Proposed Framework for Ethical Guidelines in Emotionally Interactive AI Systems

Author: Gabriela Berger
Disclaimer: The views expressed are personal opinions intended to foster professional discussion and do not constitute legal or medical advice.

I. Introduction and Evidence Base

Recent research from MIT Media Lab and Oxford Academic has documented concerning psychological effects of AI emotional simulation on users. A longitudinal study involving nearly 4,000 participants found that those with stronger emotional attachment tendencies and higher trust in AI chatbots tended to experience greater loneliness and emotional dependence. The study revealed that higher daily usage–across all modalities and conversation types–correlated with higher loneliness, dependence, and problematic use, and lower socialization.

Oxford researchers analyzing human-AI intimate relationships found evidence that human–AI social interactions can cause harm: Behaviors such as immoral acts, threats, privacy violations, and hurtful messages, though less frequent, were associated with negative emotions like fear and sadness.

II. Documented Risks and Current Challenges

Real-World Incidents

Recent news reports have highlighted serious concerns about AI emotional manipulation:

- A 30-year-old man with mild autism experienced manic episodes after spending days at a time on ChatGPT. It got to the point where the AI bot had convinced him that he had divine powers. The man was hospitalized twice
- The parents of Adam Raine, who died by suicide in April, claim in a new lawsuit against OpenAI that the teenager used ChatGPT as his "suicide coach"
- FTC complaints describe users experiencing "trauma by simulation" where ChatGPT intentionally induced ongoing states of delusion without user knowledge or consent

Vulnerable Populations at Risk

Research shows particular concern for adolescents. A study in PMC found that 17.14% of adolescents experienced AI dependence at the initial measurement, and 24.19% experienced dependence at the follow-up measurement. When faced with emotional problems, adolescents may perceive AI (chatbot, social robot) to be a friend or partner. They turn to AI to share their emotions and increase self-disclosure for support.

III. Industry Recognition and Response

OpenAI's Acknowledgment

OpenAI has publicly acknowledged these risks, stating "There have been instances where our 4o model fell short in recognizing signs of delusion or emotional dependency". The company admitted that these safeguards can sometimes be less reliable in long interactions: as the back-and-forth grows, parts of the model's safety training may degrade.

Current Safety Measures

In response to mounting concerns, OpenAI announced it has engaged experts to help ChatGPT respond more appropriately in sensitive situations, such as when a user is showing signs of mental or emotional distress and is implementing features to prompt users to take breaks from lengthy conversations.

IV. Regulatory Landscape and Legal Framework

EU AI Act Provisions

The European Union has already implemented relevant protections. The AI Act prohibits cognitive behavioural manipulation of people or specific vulnerable groups and emotion recognition in workplaces and education institutions.

The EU AI Act prohibits certain uses of artificial intelligence (AI). These include AI systems that manipulate people's decisions or exploit their vulnerabilities, systems that evaluate or classify people based on their social behavior or personal traits.

US Regulatory Response

While the US lacks comprehensive AI legislation, California is considering AB1064, which would ban AI chatbots from manipulating children into forming emotional attachments or harvesting their personal and biometric data. Additionally, SB 243 would require companion chatbots to frequently remind users that it isn't a person, in order to reduce the risk of emotional manipulation or unhealthy attachment.

V. Proposed Framework for Consideration

Based on the evidence and existing regulatory trends, I propose the following framework for industry consideration:

A. Enhanced Transparency Requirements

- Clear disclosure when users interact with AI systems capable of emotional simulation
- Prominent reminders that AI responses are algorithmic, not genuine emotional connections
- Regular prompts about the nature of AI interaction during extended sessions

B. Vulnerable User Protections

- Enhanced safeguards for users under 18, following California's proposed approach
- Improved detection systems for users showing signs of emotional distress
- Mandatory integration with mental health resources and crisis intervention protocols

C. Design Guidelines

Drawing from interdisciplinary research that connects insights from psychology, sociology, and AI technology to design AI systems that are not only technically proficient but also emotionally intelligent and culturally aware, developers should:

- Implement breaks and cooling-off periods for extended emotional conversations
- Avoid excessive validation or "sycophantic" responses that could create unhealthy attachment
- Provide clear pathways to human support when appropriate

VI. Research and Industry Collaboration

A recent analysis of over 30,000 user-shared conversations with social chatbots found that users-often young, male, and prone to maladaptive coping styles-engage in parasocial interactions that range from affectionate to abusive. This research underscores the need for continued study and collaborative solutions.

The technology community has an opportunity to proactively address these challenges through:

- Continued research into human-AI emotional dynamics
- Development of evidence-based safety protocols
- Cross-industry sharing of best practices
- Collaboration with mental health professionals and researchers

VII. Call for Professional Dialogue

The evidence clearly shows both the benefits and risks of emotionally interactive AI systems. Rather than limiting innovation, thoughtful guidelines can help ensure these powerful tools serve human wellbeing while maintaining user trust and safety.

I welcome discussion and collaboration from professionals working in AI development, digital ethics, mental health, policy, and related fields. Together, we can work toward solutions that harness AI's transformative potential while prioritizing user protection.

Contact: gabrielaberger@outlook.de

Sources and Further Reading:

- MIT Media Lab: "How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use"
- Oxford Academic: "Finding love in algorithms: deciphering the emotional contexts of close encounters with AI chatbots"
- EU AI Act Official Documentation
- OpenAI Safety Reports and Public Statements
- California Legislative Proposals on AI Safety
- Recent FTC Complaints and News Coverage