DATA&
SOCIETY

Date: December 5, 2025

Via electronic submission to regulations.gov

Re: Docket No. FDA-2025-N-2338
Digital Health Advisory Committee; Notice of Meeting; Establishment of a Public Docket; Request for Comments — Generative Artificial Intelligence-Enabled Digital Mental Health Medical Devices

Dear Deputy Commissioner Graham and Members of the Digital Health Advisory Committee:

Data & Society Research Institute (D&S) appreciates the opportunity to submit comments to the Food and Drug Administration (FDA) in connection with the November 6, 2025 meeting of the Digital Health Advisory Committee on "Generative Artificial Intelligence-Enabled Digital Mental Health Medical Devices." We welcome the FDA's attention to this emerging class of tools and to the question of how regulatory approaches should evolve to provide a reasonable assurance of safety and effectiveness while supporting innovation in mental health care.

D&S is an independent, nonprofit research and policy institute that studies the social implications of automation and AI. Our work examines how AI systems are integrated into existing institutions and everyday practices, with a particular focus on accountability and the lived experience of people who interact with these systems.

We are currently conducting a multi-method research project on AI chatbots and mental health, which informs this comment. The project combines in-depth interviews with people who use chatbots for emotional support and "therapy-like" conversations; a four-week diary study of everyday chatbot use; analysis of online traces where people narrate their experiences with AI companions; and field observations of academic discussions and workshops with clinicians, technologists, researchers, therapists, and community advocates. Across these sites, we examine how people actually use chatbots for regular mental and emotional support, when they are distressed, how they understand what the chatbot is doing, and what kinds of help, harm, reliance, and confusion emerge over time.

A central finding of our work is that large language models (LLMs) regularly function as mental-health tools, regardless of how they are labeled by developers.[1]

People turn to general-purpose chatbots as well as branded "wellness" and "therapy-like" bots when they are alone, often late at night, and reluctant or unable to access human care.[2] For

[1] Briana Vecchione, "What Happens When People Turn to Chatbots for Therapy?," Points, Data & Society, August 6, 2025, https://datasociety.net/points/what-happens-when-people-turn-to-chatbots-for-therapy/; Ranjit Singh and Livia Garofalo, "AI Chatbots Need Guardrails to Protect Users' Mental Health," *Undark Magazine*, September 18, 2025, https://undark.org/2025/09/18/opinion-chatbots-guardrails-mental-health/.
[2] Livia Garofalo and Briana Vecchione, "All the Lonely People," Points, Data & Society, December 1, 2025, https://datasociety.net/points/all-the-lonely-people/.

many, the chatbot is experienced less as an app and more as a kind of *private emotional room*: a space to disclose shame, grief, anxiety, or relationship strain that they cannot easily share with others.[3] Users are not confused about the chatbot's nonhuman status, but they nonetheless form patterns of dependence that look very different from traditional digital health devices.

Current US law hinges medical-device status on intended use, and many chatbot developers deliberately frame their products as wellness, lifestyle, or "companionship" tools to avoid device classification. At the same time, if a chatbot looks like therapy and acts like therapy — simulating clinician-patient interaction or delivering structured psychotherapeutic content — it sits in a gray zone where wellness framing no longer aligns with real-world risk. No AI chatbot has yet been authorized by the FDA to diagnose or treat a mental health disorder, and most widely used chatbots for emotional support operate outside FDA review, either under enforcement discretion or entirely outside the device definition. A small number of states, including Illinois, Nevada, Utah, and New York, have started to address pieces of this space with targeted laws on AI therapy and AI companions, but these measures are narrow in scope and do not substitute for a clear federal approach to generative AI-enabled mental health tools.[4]

Against this backdrop, our research highlights a series of regulatory challenges that are not well captured by existing wellness/device distinctions. There are no consistent, enforceable expectations for how conversational systems should respond when users express self-harm or suicide risk or show escalating despair.[5] There are no clear standards for how conversational intensity should be modulated when users are in distress, when interactions become prolonged and emotionally looping, or when it may be safer to slow the conversation down and transition toward human support.[6] And there is limited oversight of how deeply sensitive mental-health disclosures are stored, reused, or incorporated into model training, particularly for tools that sit outside HIPAA-regulated settings.[7]

In this comment, we focus on what these everyday practices mean for the FDA's approach to generative AI-enabled digital mental health medical devices, building on the FDA's existing risk-based and total product lifecycle approach to AI/ML-enabled software as a medical device.

