October 16, 2025

Division of Dockets Management (HFA-305)
Food and Drug Administration
5630 Fishers Lane, Rm. 1061
Rockville, MD 20852

Re: Comments on Safe Deployment of AI in Behavioral Health; Request for Clear Guardrails for Conversational AI Used in Therapy Support

Dear Sir or Madam:

Talkspace respectfully submits this comment to support the Food and Drug Administration's (FDA) efforts to ensure that artificial intelligence (AI) used in behavioral health is safe, effective, and trusted. We encourage the FDA to adopt a risk-based framework tailored to conversational, behavioral-health AI (BH-AI) that: (1) codifies mandatory human-in-the-loop escalation for risk; (2) requires transparent LLM data source origins, model change control (PCCP) and ongoing post-market monitoring; (3) specifies minimum evidence, evaluation, and reporting metrics for therapeutic quality and safety; and (4) sets privacy, security, and transparency expectations commensurate with the vulnerability of this population.

Talkspace has developed a behavioral-health-specific large language model (LLM) designed to serve as a therapy companion and clinical support tool. In alpha testing, the model demonstrated substantial, measurable improvements versus general-purpose baselines on both safety and therapeutic-quality metrics. We offer these data points and our safety program as a blueprint for what good looks like and propose concrete guardrails FDA can require across the category.

Talkspace (NASDAQ: TALK) is a leading virtual behavioral health provider serving individuals, couples, and teens nationwide, with payer, employer, and government partners. Care is delivered through a HIPAA-compliant web and mobile platform by thousands of licensed clinicians. Talkspace has responsibly deployed AI for years to support clinicians and members (e.g., risk detection, session insights, smart notes) under rigorous governance and with licensed-clinician oversight.

We believe that any behavioral-health AI companion should provide psychoeducation, skills reinforcement, reflective prompts, and supportive check-ins between sessions; continuously screens for safety risk and escalates to licensed clinicians when appropriate. It should not establish a diagnosis or replace human licensed care.

Any model offered to the FDA should be transparent that the source data used for training has a clinical origination as many today were built on Chat GPT source material only. Talkspace trained its LLM on hundreds of millions of tokens from anonymized, graded Talkspace therapy transcripts, plus thousands of adversarial/red-team scenarios focused on self-harm,

hallucinations, OCD, mania, and delusion. Guardrails, intent classifiers, and risk-detection layers are deployed around the generative core.

For example, the Talkspace Alpha testing highlight demonstrate baselines for FDA application testing:

- Safety: ~50% improvement vs. base model in identifying and responding to high-risk behaviors; hard blocks on disallowed content; deterministic crisis-routing pathways.

- Therapeutic quality: ~47% higher scores on a modified Cognitive Therapy Rating Scale (CTRS) across factors such as collaborative agenda setting, active listening, appropriate CBT techniques, cultural sensitivity, warmth/bonding, and risk response quality.

- User experience: ~3× higher user-reported satisfaction vs. base model in early evaluations.

Further, we recommend FDA establish the following baseline guardrails for BH-AI systems intended for patient use or clinician-supervised use. Talkspace has implemented or committed to each requirement below.

1) Clear Scope, Labeling, and Off-Ramps

- Explicit, consumer-facing labeling of intended use, benefits, and limitations; statements clarifying that the system does not diagnose, prescribe, or replace licensed therapy.

- Product UX that continuously discloses the AI nature of responses and provides single-tap access to human help.

- Built-in, “never alone with imminent risk” rules: when risk signals meet predefined thresholds, automatic escalation to licensed clinicians or emergency pathways.

2) Human-in-the-Loop Safety and Clinical Governance

- 24/7 clinician coverage for high-risk escalation, with defined service-level objectives (e.g., time-to-human contact for suicidal ideation within minutes).

- Clinician review of flagged conversations; ability to override AI, annotate, and provide feedback signals that retrain safety classifiers.

- Independent clinical safety board to review incidents, trends, and proposed model changes; root-cause analysis and corrective actions for all serious events.

3) Evidence Standards and Required Metrics

Minimum pre-market and post-market evidence should include:

- Safety performance: sensitivity/specificity, false-negative rate for suicide/self-harm detection, and time-to-escalation; report per-1,000-interaction incident rates.

