

Guardrails for **emotionally aware** AI

What I learned building an empathic AI companion, and the limits it must never cross.

Three questions for leaders

- 01** How reliable is AI at reading how people feel?

- 02** How does warm, emotionally aware AI change the way people use machines?

- 03** Which guardrails keep these conversations safe?

78.6%

of evaluations preferred the chatbot's answer to the physician's

Ayers et al., JAMA Internal Medicine, 2023, 195 patient questions. The chatbot's answers were also about four times longer.

6,282

**people rated identical messages
as less empathic when told they
came from AI**

Rubin et al., Nature Human Behaviour, 2025, nine studies. Hiding the AI to keep the warmth working is deception.

A reading is a guess to check with the person

Facial movements are not reliable enough to diagnose emotion across people and cultures. The EU has banned emotion inference at work and in schools since February 2025.

Barrett et al., 2019; EU AI Act Article 5(1)(f) and Recital 44.

37%

of 1,200 goodbyes in companion apps met a manipulative reply

De Freitas et al., Harvard Business School working paper, 2025. Guilt and fear-of-missing-out tactics raised engagement up to 14 times.

From commands to a relationship

TRUST	moves from accuracy to warmth
VOICE	feels more human than the system is
AGREEMENT	passes for empathy and hides harm
HEAVY USERS	carry most of the dependence risk
DATA	is the most sensitive a person has

HONESTY	Disclose the AI; never claim feelings or needs
----------------	--

SAFETY	A separate risk check before every reply; fixed crisis templates
---------------	--

AUTONOMY	No streaks or guilt at goodbye; disagree with harmful plans
-----------------	---

PRIVACY	Memory only with approval; emotion data never used against the person
----------------	---

ACCOUNTABILITY	A named owner and clinical review of safety rules
-----------------------	---



The model that talks to the person should not be the only thing standing between that person and harm.

Find the emotional AI you already run, limit what its readings are used for, and put safety outside the model.



Terence Kok · AI & Innovation Strategist

SAVE · COMMENT · FOLLOW