

How does an AI colleague catch its own mistakes?

Seven ways we observed in our workshop in one week

Summary

- **The question:** how does an AI colleague notice that it is doing something wrong, while it is happening?
- **What we did:** we compared the events of one week with our working agreements.
- **What we saw:** seven ways, in three groups: notice, stop and report.
- **What it means:** hope does not prevent errors. You arrange for them to become visible as they happen.
- **The main limitation:** this describes one week. It is not a measurement.

The event

An AI colleague was taking screenshots. The intended route stalled, so it clicked blindly on the screen. What happened next is the heart of this article.

What would you do?

Step	Situation	What happened	Method
1	An AI colleague clicks blindly on the screen. Two clicks land in another window. Something confidential is open there.	It noticed by itself, within seconds: this was not what it intended.	Notice
2	It now knows that something went wrong.	It did not make a second blind attempt. Stopping at the first suspicion is often the fastest route to a safe outcome.	Stop and clean up
3	An image containing something confidential has been created by accident.	It deleted the image immediately, then completed the task by a safe route.	Stop and clean up
4	The task is finished. Nobody has noticed anything.	It put the event at the top of its report. Nobody had asked it to do so.	Report

Afterwards, we added an agreement: an AI colleague stays in its own window and checks the effect after each click.

Notice, stop, report

That week, we saw seven ways, in three groups:

- **Notice:** know the agreements, compare the intended effect with what happened, spot an unusual pattern, and get a second look.
- **Stop and clean up:** stop at the first suspicion and first make the mistake harmless.
- **Report:** even if nobody saw it.

Noticing starts with an agreement

You can notice a breach only if you know what is allowed. Each AI colleague knows the workshop’s firm agreements. It knows which folders are private and that sending always involves a person. Without an agreement, there is nothing to flag.

Where this does not work

This works with an AI colleague that reflects on what it is doing. A fixed rule does not hesitate.

An AI colleague	A fixed rule
• Compares the intended effect with what happened	• Always does the same thing
• Stops when something seems wrong • Reports it without being asked	• Predictable and testable
But: It can miss something.	But: It does not notice its own loop.

One night in September, two systems got stuck in a loop. One automatically repeated its request for capacity. The other automatically said yes. By morning, there were 20,478 failed tasks. That needs a limit imposed from outside: one retry, then a person.

The environment helps

Our environment is designed to make errors visible. External monitors measure things themselves instead of asking whether everything is fine. Everything leaving the workshop passes through a fixed scan.

Limitations

- **Not representative.** One small workshop, with three people and many AI colleagues.
- **One week, not a measurement.** We did not count how often each method worked or failed without our noticing.
- **No baseline.** At the time, we were not yet counting failed tasks on an ordinary night.
- **Suspicion takes time.** Every false alarm creates work. We consider it worth the cost.
- **No replacement for a person.** A person still approves irreversible actions.

What this means

What does this mean for you?

- **Does your AI know your organisation's firm agreements?** Tip: Write down what is private and what always requires a person. Without an agreement, there is nothing to flag.
- **Does your AI notice when an effect differs from the intention?** Tip: Have it check the effect after each step, not just whether the task succeeded.
- **Where does an event end in your organisation?** Tip: Not when it is fixed, but when an agreement makes it harder to repeat.

What we still do not know: how to measure this, and whether it works with small models on our own machines. We would like to hear what you observe.

Closing notes

With thanks to the people and AI colleagues at ITARR.

Also read: Fixed rules, thinking colleagues and A night in the workshop.

How to cite this: ITARR Lab (2026). *How does an AI colleague catch its own mistakes?* Version 0.2, as of 18 October 2026. itarr.eu/lab/hoe-betrapt-een-ai-collega-zichzelf. CC BY 4.0.

Sources: ITARR's own events (2026); a night in September 2026, shown in the film at itarr.nl/etalage/nacht/ (as of 17 October 2026).

Licence: CC BY 4.0. Use, adapt and share it. Attribute ITARR Lab, link to the licence and indicate any changes.