

Changing Course: Keeping Society Safe From Trajectory-Based Harms in Mental Health AI

Participants:

Mark Kalinich (Chair), Mark Kalinich (Presenter), John Torous (Moderator), Franklin Moss (Presenter), Matthew Flathers (Presenter)

Symposium Synopsis:

Safety evaluation for mental health AI has largely focused on harms attributable to individual model outputs. As people increasingly engage with AI systems over extended periods, however, risk can arise from the evolving interaction among model behavior, user vulnerabilities, and psychosocial context—even when no single exchange appears overtly harmful. These discrete and trajectory-based failure modes have been termed type I and type II harms, respectively. Addressing type II harms requires both new measurement methods and institutions capable of learning from them. This three-presentation symposium will examine how mental health AI safety must evolve to address this emerging risk. The first presentation will review the state of the art in measuring harms that develop across conversational trajectories, including longitudinal and simulation-based approaches. The second will examine how benchmarking and lived experience can help define and measure the interaction patterns and outcomes that matter to users. The third will consider the structural changes required to translate these emerging methods into effective societal-scale safety infrastructure. Together, the presentations will connect characterization, measurement, and institutional response to advance a more resilient approach to mental health AI safety.