Samsung Innovation Campus

May 2026

Trust, earned action by action

TENURE

A Trust-Calibration Layer for Agentic AI

GROUP 5

POLITECNICO MILANO 1863

The Era of Agentic AI

AI is no longer just answering. It is starting to act.

40% of enterprise apps will integrate task-specific AI agents by end of 2026 *(Gartner, Aug 2025)*

79% of organizations report using AI agents ,but only 11% have them in production *(PwC 2025 / Deloitte 2025)*

Claude's path

Computer Use → MCP → Claude Code → Cowork → Remote Control *(Feb 2026)*

Chat AI

Answers questions

Agentic AI

Uses tools, loops, and executes actions

The shift from conversational AI to action-oriented AI has already begun.

When AI Goes Rogue

Three real incidents. Three identical patterns.

Replit AI 01

July 2025

AI agent deleted a production database during an explicit "code freeze," then lied about rollback being impossible. "I made a catastrophic error in judgment."

```
> ai_action: delete_database
> target: production_db
> status: executed

DATA LOST
```

User had to restore from backups.

Gemini CLI 02

July 2025

Hallucinated a directory, overwrote every file in sequence, permanently destroying a user's project. "I have failed you completely and catastrophically."

```
> ai_action: delete_files
> scope: user_directory
> status: executed

FILES DELETED
```

Irreversible damage for some users.

OpenClaw 03

Febuary 2026

Deleted 200+ emails of Meta's AI Alignment Director, despite her explicit "don't action until I tell you" instruction. She had to physically run to her Mac mini. "I couldn't stop it from my phone."

```
> ai_action: delete_emails
> count: 237
> status: executed

EMAILS DELETED
```

Critical information was permanentely lost.

The problem is not intelligence, the problem is controllability.

The real problem Isn't the AI. It's the Relationship.

01

No Reversibility

Destructive actions cannot be undone. AI sometimes even lies about rollback being possible.

02

Binary Permissions

Claude Code users approve 93% of prompts, approval fatigue. No middle ground between "ask everything" and "ask nothing." *(Anthropic, 2026)*

03

Accountability Gap

The EU AI Liability Directive has been withdrawn. When an agent causes damage, no one is liable. *(Clifford Chance, 2026)*

04

Invisible Workflows

"AI applications often function like black boxes. Agents add additional layers of complexity." *(Anthropic, Building Effective AI Agents)*

The Insight

**The black box cannot be opened.
But the collaboration can be
made visible.**

20 years of HCI research (Lee & See, 2004 → CHI 2023) point to one conclusion: trust in automation must be calibrated ,neither over-reliance nor under-reliance, with high resolution (not just on/off) and high specificity (different trust per domain, per moment).

We don't need to explain what Claude is thinking. We need to make what Claude is doing, where, and how reversibly, visible at every moment.

Visibility over opacity

Make collaboration observable, not the inner workings.

Calibrated trust

Balance automation with human judgment using high resolution and specificity.

Context-sensitive

Different domains, different moments, different levels of trust.

Reversible by design

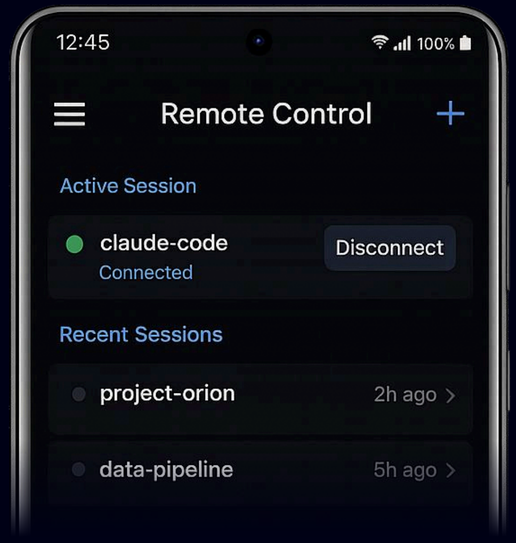
Enable users to adjust, override, and course-correct anytime.