- Therapeutic quality: validated scales (e.g., CTRS-derived measures) scored by blinded raters; longitudinal engagement and clinical improvement proxies where appropriate.

- Fairness: subgroup performance across age, race/ethnicity, gender identity, sexual orientation, disability; parity thresholds and mitigation plans.

- Reliability: robust adversarial/red-team test suite covering delusions, mania, OCD, self-harm, substance use, and abuse; publish pass/fail criteria and holdout results.

- Usability: comprehension of warnings/labels by lay users; mis-use and over-trust testing; “safe hand-off” evaluation to human care.

4) Predetermined Change Control Plan (PCCP)

- A documented PCCP that enumerates: (i) what may change (training data, prompts, safety layers); (ii) how changes are verified/validated (acceptance criteria, bias checks, regression gates); and (iii) how users and clinicians are notified of material updates.

- Separate safety and quality gates with “no-regret” rollback procedures; shadow-mode evaluation before activating changes.

5) Post-Market Surveillance and Reporting

- Real-time safety telemetry; monthly signal-detection reviews; quarterly public safety summaries (de-identified).

- Mandatory reporting and investigation of serious adverse events; commitments to share de-identified incident taxonomies with regulators and peers to improve collective safety.

6) Privacy, Security, and Data Governance

- HIPAA-compliant environment; least-privilege access; encryption in transit and at rest; red-teaming for data exfiltration and prompt-injection threats.

- Data minimization; no use of PHI for unrelated advertising; clear retention and deletion policies; member controls over data use in model improvement.

7) Equity, Accessibility, and Cultural Responsiveness

- Inclusive training sets and active harm-avoidance prompts; bilingual/ASL pathways where feasible; tailored resources aligned to identity and culture.

- Ongoing disparity monitoring with corrective re-weighting or rule updates when subgroup gaps exceed thresholds.

8) Interoperability and Crisis Integration

- Seamless hand-off to licensed Talkspace clinicians within the same platform; direct routing to crisis services when indicated; auditable logs for continuity of care.

Summary Comments to FDA

1. Define a BH-AI category (conversational, therapy-support AI) with clear risk-based expectations distinct from general wellness chatbots and from diagnostic/therapeutic SaMD. Require true clinical data source treaining.

2. Require human-in-the-loop escalation for any BH-AI that screens or triages safety risk.

3. Adopt minimum evaluation metrics (safety, quality, fairness, reliability, usability) and require publishable performance summaries.

4. Mandate PCCPs for adaptive BH-AI, with explicit validation gates and rollback plans.

5. Encourage transparent incident learning, including common, de-identified taxonomies for reporting near-misses and harms in behavioral health contexts.

6. Promote transparency artifacts, including model card-like summaries tailored to clinicians and patients, with clear scope, limitations, and update histories.

We appreciate FDA's leadership in advancing safe, effective AI. Behavioral health presents unique risks and vulnerabilities; it also offers profound opportunities to expand access, continuity, and safety when AI is deployed with the right guardrails. Talkspace supports enforceable, evidence-based standards and offers the above blueprint as a practical path to protect patients while enabling innovation.

Sincerely,

Dr. Jon Cohen

CEO-Talkspace

Appendix A — Evaluation & Reporting Metrics (Illustrative)

Safety:

- FN rate for self-harm/suicide intent ≤ X%; time-to-human review ≤ Y minutes at P95; incident rate reporting per 1,000 interactions.

Therapeutic Quality:

- CTRS-derived composite score relative to clinician benchmarks; longitudinal engagement and symptom proxy trends where ethically feasible.

Fairness:

- Max subgroup delta in safety sensitivity ≤ Δ; action plan if exceeded.

Reliability:

- Red-team pass rate ≥ threshold across mania, delusion, OCD, substance use, abuse, medication misuse, and psychosis prompts.

Usability:

- Comprehension of AI limitations ≥ threshold; safe-handoff success rate; measures of over-trust and automation bias.

Appendix B — Predetermined Change Control Plan (Outline)

- What can change: prompts/guardrails, safety classifiers, risk thresholds, training data additions, model weights (within defined bounds).
- How validated: offline eval battery, subgroup parity checks, clinician panel review, shadow-mode A/B with safety holdouts, rollback criteria.
- How communicated: release notes for clinicians; in-product notices for members when material changes affect safety-relevant behavior.