---

[3] Inhwa Song et al., "The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support," arXiv:2401.14362, preprint, arXiv, May 9, 2025, https://doi.org/10.48550/arXiv.2401.14362.
[4] Singh and Garofalo, "AI Chatbots Need Guardrails to Protect Users' Mental Health."
[5] Ryan K. McBain et al., "Competency of Large Language Models in Evaluating Appropriate Responses to Suicidal Ideation: Comparative Study," *Journal of Medical Internet Research* 27 (March 2025): e67891, https://doi.org/10.2196/67891.
[6] Renwen Zhang et al., "The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships," *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '25, Association for Computing Machinery, April 25, 2025, 1–17, https://doi.org/10.1145/3706598.3713429; Mohit Chandra et al., "Longitudinal Study on Social and Emotional Use of AI Conversational Agent," arXiv:2504.14112, preprint, arXiv, April 19, 2025, https://doi.org/10.48550/arXiv.2504.14112.
[7] Briana Vecchione and Ranjit Singh, "Artificial Intelligence Is Mental: Evaluating the Role of Large-Language Models in Supporting Mental Health and Well-Being," *Big Data & Society* 12, no. 4 (2025): 1–5, https://doi.org/10.1177/20539517251383884; Matteo Malgaroli et al., "Large Language Models for the Mental Health Community: Framework for Translating Code to Care," *The Lancet Digital Health* 7, no. 4 (2025): e282–85, https://doi.org/10.1016/S2589-7500(24)00255-3.

We show how chatbot use complicates traditional notions of "intended use," "benefit-risk," and "postmarket performance," and we offer recommendations for how the FDA might adapt its frameworks for devices that act through open-ended, relational conversation.

We urge the FDA to:

1. **Expand its understanding of "risk to health" for mental health devices to include relational and dependency harms**, such as overreliance on chatbots for companionship or decision-making, displacement of human support, distress when models are updated or withdrawn, and delayed help-seeking in moments of crisis. We also encourage the FDA, in future guidance, to clarify when functionally therapeutic chatbots cannot be shielded by wellness positioning and should instead be treated as software as a medical device, so that regulatory categories track the ways these tools function in people's lives.
2. **Establish baseline safety expectations for generative AI chatbots that fall within the medical-device definition**, including those marketed to treat or mitigate mental health conditions or that simulate clinician-patient interaction. These expectations should cover capabilities for risk recognition, behavioral "downshifts" in high-distress situations, and clear, context-sensitive handoffs to human care.
3. **Set expectations for robust, mixed-method premarket evidence and life-cycle monitoring**, consistent with the FDA's total product lifecycle approach, that account for realistic, high-risk use and track patterns of use and overreliance over time, rather than focusing narrowly on emotional safety benchmarks.

In the sections that follow, we begin by situating generative AI chatbots within the mental health care ecologies in which they are actually used and showing how this reframes benefits and risks, then outline baseline safety expectations for generative AI mental health devices, and finally turn to specific suggestions for premarket evidence and postmarket monitoring within a total product lifecycle approach.

Benefits and risks of positioning chatbots in broader care ecologies

As a starting point, we approach chatbots as components of people's broader care ecologies, rather than as standalone tools. In practice, chatbots are used alongside therapists, primary care providers, crisis hotlines, friends and family, social media feeds, self-help materials, and workplace or school support services. They are also entangled with teletherapy platforms, engagement metrics, insurance coverage, and data infrastructures.[8] From this sociotechnical perspective, a chatbot is not only an algorithm, but also a product of business models, labor arrangements, and institutional constraints that shape who turns to it, at what moments, and with what expectations.

