

AGENT DESIGN

Eight ways agent products lose trust

Failure patterns that show up across shipped products and the 2026 research.

Swipe →

Confirm only at the end

What it looks like

The agent works for twenty minutes, then asks whether it got it right.

What it costs

One error in step three means redoing all twelve.

Instead

Place checkpoints where they pay. Early ones earn the most.

In a study of 48 people at CHI 2026, **81% preferred an agent that checks in partway through**, and task time fell **13.5%**.

Transparency theatre

What it looks like

A wall of tool calls, labelled "reasoning".

What it costs

Overload reads as opacity.
Nobody checks anything.

Instead

A summary by default. Let people descend when they want to.

Microsoft Research names the balance between transparency and cognitive overload as an **overarching challenge** in human-agent communication.

One autonomy switch for everything

What it looks like

A single on or off in settings, for every task the agent does.

What it costs

One bad experience turns the whole feature off rather than dialling it back.

Instead

One dial per task type. Trust is not a single number.

People trust an agent to book a meeting long before they trust it to **send an email in their name.**

Silent irreversible action

What it looks like

It already happened. No preview before, no undo after.

What it costs

Trust rarely comes back from this one.

Instead

Preview it. Offer undo. Say plainly when undo expires.

A useful operating rule: if people reverse an action type more than **5% of the time**, turn automation off for that task.

Notification flood

What it looks like

Every background agent interrupts on its own schedule.

What it costs

People stop reading. Including the one that mattered.

Instead

One attention budget shared across every agent you run.

Too many interruptions defeat the point of delegating. Too few defeat oversight. **The ceiling is a design decision**, not a side effect.

Chat as the container for everything

What it looks like

Parallel, background and overnight work, all in one transcript.

What it costs

You cannot answer "what is happening right now" in ten seconds.

Instead

A queue of decisions for what needs you. One board for what is running.

LangChain open-sourced an **agent inbox** for exactly this. Google Antigravity ships a manager view. Both replace the transcript.

Anonymous fleets

What it looks like

Agents with no name, no owner and no stated remit.

What it costs

Nobody can say who authorised what, or which agents are still needed.

Instead

Every agent gets an owner, a scope, a spend cap and a review date.

Gartner named **agent sprawl** in April 2026: old agents accumulate, dependencies go unmapped, and nobody can say which version is running.

Confident guessing

What it looks like

Ambiguity resolved quietly into something plausible.

What it costs

The error arrives finished, formatted and easy to believe.

Instead

Three ways out: tell me, ask me one thing, or show me before you act.

An agent that says "**I am not sure how to proceed**" costs you thirty seconds. One that guesses can cost you the relationship.

If you change one thing

Make the agent hand you receipts

Why this one

Plans, screenshots, source lists, a note on what it could not verify. Reviewable in a minute, instead of a log nobody reads.

Google calls these **Artifacts**. It is the strongest pattern of the last year, and it has barely left developer tools.

Go deeper

The whole pattern library

On the site

Seven families, forty-one patterns.
What each one prevents, and the
product or the paper it came from.
Free, and no sign-up.

thinkbigleaders.com/resources/agent-patterns

Or talk it through

Curious, and nowhere near a decision?
Good. Have a use case in mind? Also
good. Already building one, or want
agents inside your own product? Bring
what you have.



Thirty minutes with Think Big Leaders.

Scan the code to pick a time.