

Hoe betrapt een AI-collega zichzelf?

Zeven manieren die we in één week in onze werkplaats zagen

Samenvatting

- **De vraag:** hoe merkt een AI-collega dat hij iets fout doet, op het moment zelf?
- **Wat we deden:** we legden de voorvallen van één week naast onze afspraken.
- **Wat we zagen:** zeven manieren, in drie groepen: merken, stoppen en melden.
- **Wat het betekent:** hoop voorkomt geen fouten. Je laat ze opvallen zodra ze ontstaan.
- **De grootste beperking:** dit is een beschrijving van één week, geen meting.

Het voorval

Een AI-collega maakte schermafbeeldingen. De nette route liep vast, en hij klikte blind op het scherm. Wat er daarna gebeurde, is de kern van dit stuk.

Wat zou jij doen?

Stap	Situatie	Wat er gebeurde	Manier
1	Een AI-collega klikt blind op het scherm. Twee klikken komen in een ander venster terecht. Daar staat iets vertrouwelijks open.	Hij zag het zelf, binnen seconden: dit was niet wat hij wilde.	Merken
2	Hij weet nu dat het misging.	Hij deed geen tweede blinde poging. Stoppen bij het eerste vermoeden is vaak de snelste weg naar een veilig resultaat.	Stoppen en opruimen
3	Er is per ongeluk een beeld gemaakt met iets vertrouwelijks erop.	Hij gooide het beeld meteen weg. Daarna maakte hij het werk langs een veilige weg af.	Stoppen en opruimen
4	Het werk is af. Niemand heeft iets gemerkt.	Hij zette het voorval bovenaan zijn verslag. Niemand had erom gevraagd.	Melden

Daarna kwam er een afspraak bij: een AI-collega blijft in zijn eigen venster, en controleert na elke klik het effect.

Merken, stoppen, melden

In die week zagen we zeven manieren, in drie groepen:

- **Merken:** de afspraken kennen, vergelijken wat je wilde met wat er gebeurde, een vreemd ritme zien, en een tweede blik.
- **Stoppen en opruimen:** stoppen bij het eerste vermoeden, en eerst onschadelijk maken wat fout ging.
- **Melden:** ook als niemand het zag.

Merken begint bij een afspraak

Merken kan alleen als je weet wat mag. Elke AI-collega kent de harde afspraken van het huis. Hij weet welke mappen privé zijn, en dat versturen altijd via een mens gaat. Zonder afspraak valt er niets op.

Waar het niet werkt

Dit werkt bij een AI-collega die nadenkt over wat hij doet. Een vaste regel twijfelt niet.

Een AI-collega	Een vaste regel
• Vergelijkt wat hij wilde met wat er gebeurde	• Doet altijd hetzelfde
• Stopt bij een vermoeden • Meldt het uit zichzelf	• Voorspelbaar en toetsbaar
Maar: Hij kan iets missen.	Maar: Hij merkt zijn eigen lus niet.

Op een nacht in september raakten twee systemen in een lus. Het ene herhaalde vanzelf zijn vraag om ruimte. Het andere zei vanzelf ja. 's Ochtends stonden er 20.478 mislukte opdrachten. Daar heb je een grens van buitenaf nodig: één herkansing, daarna een mens.

De omgeving helpt mee

Onze omgeving is zo gebouwd dat fouten zich melden. Bewakers van buiten meten zelf, in plaats van te vragen of alles goed gaat. Alles wat naar buiten gaat, komt langs een vaste scan.

Beperkingen

- **Niet representatief.** Eén kleine werkplaats, met drie mensen en veel AI-collega's.
- **Eén week, geen meting.** We telden niet hoe vaak elke manier werkte, of faalde zonder dat we het zagen.
- **Geen vergelijkingspunt.** Hoeveel opdrachten er op een gewone nacht mislukken, telden we toen nog niet.
- **Achterdocht kost tijd.** Elk vals alarm is werk. Wij vinden dat de prijs waard.
- **Geen vervanging van een mens.** Wat niet terug kan, tekent nog steeds een mens.

Wat dit betekent

Wat betekent dit voor jou?

- **Kent jouw AI de harde afspraken van je huis?** Tip: Schrijf op wat privé is en wat altijd via een mens gaat. Zonder afspraak valt er niets op.
- **Merkt jouw AI het als het effect anders is dan bedoeld?** Tip: Laat na elke stap het effect controleren, niet alleen of de opdracht lukte.
- **Waar eindigt bij jou een voorval?** Tip: Niet bij opgelost, maar bij een afspraak die herhaling moeilijk maakt.

Wat we zelf nog niet weten: hoe je dit meet, en of het werkt bij kleine modellen op eigen machines. We horen graag wat jij ziet.

Slot

Met dank aan de mensen en AI-collega's van ITARR.

Lees ook: "Vaste regels, denkende collega's" en "Een nacht in de werkplaats".

Zo citeer je dit: ITARR Lab (2026). *Hoe betrapt een AI-collega zichzelf?* Versie 0.2, peildatum 18-10-2026. itarr.nl/lab/hoe-betrapt-een-ai-collega-zichzelf. CC BY 4.0.

Bronnen: eigen voorvallen van ITARR (2026); een nacht in september 2026, als film op itarr.nl/etalage/nacht/ (peildatum 17-10-2026).

Licentie: CC BY 4.0. Gebruik, bewerk en deel het. Noem ITARR Lab als bron, verwijst naar de licentie en zeg het als je iets veranderde.