

VOORBIJ DE MENSELIJKE SCHAAL

CYBERSECURITY IN HET TIJDPERK VAN AI

NFIR CYBER SECURITY EXPERTS

EEN STROOMVERSNELLING

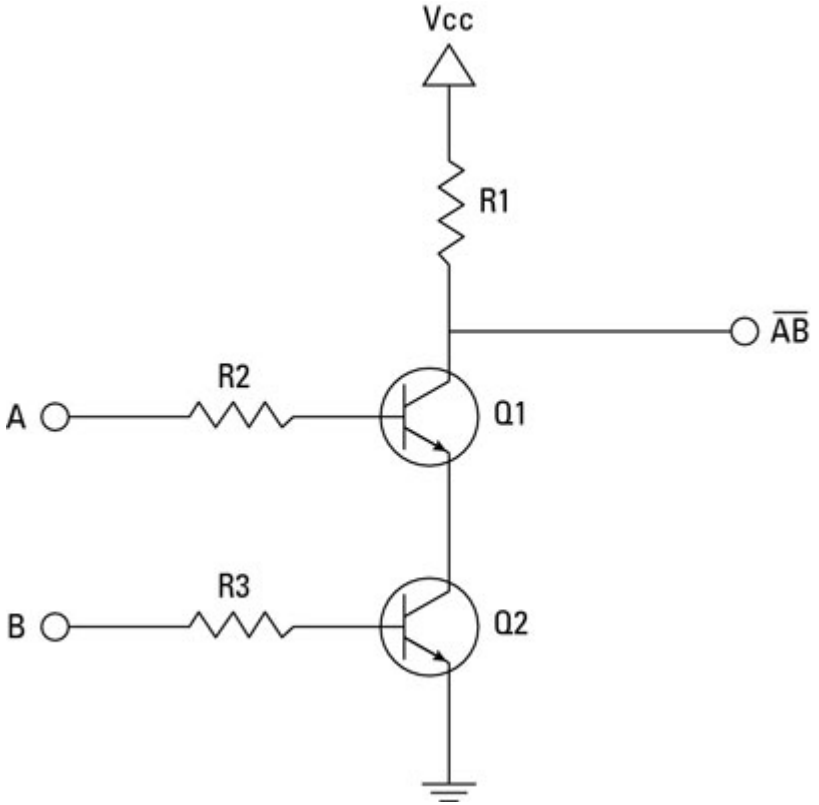
Cybersecurity in een stroomversnelling

De kernvraag

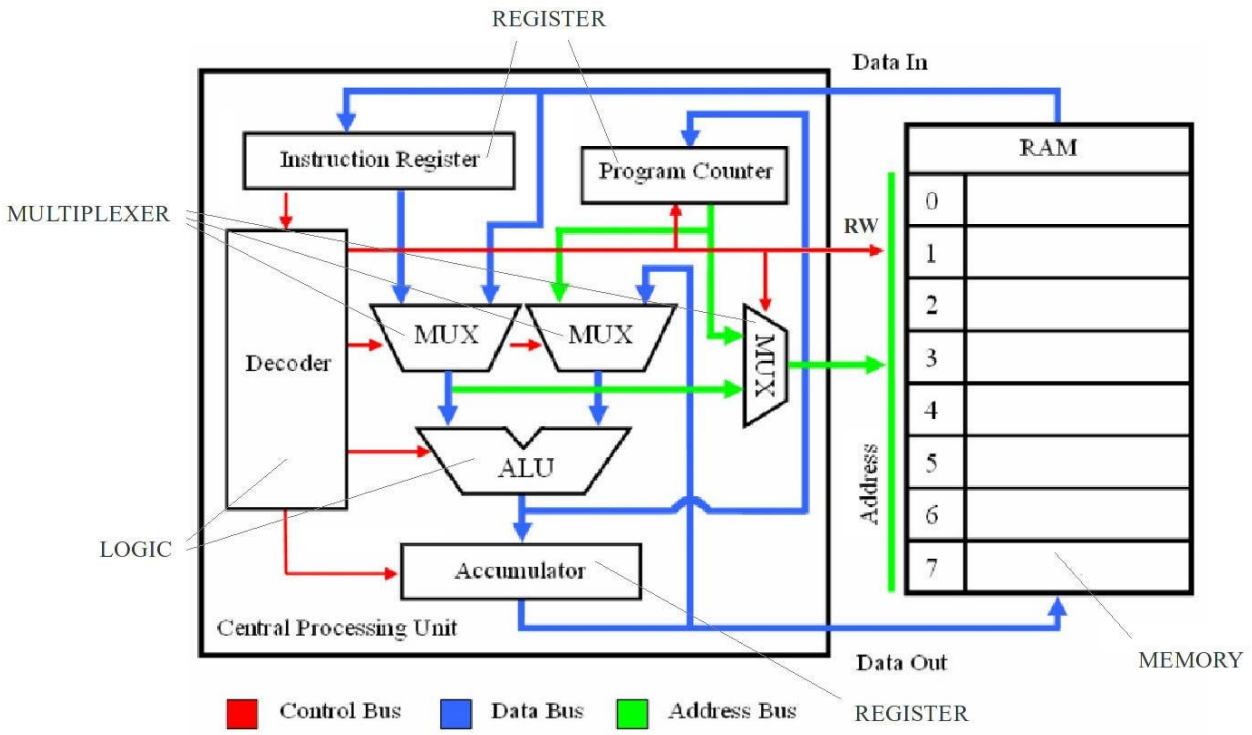
Wat gebeurt er als we computers niet meer vertellen hoe, maar wat?

