review," vol. 12, no. 1, p. e80739. [Online]. Available: https://mental.jmir.org/2025/1/e80739

[16] Artificial intelligence. European Commission - European Commission. [Online]. Available: https://ec.europa.eu/commission/presscorner

[17] The regulation of artificial intelligence as a medical device: Government response to the Regulatory Horizons Council. GOV.UK. [Online]. Available: https://www.gov.uk/government/publications/the-regulation-of-artificial-intelligence-as-a-medical-device-government-response-to-the-rhc/the-regulation-of-artificial-intelligence-as-a-medical-device-government-response-to-the-regulatory-horizons-council

[18] WHO releases AI ethics and governance guidance for large multi-modal models. [Online]. Available: https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models

[19] M.-S. Burrett, C. Gordon, and J. McQuillen. Australia introduces a national AI plan: Four things leaders need to know. Lexology. [Online]. Available: https://www.lexology.com/library/detail.aspx?g=a200e0e5-1829-4b27-89de-e426a345a201

[20] E. a. S. D. Canada. Compendium of best practices for a human-centered development and use of Artificial Intelligence in the world of work. [Online]. Available: https://www.canada.ca/en/employment-social-development/corporate/reports/2025-best-practices-artificial-intelligence.html

[21] F. C. Ohu, D. N. Burrell, and L. A. Jones, "Public Health Risk Management, Policy, and Ethical Imperatives in the Use of AI Tools for Mental Health Therapy," vol. 13, no. 21, p. 2721. [Online]. Available: https://www.mdpi.com/2227-9032/13/21/2721

[22] T. Tavory, "Regulating AI in Mental Health: Ethics of Care Perspective," vol. 11, no. 1, p. e58493. [Online]. Available: https://mental.jmir.org/2024/1/e58493

[23] A. Golden and E. Aboujaoude, "Describing the Framework for AI Tool Assessment in Mental Health and Applying It to a Generative AI Obsessive-Compulsive Disorder Platform: Tutorial," vol. 8, no. 1, p. e62963. [Online]. Available: https://formative.jmir.org/2024/1/e62963

[24] Building a Comprehensive AI Governance Framework in Life Sciences. JD Supra. [Online]. Available: https://www.jdsupra.com/legalnews/building-a-comprehensive-ai-governance-7275037/

[25] visceral_dev_admin. Transparency in Health Care AI: A Conversation with Experts. Bipartisan Policy Center. [Online]. Available: https://bipartisanpolicy.org/article/transparency-in-health-care-ai-conversation-with-experts/

[26] G. L. Group. International Comparative Legal Guides. International Comparative Legal Guides International Business Reports. [Online]. Available: https://iclg.com/practice-areas/digital-health-laws-and-regulations/usa

[27] ——. International Comparative Legal Guides. International Comparative Legal Guides International Business Reports. [Online]. Available: https://iclg.com/practice-areas/digital-health-laws-and-regulations/03-artificial-intelligence-tools-in-health-services-an-overview-of-current-and-evolving

[28] (1) Regulating Generative AI-Enabled Medical Devices - Part 1 - Premarket Performance Evaluation — LinkedIn. [Online]. Available: https://www.linkedin.com/pulse/regulating-generative-ai-enabled-medical-devices-1-conjeti-phd-00rfe/

[29] (1) Regulating Generative AI-Enabled Medical Devices - Part 3 - Post Market Monitoring — LinkedIn. [Online]. Available: https://www.linkedin.com/pulse/regulating-generative-ai-enabled-medical-devices-3-conjeti-phd-xzt5e/

[30] Advocacy Insights: The current and future landscape of health care AI policy. American Medical Association. [Online]. Available: https://www.ama-assn.org/member-benefits/events/advocacy-insights-current-and-future-landscape-health-care-ai-policy

[31] S. Joshi, "Regulatory Reform for Agentic AI: Addressing Governance Challenges in Federal AI Adoption." [Online]. Available: https://zenodo.org/records/17808694

[32] ——. Framework for Government Policy on Agentic and Generative AI in Healthcare: Governance, Regulation, and Risk Management of Open-Source and Proprietary Models. [Online]. Available: https://www.preprints.org/manuscript/202509.1087

[33] ——, "A Comprehensive Framework for U.S. AI Export Leadership: Analysis, Implementation, and Strategic Recommendations." [Online]. Available: https://zenodo.org/records/17823269

About the Author

Satyadhar Joshi is a quantitative analyst with expertise in financial risk, data science, machine learning, and artificial intelligence. He currently serves as an Assistant Vice President at Bank of America. His independent research explores AI-driven risk assessment, financial modeling, and big data analytics, with particular interest in improving uncertainty modeling tools for AI development for US National Interest.