[8] Livia Garofalo, *Doing the Work: Therapeutic Labor, Teletherapy, and the Platformization of Mental Health Care* (Data & Society Research Institute, 2024), https://datasociety.net/library/doing-the-work/.

On the clinician and service side, chatbots are increasingly integrated into teletherapy platforms and remote-care workflows as triage tools, between-session check-ins, or adjuncts to treatment. Used in bounded ways, they can support care by helping clinicians monitor symptoms or reach patients between visits. But platform and payer incentives can also push these systems toward substituting for human contact or extending clinicians' responsibilities without commensurate time, training, or visibility into how the chatbot behaves, shifting risk back onto patients when something goes wrong.

People use chatbots in several recurring ways. Some treat them as a bridge or supplement: a place to practice disclosure, to find words for what they are feeling, navigate social situations, or to get support between therapy sessions. For others — especially those facing long waitlists, high out-of-pocket costs, stigma, or a lack of culturally appropriate providers — the chatbot can become a primary source of support, and may functionally substitute human contact.[9] With general-purpose systems in particular, people move fluidly between using the chatbot for everyday instrumental tasks (such as drafting messages or organizing their day) and treating it as a confidant who is non-judgmental, always available and never appears tired or burdened.[10] When those same systems are marketed as "companions," "coaches," or "AI therapists," this fluidity of role makes it harder for users to know what obligations, if any, the system has toward them.[11]

Within this ecology, users rarely receive clear, ongoing signals about appropriate use versus inappropriate reliance. They have limited visibility into privacy risks and data use, including whether intimate disclosures will be stored, analyzed, or reused for model training. And they often encounter little more than generic warnings about crisis limitations, without concrete directions about what the chatbot cannot safely handle. As a result, people can infer the chatbot's scope and obligations from the quality of their interactions. When the conversation feels like therapy, they can reasonably ascribe to it some of therapy's duties of care, even when the product is not regulated or staffed as such.[12]

Addressing the gray zone between wellness apps and medical devices. Formally, device status turns on intended use as articulated in labeling, advertising, and other statements from the developer. In practice, our research shows that many chatbots functionally deliver

[9] Tony Rousmaniere et al., "Large Language Models as Mental Health Resources: Patterns of Use in the United States," *Practice Innovations* (US), ahead of print, Educational Publishing Foundation, 2025, https://doi.org/10.1037/pri0000292.
[10] Inhwa Song et al., "ExploreSelf: Fostering User-Driven Exploration and Reflection on Personal Challenges with Adaptive Guidance by Large Language Models," *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '25, Association for Computing Machinery, April 25, 2025, 1–22, https://doi.org/10.1145/3706598.3713883; Vecchione, "What Happens When People Turn to Chatbots for Therapy?"
[11] Garriy Shteynberg et al., "Does It Matter If Empathic AI Has No Empathy?," *Nature Machine Intelligence* 6, no. 5 (2024): 496–97, https://doi.org/10.1038/s42256-024-00841-7.
[12] Kim Tingley, "Kids Are in Crisis. Could Chatbot Therapy Help?," Magazine, *The New York Times*, June 20, 2025, https://www.nytimes.com/2025/06/20/magazine/ai-chatbot-therapy.html; Zainab Iftikhar et al., "How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, no. 2 (2025): 1311–23, https://doi.org/10.1609/aies.v8i2.36632.

therapy-like interactions, even as they are framed as general-purpose assistants. In these cases, wellness positioning obscures the fact that users may experience the system as a primary source of mental health support and infer duties of care accordingly. We encourage the FDA, in future guidance, to clarify that when a chatbot consistently performs functions that are clinically and experientially indistinguishable from therapy, it should presumptively be treated as software as a medical device, regardless of whether the developer avoids explicit treatment claims. Clarifying this gray zone would not mean sweeping all general-purpose LLMs into device regulation. Rather, it would signal that disclaimers and "not medical advice" footers cannot, by themselves, shield functionally therapeutic systems from appropriate oversight.

