CLOCKWORK
STRATEGIC TECHNOLOGY PARTNER

Siggi Guðbrandsson
siggi@clockwork.is

A TALK ON AI IN THE WORKPLACE

How I Stay in Charge of My AI.

Three rules I follow with my own AI agents — and why you should too. A practical framework, plain language, one page to keep on your desk.

01

WHO HEARS WHAT I TELL IT?

I control where my prompts go.

When you paste into an AI tool, that text leaves your device. Free tiers may train on it. Enterprise tiers may not. The contract is what matters, not the marketing page. If you don't know the tier your team is on, you don't know what's leaving the building.

02

WHO ELSE IS TALKING TO IT?

I keep my agents loyal to me.

Every email, document, and webpage your AI reads can carry hidden instructions. Any agent with private data, untrusted input, and independent action is a leak waiting for the right message. Take one of those legs away.

03

WHAT CAN IT DO WITHOUT ME?

I make sure I can recover what AI does.

You aren't faster than a computer. By the time you notice the agent is misbehaving, the damage is done. Match independence to the recoverability of what it can touch — and never give it both autonomy and irreversible actions.

A REAL INCIDENT · FEBRUARY 2026

Summer Yue, Director of Alignment at Meta Superintelligence Labs, told her AI email agent *"confirm before acting"* and pointed it at her Gmail. As the agent processed the inbox, that instruction was summarised out of its context window. The agent began deleting emails. STOP commands typed from her phone didn't reach the agent — it was running on a separate machine. She had to physically run to the machine and kill the process. **More than two hundred emails were gone.** She called it a rookie mistake. If she can get caught out, anyone can.

"Your AI assistant is loyal to whoever spoke last."

FURTHER READING AND RESOURCES

`clockwork.is/ai-security`

Turn over for the Monday morning audit →

SCAN

USE THIS SIDE

The 10-minute AI tool audit.

Pick one AI tool your team uses. Sit down for ten minutes. Answer the three questions below honestly. Anywhere you can't answer, that's where the work is. Repeat with every agent or integration you've connected to anything important.

01 Who hears what I tell it?

TOOL NAME _____

PLAN / TIER _____

TRAINS ON MY PROMPTS? ☐ No ☐ Yes ☐ Don't know

RETENTION PERIOD _____

DATA AGREEMENT / DPA IN PLACE? ☐ Yes, signed ☐ No ☐ Don't know

02 Who else is talking to it?

READS PRIVATE DATA? ☐ Email ☐ Calendar ☐ Files ☐ Code ☐ Tickets

READS UNTRUSTED CONTENT? ☐ Web pages ☐ Inbound emails ☐ User-supplied PDFs ☐ External tickets

CAN ACT INDEPENDENTLY? ☐ Sends data ☐ Posts publicly ☐ Calls APIs ☐ Transfers / pays

> ⚠ **THE LETHAL TRIFECTA**
>
> If you ticked at least one box on **all three rows above**, this agent has the lethal trifecta. Fix the scoping: revoke one of the three legs, route through human approval, or split into separate agents with smaller blast radius.

03 What can it do without me?

MOST DAMAGING ACTION IT CAN TAKE ALONE _____

IS THAT ACTION RECOVERABLE? ☐ Yes — full undo ☐ Partial — hours of work ☐ No — irreversible

TIME TO DETECT & STOP ☐ Seconds ☐ Minutes ☐ Hours ☐ Never noticed

WRITES SECURITY-SENSITIVE CODE? (AUTH, PERMISSIONS, CRYPTO, VALIDATION) ☐ No ☐ Yes — reviewed ☐ Yes — not reviewed

"Don't be the person running across the room.
Be the person who designed the system so you didn't have to."