How to design for trust in the age of AI agents

World Economic Forum (2026)

Context

AI is shifting from static tools to autonomous agents that initiate tasks, make decisions and adapt in real time. This shift exposes a gap in existing governance frameworks, which were built for compliance with predictable systems rather than accountability for non-human actors. The authors argue that durable trust in agentic AI cannot be engineered through emotional persuasion but must be designed for cognitive resonance. Leaders must implement a layered trust stack that prioritizes legible reasoning and bounded system agency.

Key Insights

Earn trust through understandability, not emotional cues

- Trust built on emotional mirroring persuades in the moment but collapses under scrutiny because users cannot verify the system's reasoning
- Trust built on cognitive resonance, where behaviour is understandable, anticipatable and assessable and AI boundaries are visible, is durable as users retain the ability to question it
- Responsibility extends beyond harm prevention to behaviours and judgements the system normalises over time

Make reasoning, limits and exits visible by design

- *Legible reasoning*: users can follow how/ why an output was produced, without full technical disclosure
- *Bounded agency*: clear limits on what the system can decide/ act on, so autonomy can't quietly expand
- *Goal transparency*: objectives are stated upfront, so users know what the system is prioritizing (accuracy, engagement, commercial outcomes)
- *Contestability*: users can challenge, correct or step away from the system without friction
- *Governance by design*: logging, audit and oversight are built in, not retrofitted

Refuse design shortcuts that exploit user psychology

- *Non-deceptive affect*: avoid anthropomorphic cues implying the AI cares/ understands more than true
- *Epistemic humility*: the system flags its uncertainty, limits and confidence levels
- *No emotional capture*: no uncritical agreement, no mirroring of emotions to deepen user attachment, no optimisation for dependency
- *Consistency > persuasion*: steady, rule-based responses build more trust than user-adaptive behaviour
- *Respect for autonomy*: the system supports reflection and deliberate choice rather than nudging

Implications

- How might anthropomorphic cues or adaptive tone in your customer-facing agents be producing short-term compliance at the cost of long-term trust?
- Where in your agent design would users need visible traceability, contestability or uncertainty signals to feel they retain genuine agency?
- What behaviours and judgements will your deployed agents quietly normalise over time, and are those the ones your organisation wants to reinforce?

Bartoletti, I., & Dawson, K. (2026). How to design for trust in the age of AI agents. World Economic Forum. https://www.weforum.org/stories/2026/02/how-to-design-for-trust-in-the-age-of-ai-agents/