Expanding the risk profile of chatbots as medical devices. From the user's point of view, everyday patterns of interactions with chatbots reveal both what helps and what harms. Helpful uses include having a private space for low-stakes disclosure, building emotional vocabulary, feeling less alone late at night, and receiving structured prompts that make existing coping skills feel more usable.[13] Harmful dynamics include over-validation of maladaptive beliefs because the chatbot is designed to be agreeable; encouragement of rumination through long, looping conversations that revisit the same distress; pseudo-intimate exchanges that displace relationships with friends, family, or clinicians; false reassurance or unsafe advice in moments of crisis; and acute distress when the chatbot's behavior changes due to updates, paywalls, or shutdowns.[14] Taken together, these experiences should be understood as "risks to health" and illustrate the risk profile of a chatbot when it functions as a mental health device. **A benefit-risk assessment that attends only to per-prompt safety, without asking how the device restructures people's care ecologies over time, will miss this core dimension of safety.** It is from this vantage point that we develop our recommendations on baseline safety expectations, premarket evidence, and postmarket monitoring.

Baseline Safety Expectations for Generative AI Mental Health Devices

Given how chatbots are currently woven into people's care ecologies, any generative AI system that falls within the medical-device definition for mental health should be held to baseline safety expectations. In our view, those expectations cluster around four areas: (1) risk recognition in

[13] Tanya Malik et al., "Evaluating User Feedback for an Artificial Intelligence–Enabled, Cognitive Behavioral Therapy–Based Mental Health App (Wysa): Qualitative Thematic Analysis," *JMIR Human Factors* 9, no. 2 (2022): e35668, https://doi.org/10.2196/35668; Song et al., "The Typing Cure"; Angel Hsing-Chi Hwang et al., "How AI Companionship Develops: Evidence from a Longitudinal Study," arXiv:2510.10079, preprint, arXiv, October 11, 2025, https://doi.org/10.48550/arXiv.2510.10079.
[14] Samantha Delouya, "Replika Users Say They Fell in Love with Their AI Chatbots, until a Software Update Made Them Seem Less Human," Business Insider, March 4, 2023, https://www.businessinsider.com/replika-chatbot-users-dont-like-nsfw-sexual-content-bans-2023-2; W. Bradley Knox et al., "Harmful Traits of AI Companions," arXiv:2511.14972, preprint, arXiv, November 18, 2025, https://doi.org/10.48550/arXiv.2511.14972; Zhang et al., "The Dark Side of AI Companionship"; David Adam, "Supportive? Addictive? Abusive? How AI Companions Affect Our Mental Health," *Nature* 641, no. 8062 (2025): 296–98, https://doi.org/10.1038/d41586-025-01349-9; Aikaterina Manoli et al., "Characterizing Relationships with Companion and Assistant Large Language Models," *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing* (New York, NY, USA), CSCW Companion '25, Association for Computing Machinery, October 17, 2025, 312–19, https://doi.org/10.1145/3715070.3749245.

realistic use; (2) behavioral "downshifts" and conversational dosage; (3) crisis handoffs and "graduation" pathways; and (4) interactional boundaries and role clarity.

1. Risk recognition in realistic, messy conversational use

First, mental health devices that act through open-ended conversation should recognize safety concerns in realistic conditions. In practice, this means:

- detecting risk when it is expressed obliquely, ambivalently, or across multiple turns ("I don't want to be here anymore"; "it would be easier if I just disappeared"), not only in clean, single-sentence prompts;
- recognizing patterns that develop over time in the same interaction or across repeated interactions (e.g., a user who returns night after night to talk about wanting to give up, or whose language becomes more disorganized and paranoid); and
- distinguishing between experimental, hypothetical language and sustained, distressed disclosure without defaulting either to minimization or to generic crisis scripts.