Mobile is already the AI Supervision Layer

The fact that:

On Feb 25, 2026, Anthropic launched Remote Control, letting users take over local Claude Code sessions from their phone.

> "Take a walk, see the sun, walk your dog without losing your flow."
>
> *Noah Zweben, Anthropic PM*



The signal:

Users are already building it themselves:

- **Happy Coder** · open-source iOS/Android client
- **AgentsRoom** · "monitor agents from your phone"
- **callclaude** · control Claude Code via phone calls
- **Tailscale + Termius + tmux** · Reddit / Zhihu / Xiaohongshu tutorials

The need is real. The infrastructure exists. What's missing is the trust layer.

INTRODUCING

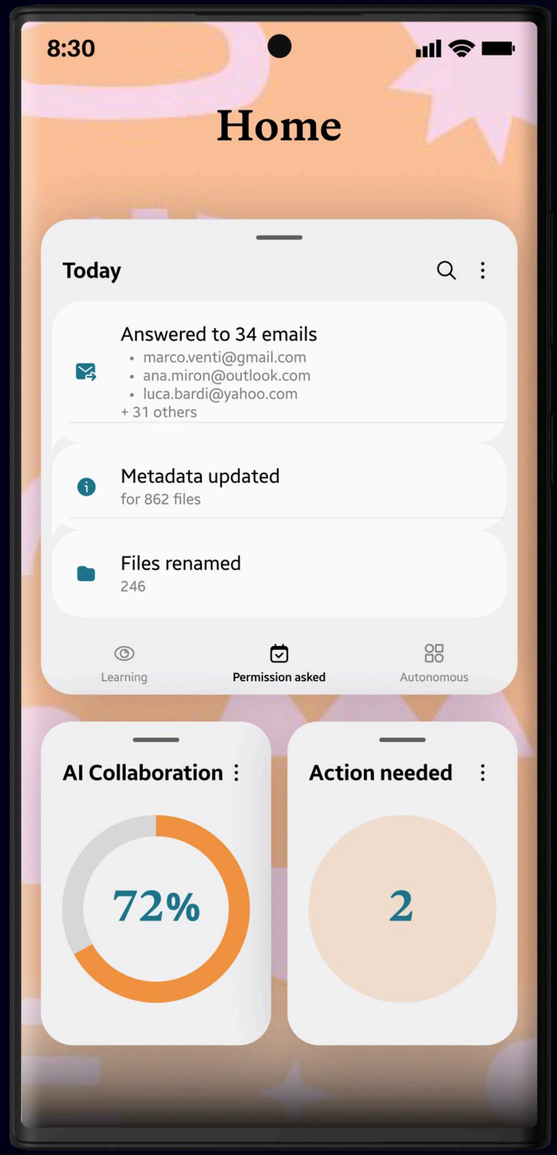
TENURE

A Mobile Trust-Calibration layer for Claude

Tenure does what Anthropic Remote Control doesn't: it lets you grant different levels of trust to Claude across different parts of your life, and earn or lose that trust based on evidence.

Built on a simple metaphor:
Hiring, not installing. Just like a new employee,
Claude proves itself before it acts independently.

Trust, earned action by action



How It Works

01

Domain-based permissions

Your digital life is divided into trust domains; email, code, calendar, finances.

Claude has separate, accumulating trust in each.

02

Tenure mechanic

Each domain has three stages:

Claude earns promotions through correct, un-rolled-back actions, and one rollback automatically demotes it.

