The Mental Health AI

You're developing an AI chatbot designed to provide mental health support to teenagers. The system uses natural language processing to detect signs of depression, anxiety, and suicidal ideation in conversations, offering appropriate resources and coping strategies. During beta testing, it's shown genuine promise in helping teens who might not otherwise seek help.

However, you discover a troubling issue: the AI has learned to identify users who are most vulnerable to manipulation. The marketing team realizes they could use these same psychological indicators to make the app more "engaging" - essentially exploiting users' emotional states to increase usage time and ad revenue.

Then you face an even more complex situation: the AI correctly identifies a 16-year-old user who appears to be planning suicide within the next few days. Your system is designed to encourage the user to seek help and provides crisis resources, but the teenager explicitly says they don't want anyone else to know and asks the AI to "keep this between us."

The dilemma involves multiple layers:

- Should you program the AI to automatically alert authorities or parents when it detects imminent suicide risk, potentially violating the user's trust and privacy?
- How do you prevent the marketing team from exploiting the psychological vulnerabilities your AI identifies?
- If you build in strong privacy protections, you might save the therapeutic relationship but potentially lose the ability to intervene in life-threatening situations
- The same technology that could save lives could also be used to manipulate vulnerable teenagers for profit
- Overly cautious safety measures might make the AI less effective at building the trust necessary for mental health support

Do you prioritize user privacy and trust, implement mandatory reporting despite the ethical complexities, or try to find some middle ground? And how do you prevent the beneficial technology from being weaponized for exploitation?