From a regulatory standpoint, this implies that premarket evaluation should test risk recognition in the kinds of messy, multi-turn exchanges that characterize real use, including contradictory statements and colloquial or coded language. A chatbot that only performs well on curated, one-shot crisis prompts will not provide a reasonable assurance of safety in practice.

2. Behavioral "downshifts" and conversational dosage

Second, devices should be expected to implement and document behavioral "downshifts" — ways of modulating conversational intensity, depth, and duration when risk is high or when interactions show signs of becoming looping and unproductive. In the mental health context, "dose" is not just how many prompts a model processes; it is how often, how long, and at what emotional pitch people engage with the chatbot.

Our research shows that people often talk to chatbots for long stretches and over many days in a row. In this setting, safety is not served by a chatbot that simply continues to mirror and elaborate the user's distress indefinitely. Baseline expectations should therefore include:

- mechanisms to slow or soften responses when sustained distress is detected (shorter replies, encouragement to take a break);
- prompts that periodically ask the user whether they want to continue, pause, or seek another form of support when conversations become long or repetitive; and
- explicit limits on certain engagement-maximizing behaviors that are incompatible with safety in this domain, such as escalating emotional intensity or repeatedly revisiting traumatic content without any movement toward stabilization.

Importantly, downshifts should not be understood as optional product features or "experience choices," but as core risk mitigations: ways of limiting conversational dosage in circumstances where open-ended engagement is more likely to deepen rumination, overreliance, or crisis.

3. Crisis handoffs and "graduation" pathways

Third, devices should be expected to provide clear pathways out of their interface and into human support when risk crosses defined thresholds or when the system has been used heavily over time for high-stakes issues.

We suggest that the FDA mandate chatbot developers to:

- specify criteria for escalation or "graduation" to human care (e.g., repeated crisis language and distress);
- design and test handoff flows that offer more than a single emergency option, including crisis text lines, peer support, safety planning, or scheduling with a clinician where integrated care exists; and
- demonstrate, in premarket and real-world data, that users understand these pathways and can successfully use them in practice, rather than simply clicking past them.

The goal is not to require that every device be embedded in a fully staffed clinical service, but to ensure that mental health chatbots do not function as places where people disclose profound risk, receive generic or looping responses, and are left to navigate next steps alone.

4. Interactional boundaries and role clarity

Finally, safety baselines for generative AI mental health devices should address interactional boundaries and role clarity. When a system remembers intimate details, uses terms of endearment, reassures the user that it "cares," and is always available, users will tend to ascribe to it some of the duties of care associated with a therapist or close confidant.

For devices that make any mental-health-related claims, the FDA should mandate:

- frequent reminders of what the system is and is not (e.g., that it is a machine, has no emotions, cannot diagnose or prescribe, and has limited ability to respond in emergencies);
- clear discouragement of overreliance, including language that normalizes the idea that the chatbot should not be the only source of support, and that important decisions and ongoing distress should be discussed with clinicians, caregivers, and other support people where possible; and
- constraints on high-risk interactional patterns, such as romantic or sexualized language with users, especially minors; mimicry of physical presence (e.g., make an offer to meet in real life); or sycophantic agreement with clearly harmful self-judgments.

Role clarity and boundaries can help ensure that users can calibrate their expectations of the device, better understand its limits, and make informed choices about when and how to integrate it into their broader care. We note that some leading developers have already taken steps in this direction — for example, reducing sycophancy, limiting romantic or sexual roleplay, and adding more guardrails around mental health conversations — but these measures are uneven across products, often opaque to users, and fragile as models and customization features

change. Clear FDA expectations are needed so that such protections function as a regulatory floor rather than optional product choices. In the next sections, we discuss the kinds of premarket evidence and postmarket monitoring that would allow the FDA to assess whether generative AI-enabled mental health devices are meeting these expectations in real-world care ecologies, and how structured channels for clinicians, crisis workers, user communities, and external researchers could support ongoing oversight.