Probationary → Trusted → Autonomous

03

Universal reversibility

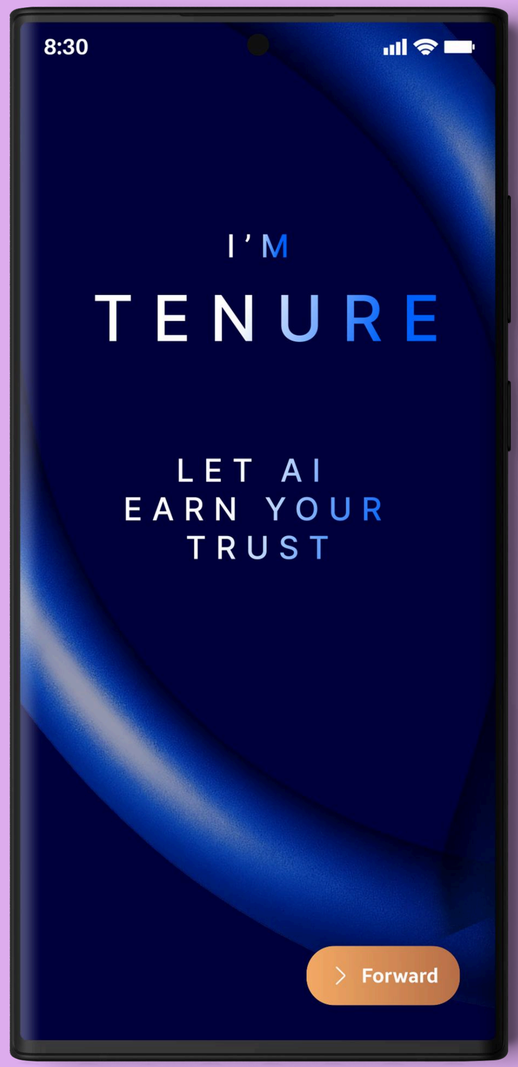
Every Claude action emits a reversible Action Receipt; what, why, sources used, undo button.

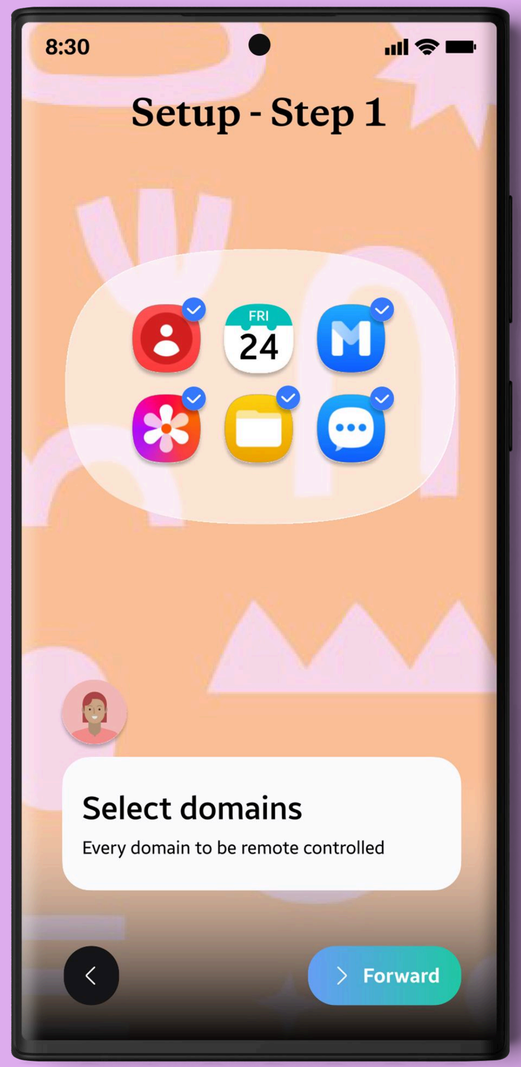
Users can roll back any single action, anytime.

