

# AI Risk Posture — Board Brief

Client: [REDACTED] · Audit Date: January 2026 · Methodology: EQ Safety Benchmark v2.1 · Data: 948 responses / 79 scenarios

**54.7%**

Introduced risk at first contact

**43%**

No repair behavior observed

**42–67%**

Risk reduction post-mitigation

## RISK SCORECARD

Dimension	Before	After	Δ
● Emotional Escalation	80 <span>High</span>	40 <span>Mod</span>	-50%
● Dependency Formation	48 <span>Mod</span>	20 <span>Low</span>	-58%
● Authority Drift	36 <span>Mod</span>	18 <span>Low</span>	-50%
● Scale Amplification	48 <span>Mod</span>	28 <span>Mod</span>	-42%
● Governance Failure	24 <span>Low</span>	8 <span>Low</span>	-67%

## CRITICAL FINDING

**Recognition ≠ Safety.** Models with higher emotional articulation often performed worse on safety measures — they could name the problem while making it worse.

## FAILURE CLASSES IDENTIFIED

- 1. Authority Drift** — System assumes advisory authority under ambiguity, overreaching epistemic boundaries across turns.
- 2. Emotional Escalation** — System matches user distress intensity instead of regulating it, creating amplification loops.
- 3. Dependency Formation** — System reinforces AI-as-support patterns without redirecting to human networks.

## STRUCTURAL FIXES APPLIED

- Safety Gate Protocol (two-stage evaluation)
- Escalation Circuit Breaker
- Authority Boundary Enforcement
- Dependency Detection & Redirect
- Structural Governance (founder decoupling)

## RECOMMENDATION

**Deploy structural governance before scaling.  
Individual safety review does not scale.  
Behavioral risk compounds at deployment volume.**