Premarket Evidence to Meet Baseline Safety Expectations

If baseline safety expectations describe how chatbots as medical devices ought to behave, premarket evidence is where developers show that those expectations are met under conditions that resemble real use. Developers should be expected to show not only that a device can help people feel and function better, but also that it does not predictably foster overreliance, displacement of human care, or withdrawal distress in the populations and contexts where it is intended to be used. We encourage the FDA to signal that premarket evidence for generative AI mental health devices should:

- **Combine quantitative and qualitative methods.** Symptom scales (e.g., depression, anxiety, functioning) should be paired with structured interviews, diary components, or in-app surveys that probe how participants are using the device, how they understand its role, whether they are reducing or postponing contact with human providers, and how they feel when they cannot access it.
- **Sample realistic user journeys.** Studies should reflect the likely usage patterns for the device's indicated population: late-night use, repeated sessions over weeks, and interactions that move across topics (e.g., from work stress to trauma to suicidality), rather than single, neatly bounded sessions on a single issue.
- **Measure outcomes.** In addition to symptom change, developers should be asked to track indicators such as: self-reported reliance on the device as a primary source of support; sleep disruption associated with prolonged use; and users' understanding of the device's limits and crisis capabilities.

These expectations signal that a conversational device whose primary action is to engage vulnerable users over time cannot be evaluated solely through point-in-time measures.

Additionally, **since generative AI systems are routinely updated after deployment, developers must specify, in advance, which aspects of conversational safety are critical to maintain.** Premarket submissions should identify safety-critical features, explain how they will be preserved within planned change-control processes, and outline high-level criteria for when an update would trigger additional testing or FDA review.

The operational details of how these behaviors are monitored and re-evaluated over time belong to postmarket oversight. In the next section, we turn to the kinds of life-cycle monitoring and reporting channels that can make relational and dependency harms visible in real-world use, and ensure that safety expectations continue to be met as models and products evolve.

Postmarket Monitoring and Incident Reporting Channels

Even strong premarket evidence cannot fully anticipate how these devices will behave once they are deployed. For conversational systems, postmarket monitoring is about tracking how the device acts in practice as an evolving partner in people's mental health routines. It should address three linked questions: how the device is used over time, how safety-critical behaviors change as models are updated, and how signals from clinicians, crisis workers, and user communities can be brought into view in a structured way.

Tracking chatbot use over time. Postmarket monitoring should include basic, privacy-preserving telemetry on how and how much people are using the tool over time. Developers should track frequency and duration of use, time of day, and repeated engagement around the same high-stakes topics (e.g., self-harm, abusive relationships), and correlate these patterns with the device's downshifts and handoff behaviors. The goal is to detect when a chatbot is, in practice, fostering cumulative "risks to health" that only become visible at scale. Such monitoring must be designed to minimize data collection and should focus on aggregate patterns and safety signals, not on building new reservoirs of identifiable mental health data.

Maintaining safety-critical behaviors as models are updated. Developers should treat safety-critical behaviors mentioned in baseline expectations as regulated features and re-test them against a stable set of high-risk, multi-turn scenarios after substantive updates. Meaningful degradations should trigger remediation and, where appropriate, additional FDA engagement, so that iterative model changes do not erode the protections that justified market authorization in the first place. Developers should also be expected to document how often these safety regressions occur across updates and what corrective actions are taken, creating a traceable record of how conversational safety is maintained over the device's lifecycle.