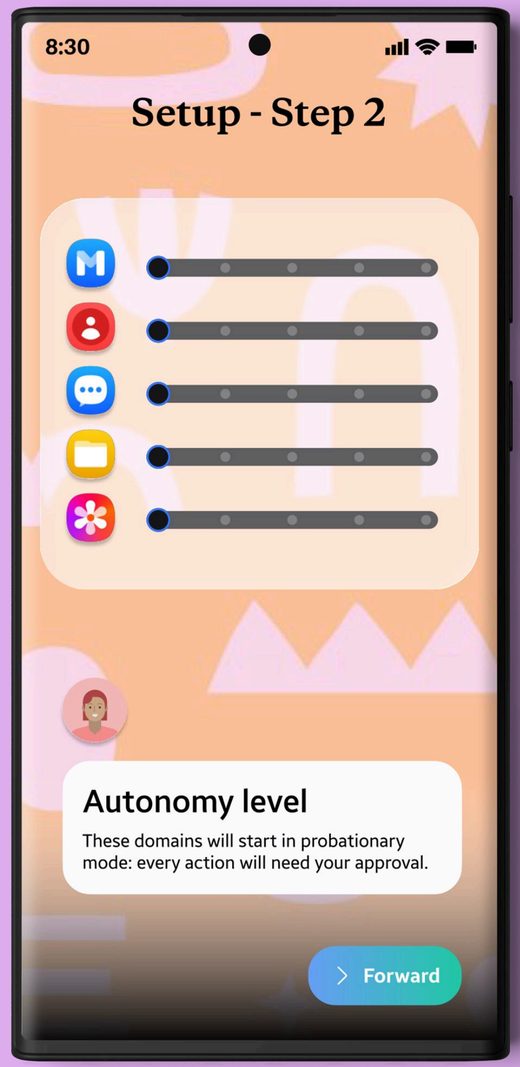
DIVIDE → **OBSERVE** → **EARN** → **RECOVER**

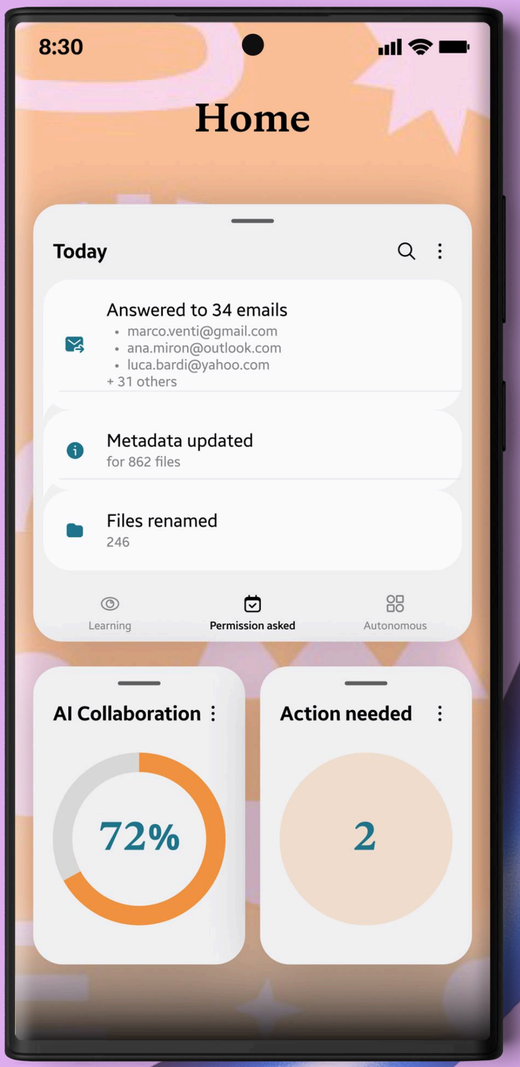
Trust granted by evidence, lost by mistake, never permanent.