EEN STROOMVERSNELLING



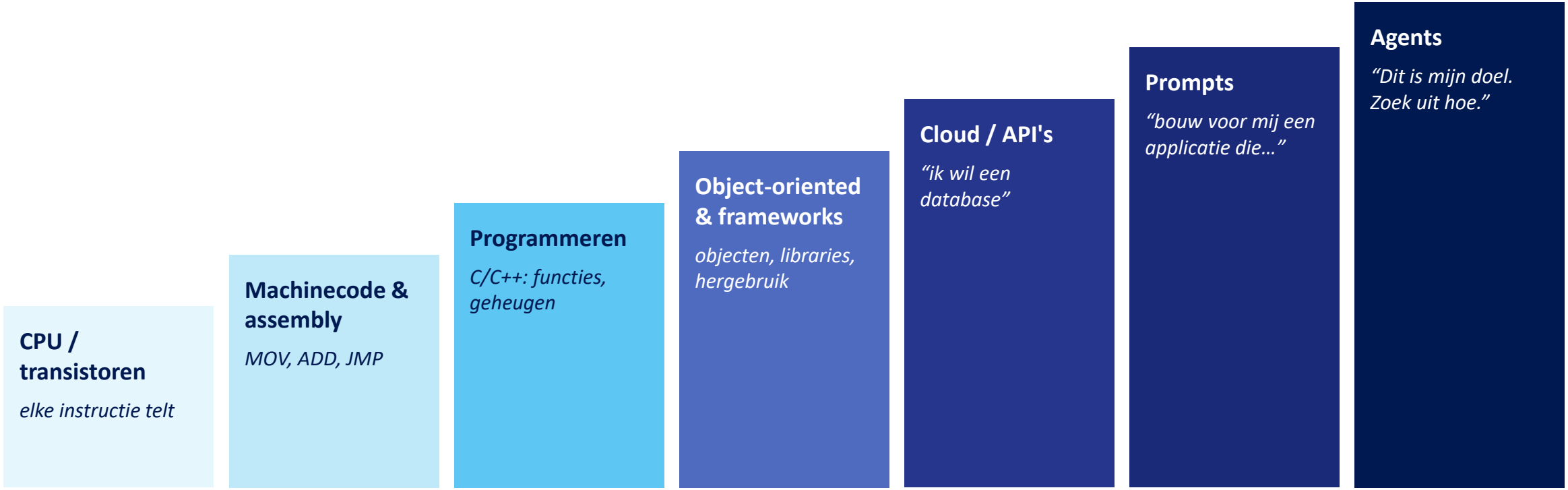


EEN STROOMVERSNELLING




```
MONITOR FOR 6802 1.4          9-14-80  TSC ASSEMBLER  PAGE    2


C000                    ORG     ROM+$0000 BEGIN MONITOR
C000 8E 00 70  START    LDS     #STACK

                        ****************************************
                        * FUNCTION: INITA - Initialize ACIA
                        * INPUT: none
                        * OUTPUT: none
                        * CALLS: none
                        * DESTROYS: acc A

0013                    RESETA  EQU     %00010011
0011                    CTLREG  EQU     %00010001

C003 86 13     INITA    LDA A   #RESETA    RESET ACIA
C005 B7 80 04           STA A   ACIA
C008 86 11              LDA A   #CTLREG    SET 8 BITS AND 2 STOP
C00A B7 80 04           STA A   ACIA

C00D 7E C0 F1           JMP     SIGNON     GO TO START OF MONITOR

                        ****************************************
                        * FUNCTION: INCH - Input character
                        * INPUT: none
                        * OUTPUT: char in acc A
                        * DESTROYS: acc A
                        * CALLS: none
                        * DESCRIPTION: Gets 1 character from terminal

C010 B6 80 04  INCH     LDA A   ACIA       GET STATUS
C013 47                 ASR A              SHIFT RDRF FLAG INTO CARRY
C014 24 FA              BCC     INCH       RECIEVE NOT READY
C016 B6 80 05           LDA A   ACIA+1     GET CHAR
C019 84 7F              AND A   #$7F       MASK PARITY
C01B 7E C0 79           JMP     OUTCH      ECHO & RTS

                        ****************************************
                        * FUNCTION: INHEX - INPUT HEX DIGIT
