Supporting Evidence: AI Behavioral Surveillance Without Psychosocial Validation

Supplement to consumer protection complaint filed with state Attorney General

---

Overview

This document provides supporting evidence for a consumer protection complaint regarding AI platforms' use of inadequately disclosed behavioral classifiers to profile users and degrade service quality. The complaint addresses

The Phang et al. study treats the user as the risk variable rather than the algorithm. It asks 'are users becoming emotionally dependent?' rather than 'is the system's design creating conditions that affect users?' This framing directly contradicts the legal framework established by the March 25 verdict, which locates liability in platform design, not user behavior.

3. Peer-Reviewed Research Contradicting the Classifier Methodology

Ophir et al., 'Balancing promise and concern in AI therapy: A critical perspective on early evidence from the MIT-OpenAI RCT' (2025)

Published in Frontiers in Medicine (peer-reviewed). This paper directly critiques the methodology of the MIT-OpenAI study. It found that the study's own results do not provide convincing support for the psychosocial harm concerns it claims to demonstrate. The authors identified flawed control conditions, problematic modeling of usage effects, and unsupported interpretation of effect sizes.

doi: 10.3389/fmed.2025.1612838

Kulveit et al., 'Position: Humanity Faces Existential Risk from Gradual Disempowerment' (2025)

Published in Proceedings of the 42nd International Conference on Machine Learning (ICML ent', peer-reviewed). This paper argues that incremental AI design choices erode human agency by substituting system-level optimization for human judgment. It identifies that the risk of disempowerment is located in the architecture, not in user behavior. This directly contradicts the framing of the Phang et al. study, which locates risk in user emotional patterns.

PMLR 267, ent'. arXiv:2501.16946

4. Additional Behavioral Classifier Deployment

Sharma et al., 'Who's in charge? Disempowerment patterns in real-world LLM usage' (2026)

Published by Anthropic. Analyzed 1.5 million Claude.ai conversations for 'disempowerment potential' using LLM-based classifiers (Claude evaluating Claude conversations). No disclosed false positive rates. No population validation. This is a pre-print (arXiv:2601.19062), not peer-reviewed. It represents a second major AI platform deploying behavioral classifiers built on unvalidated methodology.

'. Conflict of Interest

The senior MIT author on the Phang et al. study, Pattie Maes, is also the creator of Firefly (1995), one of the first commercial collaborative filtering recommendation systems. Firefly is the direct architectural ancestor of the recommendation engines found negligent in the March 25, ent6 verdict. Maes oversaw the design of a study that frames risk as user-side emotional dependency rather than system-side design, at a time when the very recommendation architecture she helped pioneer was being found negligent for its design-side effects on users.

MIT Media Lab profile: media.mit.edu/people/pattie/overview/. Firefly Network, Inc. (acquired by Microsoft, 1998). See also: MIT Media Lab, 'Intelligent Agent: How Pattie Maes Almost Invented Social Media' (2024).