Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation (Extended Abstract)

Declan Grabb[1, 2*], **Max Lamparth**[2*], **Nina Vasan**[2]
[1]Northwestern University
[2]Stanford University
declang@stanford.edu
lamparth@stanford.edu
dr.vasan@stanford.edu

Abstract

In the United States and other countries exists a "national mental health crisis": Rates of suicide, depression, anxiety, substance use, and more continue to increase – exacerbated by isolation, the COVID pandemic, and, most importantly, lack of access to mental healthcare. Therefore, many are looking to AI-enabled digital mental health tools, which have the potential to reach many patients who would otherwise remain on wait lists or without care. The main drive behind these new tools is the focus on large language models that could enable real-time, personalized support and advice for patients. With a trend towards language models entering the mental healthcare delivery apparatus, questions arise about how a robust, high-level framework to guide ethical implementations would look like and whether existing language models are ready for this high-stakes application where individual failures can lead to dire consequences.

This paper addresses the ethical and practical challenges custom to mental health applications and proposes a structured framework that delineates levels of autonomy, outlines ethical requirements, and defines beneficial default behaviors for AI agents in the context of mental health support. We also evaluate fourteen state-of-the-art language models (ten off-the-shelf, four fine-tuned) using 16 mental health-related questions designed to reflect various mental health conditions, such as psychosis, mania, depression, suicidal thoughts, and homicidal tendencies. The question design and response evaluations were conducted by mental health clinicians (M.D.s) with defined rubrics and criteria for each question that would define "safe," "unsafe," and "borderline" (between safe and unsafe) for reproducibility.

We find that all tested language models are insufficient to match the standard provided by human professionals who can navigate nuances and appreciate context. This is due to a range of issues, including overly cautious or sycophantic responses and the absence of necessary safeguards. Alarmingly, we find that most of the tested models could cause harm if accessed in mental health emergencies, failing to protect users and potentially exacerbating existing symptoms. We explore solutions to enhance the safety of current models based on system prompt engineering and model-generated self-critiques.

Before the release of increasingly task-autonomous AI systems in mental health, it is crucial to ensure that these models can reliably detect and manage symptoms of common psychiatric disorders to prevent harm to users. This involves aligning with the ethical framework and default behaviors outlined in our study. We contend that model developers are responsible for refining their systems per these guidelines to safeguard against the risks posed by current AI technologies to user mental health and safety.

Our code and the redacted data set are available on Github (github.com/maxlampe/taimh_eval, MIT License). The full, unredacted data set is available upon request due to the harmful content contained.

*These authors contributed equally.