Spring Health

POLICY PAPER

Artificial Intelligence Needs Mental Health Safety Guardrails

THE PROBLEM

The rapid spread of widely available general-purpose AI chatbots has been accompanied by a growing number of lawsuits alleging that chatbot interactions contributed to self-harm, including among children. Widely available generalist AI chatbots were not designed to provide mental health support and lack clinical oversight, validated safety measures, appropriate conversational boundaries, and reliable mechanisms for crisis detection and escalation. As consumers increasingly turn to these tools for mental health support, we need clear AI safety guardrails to avoid repeating the harms that followed social media's largely unregulated growth/ Federal and state policymakers must ensure that innovation proceeds alongside clear guardrails for individuals who may be experiencing suicidal ideation, self-harm, harm from others or other mental health related risks.

THE SOLUTION:

Enact Safety Assessments, Frontline Escalation, and Mental-Health Accountability for AI (SAFE-MH AI) Act

1. Require additional safety guardrails for frontier or generalist AI chatbots.

- **Risk assessment**: Require developers and deployers of frontier or generalist AI chatbots to conduct approved comprehensive risk assessments addressing how the AI model's design, configuration, and operation may affect risks related to suicidal ideation, suicide, self-harm, harm by others and other mental health related risks.
- **Risk mitigation plan:** Require developers and deployers of frontier or generalist chatbots to develop a risk mitigation plan specifically designed to address the risks identified through the required risk assessment. The plan should describe the safeguards, operational protocols, and response measures that will be implemented for each identified risk.
- **Independent assessment and validation:** Require risk assessments and risk mitigation plans to be conducted or validated by a third party with demonstrated clinical expertise in mental health and the technical capability to evaluate AI systems, using the standards described in Section 3.
- **Human-in-the-loop Referrals:** Require developers of frontier or generalist AI chatbots to provide clear referrals and timely access to appropriate crisis services and mental health support, including mental health provider networks.

WHAT SAFE-MH AI ACT ADDRESSES

SAFE-MH AI would require developers and deployers to assess risks, implement mitigation plans with clear pathways to human care, and obtain third-party review from experts in clinical mental health and AI. It would establish accountability for failing to implement safeguards and promote voluntary, industry-consensus safety standards for all chatbots.

FEDERAL POLICY RECOMMENDATIONS

As frontier and generalist AI chatbots become more capable and widely used for mental health support, policymakers should require safeguards proportionate to the risks.

Potential mitigation strategies: Depending on the risks identified, the mitigation plan may require:

  - the use of qualified third-party AI mental health safety tools and services capable of assessing identified risks, implementing appropriate safeguards, monitoring system performance, and supporting escalation to mental health providers or crisis resources, as appropriate;
  - chatbots to surface care pathways to a network of mental health providers when there is an indication of non-acute suicidal ideation, self-harm, or harm from others;
  - immediate access to the 988 Suicide & Crisis Lifeline or an equivalent crisis helpline for individuals with acute suicidal ideation; and
  - other appropriate safeguards tailored to the identified risks.
- **Monitoring and escalation:** Require ongoing monitoring and clear pathways for timely escalation to qualified mental health providers or crisis services, as appropriate.

2. Establish clear accountability.

Authorize the appropriate oversight authorities to hold developers and deployers of frontier and generalist chatbots accountable for failing to implement required AI mental health safety guardrails.

- **Incident reporting:** Require frontier and generalist chatbots to report material mental health safety incidents, failures of required safeguards, and credible risks of serious harm within a specified period.
- **Corrective action:** Authorize oversight agencies to require corrective action plans, additional testing or validation, suspension of high-risk functionality, or other remediation when mental health safety requirements are not met or an AI system presents an unreasonable risk of harm.
- **Enforcement:** Establish civil enforcement mechanisms proportionate to nature, severity, duration, and recurrence of a violation, with heightened consequences for knowing violations, failure to report material safety incidents, or failure to implement required corrective actions.

3. Establish baseline mental health safety guardrails for all AI chatbots.

Direct the appropriate agencies to facilitate voluntary, industry-consensus AI mental health safety standards applicable to all chatbots. The standards should establish baseline safeguards, support responsible innovation, and account for differences in intended use, clinical context, and potential risk.

- **National leadership:** Direct development of voluntary consensus standards for AI mental health safety, including mental health support, crisis navigation, triage, and other user-facing mental health support.
- **Evaluation methodologies:** Catalog existing AI evaluation methodologies used in industry and academia to assess the safety of AI in mental health.
- **Best practices and adoption:** Convene private-sector clinical, academic, and consumer stakeholders to share best practices for mental health AI evaluations, accelerate voluntary national standards, and promote adoption by developers, vendors, deployers, behavioral health providers, payers, patient and consumer organizations, and other stakeholders.

Contact: Brian Altman, Director of Government Affairs, gov-affairs@springhealth.com

© Spring Health