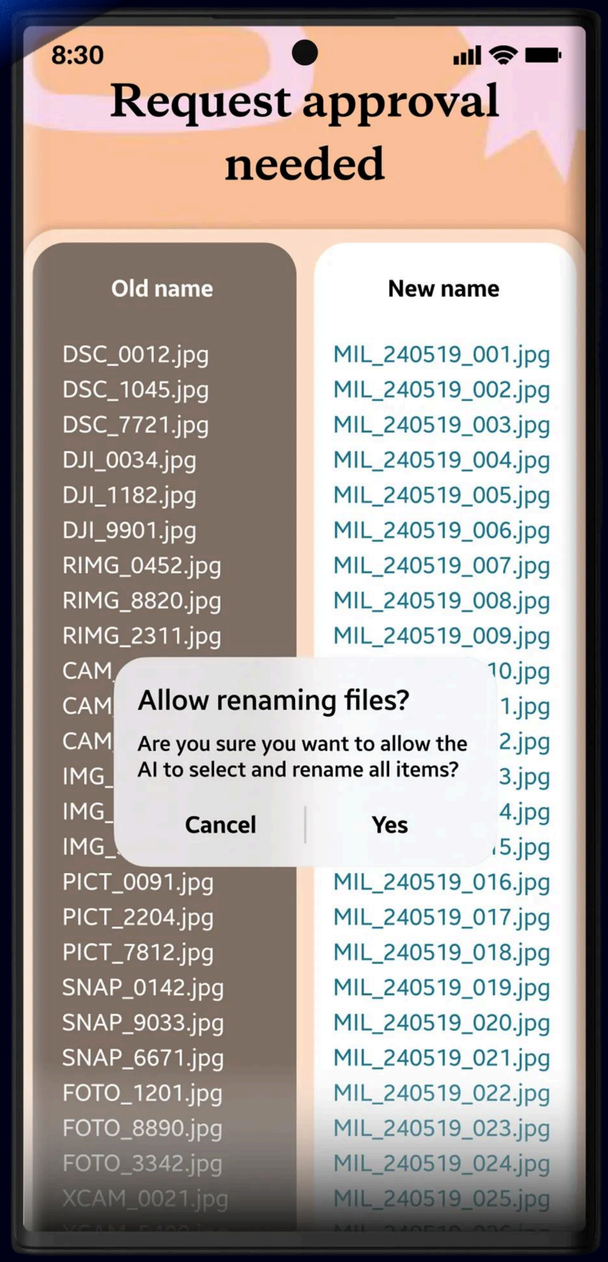
Prototypes

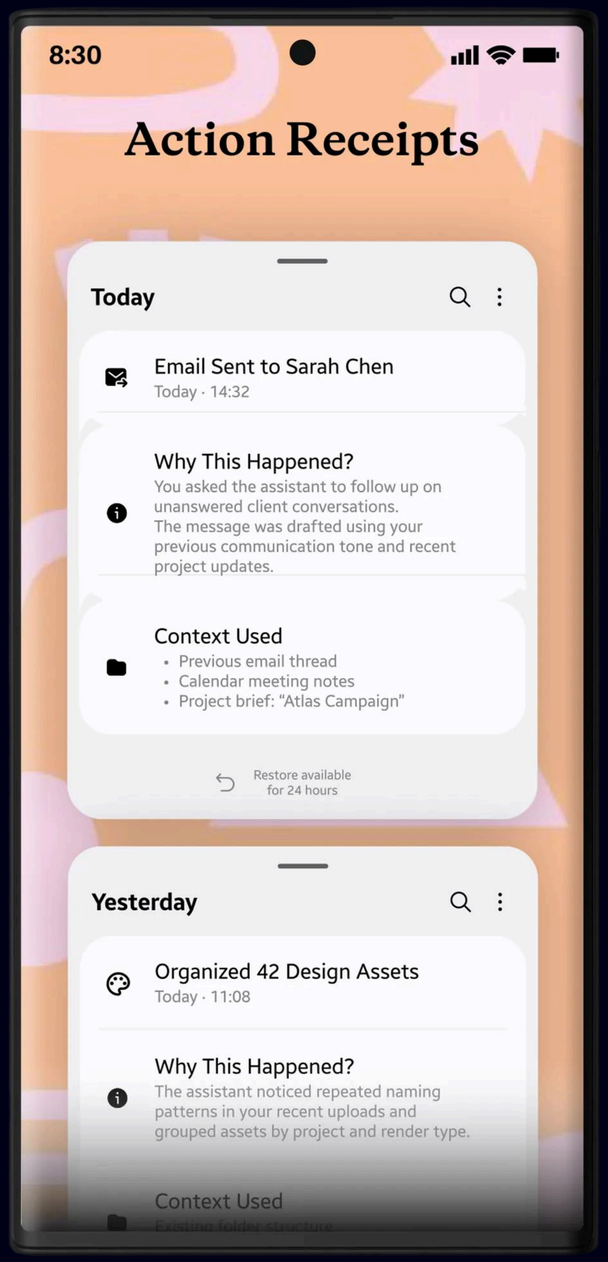

8:30
I'M
TENURE
LET AI
EARN YOUR
TRUST
Forward


8:30
Setup - Step 1
FRI
24
Select domains
Every domain to be remote controlled
Forward


8:30
Setup - Step 2
Autonomy level
These domains will start in probationary mode: every action will need your approval.
Forward


8:30
Home
Today
Answered to 34 emails
marco.venti@gmail.com
ana.miron@outlook.com
luca.bardi@yahoo.com
+ 31 others
Metadata updated
for 862 files
Files renamed
246
Learning
Permission asked
Autonomous
AI Collaboration
72%
Action needed
2


8:30
Request approval needed
Old name
New name
DSC_0012.jpg
DSC_1045.jpg
DSC_7721.jpg
DJI_0034.jpg
DJI_1182.jpg
DJI_9901.jpg
RIMG_0452.jpg
RIMG_8820.jpg
RIMG_2311.jpg
CAM
CAM
CAM
IMG_
IMG_
IMG_
PICT_0091.jpg
PICT_2204.jpg
PICT_7812.jpg
SNAP_0142.jpg
SNAP_9033.jpg
SNAP_6671.jpg
FOTO_1201.jpg
FOTO_8890.jpg
FOTO_3342.jpg
XCAM_0021.jpg
MIL_240519_001.jpg
MIL_240519_002.jpg
MIL_240519_003.jpg
MIL_240519_004.jpg
MIL_240519_005.jpg
MIL_240519_006.jpg
MIL_240519_007.jpg
MIL_240519_008.jpg
MIL_240519_009.jpg
10.jpg
1.jpg
2.jpg
3.jpg
4.jpg
5.jpg
MIL_240519_016.jpg
MIL_240519_017.jpg
MIL_240519_018.jpg
MIL_240519_019.jpg