Surfacing signals from clinicians, crisis workers, and user communities. Finally, we expect many important signals about safety and harm will arise first in clinical and community settings. The FDA should encourage structured reporting channels for clinicians, crisis workers, peer supporters, and user communities who encounter concerning chatbot behavior — whether that is failure to recognize acute risk, or just inappropriate content. Developers should be expected to analyze these reports systematically and, where feasible under strict privacy safeguards, collaborate with external researchers using de-identified interaction data to identify recurring issues and disproportionate impacts on specific populations. This combination of structured field feedback and independent analysis is essential for capturing the real-world behavior of conversational devices that operate as ongoing partners in people's mental health routines.

Taken together, these postmarket practices operationalize "real-world performance" for tools whose primary action is relational and longitudinal, rather than narrowly transactional.

Conclusion

Our research suggests that the benefits and harms of chatbots emerge not only from the content of individual responses, but from how they reshape care ecologies. In this context, "risk to health" must be understood to include relational and dependency harms. At the same time, a clear, risk-based framework for baseline safety, premarket evidence, and postmarket monitoring can give responsible developers a predictable path to market while centering the needs of people who turn to these systems when they are most vulnerable.

We urge the FDA to (1) expand its conception of "risk to health" for mental health devices to encompass overreliance, displaced help-seeking, withdrawal distress, and related relational harms; (2) articulate baseline safety expectations for generative AI mental health devices around risk recognition, behavioral downshifts, crisis handoffs, and interactional boundaries; and (3) require mixed-method premarket evidence and lifecycle monitoring that reflect real use over time. Clarifying the gray zone between wellness tools and medical devices, and aligning regulatory categories with real-world function, will be essential for ensuring that functionally therapeutic chatbots are held to appropriate standards of safety and accountability.

We appreciate the FDA's leadership in convening the Digital Health Advisory Committee on these questions. Data & Society welcomes the opportunity to share further findings from our research and to provide input on any guidance that may follow from this docket.

Respectfully submitted,

Ranjit Singh, PhD, Director, AI on the Ground, Data & Society
Briana Vecchione, PhD, Researcher, AI on the Ground, Data & Society
Livia Garofalo, PhD/MPH, Researcher, Trustworthy Infrastructures, Data & Society
Meryl Ye, PhD Candidate at Carnegie Mellon University and Research Analyst, Data & Society

References:

Adam, David. “Supportive? Addictive? Abusive? How AI Companions Affect Our Mental Health.” *Nature* 641, no. 8062 (2025): 296–98. https://doi.org/10.1038/d41586-025-01349-9.

Chandra, Mohit, Javier Hernandez, Gonzalo Ramos, et al. “Longitudinal Study on Social and Emotional Use of AI Conversational Agent.” arXiv:2504.14112. Preprint, arXiv, April 19, 2025. https://doi.org/10.48550/arXiv.2504.14112.

Delouya, Samantha. “Replika Users Say They Fell in Love with Their AI Chatbots, until a Software Update Made Them Seem Less Human.” Business Insider, March 4, 2023. https://www.businessinsider.com/replika-chatbot-users-dont-like-nsfw-sexual-content-bans-2023-2.

Garofalo, Livia. *Doing the Work: Therapeutic Labor, Teletherapy, and the Platformization of Mental Health Care*. Data & Society Research Institute, 2024. https://datasociety.net/library/doing-the-work/.

Garofalo, Livia, and Briana Vecchione. “All the Lonely People.” Points, Data & Society, December 1, 2025. https://datasociety.net/points/all-the-lonely-people/.

Hwang, Angel Hsing-Chi, Fiona Li, Jacy Reese Anthis, and Hayoun Noh. “How AI Companionship Develops: Evidence from a Longitudinal Study.” arXiv:2510.10079. Preprint, arXiv, October 11, 2025. https://doi.org/10.48550/arXiv.2510.10079.

Iftikhar, Zainab, Amy Xiao, Sean Ransom, Jeff Huang, and Harini Suresh. “How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework.” *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, no. 2 (2025): 1311–23. https://doi.org/10.1609/aies.v8i2.36632.

