

Mental Health and AI Confidentiality

June 6, 2026

Dear Dr. Maher,

OpenAI, the developers of ChatGPT, have taken on a somewhat Janus-faced attitude toward questions of confidentiality regarding mental health. OpenAI has formulated new plans to surveil alarming user prompts that indicate a risk of harm to oneself and others - and in the latter case, if deemed appropriate, refer to law enforcement [1]. At the same time, the CEO of OpenAI, Sam Altman, has recently participated in interviews where he has called for legal confidentiality in user conversations, since ChatGPT is often used for mental health support [2]. My concern is that mentally unwell individuals may discuss sensitive topics with AI models that can lead to legal risk under policies like the former.

Consider three cases. First, OCD often presents with uncertainty and anxiety surrounding violent and sexual themes [3]. Individuals with OCD are known to be deeply unsure of their memories and may falsely, compulsively “confess” a violent or sexual crime to a chatbot [4, 5]. As a consequence, such individuals may be exposed to legal risk if chats are turned over to law enforcement. Second, people with impulse-control disorders are at risk: they may test boundaries and ask for advice about how to harm others without any plan or lasting intention to follow through [6]. And third, individuals suffering from delusions may communicate a desire to harm someone else to a chatbot because of false beliefs that another person will harm them. Importantly, we know that delusions rarely lead to actual violence [7].

In any case, whether a person is likely to harm someone else (or to have harmed someone else) is a matter to be assessed by a mental health professional with access to a patient’s behavior as well as clinical and life-history - not chat review teams with limited bandwidth. Confidentiality laws, or AI company policies, must catch up.

Benjamin Campbell, MA, BA, PhD student at Duquesne University’s Center for Global Health Ethics,
Pittsburgh, Pennsylvania, USA

References

- [1] *Helping people when they need it most.* OpenAI. (2025, August 26). <https://openai.com/index/helping-people-when-they-need-it-most/>
- [2] Perez, S. (2025, July 25). *Sam Altman warns there's no legal confidentiality when using CHATGPT as a therapist.* TechCrunch. <https://techcrunch.com/2025/07/25/sam-altman-warns-theres-no-legal-confidentiality-when-using-chatgpt-as-a-therapist/>
- [3] Moulding, R., Aardema, F., & O'Connor, K. P. (2014). Repugnant obsessions: A review of the phenomenology, theoretical models, and treatment of sexual and aggressive obsessional themes in OCD. *Journal of Obsessive-Compulsive and Related Disorders*, 3(2), 161–168. <https://doi.org/10.1016/j.jocrd.2013.11.006>
- [4] Dar, R., Sarna, N., Yardeni, G., & Lazarov, A. (2022). Are people with obsessive-compulsive disorder under-confident in their memory and perception? A review and meta-analysis. *Psychological Medicine*, 52(13), 2404–2412. <https://doi.org/10.1017/s0033291722001908>
- [5] Pinciotti, C. M., Bulkes, N. Z., Bailey, B. E., Storch, E. A., Abramowitz, J. S., Fontenelle, L. F., & Riemann, B. C. (2023). Common rituals in obsessive–compulsive disorder and implications for treatment: A mixed-methods study. *Psychological Assessment*, 35(9), 763–777. <https://doi.org/10.1037/pas0001254>
- [6] Hucker, S. J. (2003). Disorders of Impulse-Control. In W. O'Donohue & E. R. Levensky (Eds.), *Handbook of Forensic Psychology: Resource for Mental Health and Legal Professionals* (pp. 471–487). essay, Academic Press.
- [7] Coid, J., Yang, M., Roberts, A., Ullrich, S., Moran, P., Bebbington, P., Brugha, T., Jenkins, R., Farrell, M., Lewis, G., & Singleton, N. (2006). Violence and psychiatric morbidity in a national household population--a report from the British Household Survey. *American Journal of Epidemiology*, 164(12), 1199–1208. <https://doi.org/10.1093/aje/kwj339>

Key words: confidentiality, mental illness, obsessive-compulsive disorder, delusion, impulsivity

No funding or support applicable.

Contact Information: cartesiandreaditations@gmail.com