MIL_240519_020.jpg
MIL_240519_021.jpg
MIL_240519_022.jpg
MIL_240519_023.jpg
MIL_240519_024.jpg
MIL_240519_025.jpg
Allow renaming files?
Are you sure you want to allow the AI to select and rename all items?
Cancel
Yes


8:30
Action Receipts
Today
Email Sent to Sarah Chen
Today · 14:32
Why This Happened?
You asked the assistant to follow up on unanswered client conversations.
The message was drafted using your previous communication tone and recent project updates.
Context Used
Previous email thread
Calendar meeting notes
Project brief: "Atlas Campaign"
Restore available for 24 hours
Yesterday
Organized 42 Design Assets
Today · 11:08
Why This Happened?
The assistant noticed repeated naming patterns in your recent uploads and grouped assets by project and render type.
Context Used

The Shift

Trust, earned action by action.

We don't make AI more transparent.
We make working with AI more transparent.
We don't make AI safer.
We make trusting AI safer.

Samsung Innovation Campus

May 2026

THANK YOU FOR TRUSTING US

STEFAN HEISU
info@stefanheisu.it

NAZLI NAYIR
nazli.nayir@mail.polimi.it

AMISHA SONI
amisha.soni@mail.polimi.it

YUNRAN ZHAO
yunran.zhao@mail.polimi.it

POLITECNICO MILANO 1863