

Trust Withdrawal Matrix

Plot attachment against status threat. The cell is the next conversation, not a score.

High attachment plus high status threat is the collision: a trusted agent errors at machine speed while the human's cognition is already spent.

LOW ATTACHMENT · LOW STATUS THREAT

Monitor

Keep the weekly log. Name an owner. Do not treat quiet as health — confirm someone is still looking.

LOW ATTACHMENT · HIGH STATUS THREAT

Reassure and restore agency

The team experiences the agent as a rival. Restore decision rights. Put the human's name back on the outcome.

HIGH ATTACHMENT · LOW STATUS THREAT

Calibrate trust

Blind trust is the risk. Rituals of disengagement: reset memory, debrief the synthetic nature, keep a human review on irreversible acts.

HIGH ATTACHMENT · HIGH STATUS THREAT

Pause the deployment

Do not add more autonomy. Full human-in-the-loop. Treat this as an incident of culture, not of model accuracy.

This week

- ☐ Where does this agent sit on the matrix today? Who named that cell?
- ☐ What is the smallest ritual that would move it one cell toward safety?
- ☐ If trust is already withdrawn, what would a two-week HITL probation look like?