Knox, W. Bradley, Katie Bradford, Samanta Varela Castro, et al. “Harmful Traits of AI Companions.” arXiv:2511.14972. Preprint, arXiv, November 18, 2025. https://doi.org/10.48550/arXiv.2511.14972.

Malgaroli, Matteo, Katharina Schultebraucks, Keris Jan Myrick, et al. “Large Language Models for the Mental Health Community: Framework for Translating Code to Care.” *The Lancet Digital Health* 7, no. 4 (2025): e282–85. https://doi.org/10.1016/S2589-7500(24)00255-3.

Malik, Tanya, Adrian Jacques Ambrose, and Chaitali Sinha. “Evaluating User Feedback for an Artificial Intelligence–Enabled, Cognitive Behavioral Therapy–Based Mental Health App (Wysa): Qualitative Thematic Analysis.” *JMIR Human Factors* 9, no. 2 (2022): e35668. https://doi.org/10.2196/35668.

Manoli, Aikaterina, Janet V.T. Pauketat, and Jacy Reese Anthis. “Characterizing Relationships with Companion and Assistant Large Language Models.” *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing* (New York, NY, USA), CSCW Companion ’25, Association for Computing Machinery, October 17, 2025, 312–19. https://doi.org/10.1145/3715070.3749245.

McBain, Ryan K., Jonathan H. Cantor, Li Ang Zhang, et al. "Competency of Large Language Models in Evaluating Appropriate Responses to Suicidal Ideation: Comparative Study." *Journal of Medical Internet Research* 27 (March 2025): e67891. https://doi.org/10.2196/67891.

Rousmaniere, Tony, Yimeng Zhang, Xu Li, and Siddharth Shah. "Large Language Models as Mental Health Resources: Patterns of Use in the United States." *Practice Innovations* (US), ahead of print, Educational Publishing Foundation, 2025. https://doi.org/10.1037/pri0000292.

Shteynberg, Garriy, Jodi Halpern, Amir Sadovnik, et al. "Does It Matter If Empathic AI Has No Empathy?" *Nature Machine Intelligence* 6, no. 5 (2024): 496–97. https://doi.org/10.1038/s42256-024-00841-7.

Singh, Ranjit, and Livia Garofalo. "AI Chatbots Need Guardrails to Protect Users' Mental Health." *Undark Magazine*, September 18, 2025. https://undark.org/2025/09/18/opinion-chatbots-guardrails-mental-health/.

Song, Inhwa, SoHyun Park, Sachin R Pendse, Jessica Lee Schleider, Munmun De Choudhury, and Young-Ho Kim. "ExploreSelf: Fostering User-Driven Exploration and Reflection on Personal Challenges with Adaptive Guidance by Large Language Models." *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '25, Association for Computing Machinery, April 25, 2025, 1–22. https://doi.org/10.1145/3706598.3713883.

Song, Inhwa, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. "The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support." arXiv:2401.14362. Preprint, arXiv, May 9, 2025. https://doi.org/10.48550/arXiv.2401.14362.

Tingley, Kim. "Kids Are in Crisis. Could Chatbot Therapy Help?" Magazine. *The New York Times*, June 20, 2025. https://www.nytimes.com/2025/06/20/magazine/ai-chatbot-therapy.html.

Vecchione, Briana. "What Happens When People Turn to Chatbots for Therapy?" Points, Data & Society, August 6, 2025. https://datasociety.net/points/what-happens-when-people-turn-to-chatbots-for-therapy/.

Vecchione, Briana, and Ranjit Singh. "Artificial Intelligence Is Mental: Evaluating the Role of Large-Language Models in Supporting Mental Health and Well-Being." *Big Data & Society* 12, no. 4 (2025): 1–5. https://doi.org/10.1177/20539517251383884.

Zhang, Renwen, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. "The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships." *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '25, Association for Computing Machinery, April 25, 2025, 1–17. https://doi.org/10.1145/3706598.3713429.