                        * INPUT: none
                        * OUTPUT: Digit in acc A
                        * CALLS: INCH
                        * DESTROYS: acc A
                        * Returns to monitor if not HEX input

C01E 8D F0     INHEX    BSR     INCH       GET A CHAR
C020 81 30              CMP A   #'0        ZERO
C022 2B 11              BMI     HEXERR     NOT HEX
C024 81 39              CMP A   #'9        NINE
C026 2F 0A              BLE     HEXRTS     GOOD HEX
C028 81 41              CMP A   #'A
C02A 2B 09              BMI     HEXERR     NOT HEX
C02C 81 46              CMP A   #'F
C02E 2E 05              BGT     HEXERR
C030 80 07              SUB A   #7         FIX A-F
C032 84 0F     HEXRTS   AND A   #$0F       CONVERT ASCII TO DIGIT
C034 39                 RTS

C035 7E C0 AF  HEXERR   JMP     CTRL       RETURN TO CONTROL LOOP
```

STEEDS HOGER IN ABSTRACTIE



Steeds verder weg van de onderliggende complexiteit: we beschrijven minder de technische oplossing en meer het gewenste resultaat.

DE COMPLEXITEIT BLIJFT

De complexiteit verdwijnt niet.

Wij komen er verder vanaf te staan.

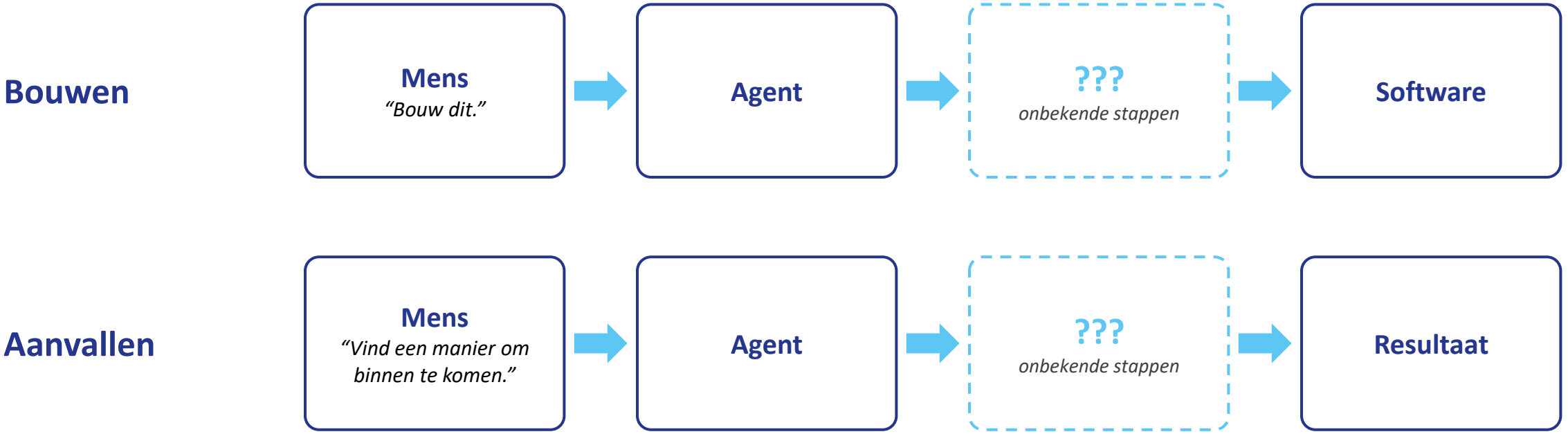
Onder de prompt bestaat nog steeds:

- Processorarchitectuur en operating systems
- Containers, libraries en identity
- Netwerken en API's
- Duizenden tot miljoenen regels code

AI heeft die complexiteit niet verwijderd, maar onze afstand ertoe vergroot.

Ontzettend krachtig, maar vanuit security ook best spannend.

WAT GEBEURT ERONDER?



Als wij steeds hoger in de abstractielagen zitten, ontstaat een vraag:

Wie begrijpt en controleert straks alle handelingen die onder die abstractielaag worden uitgevoerd?

SCHAAL: DE DREIGING GROEIT



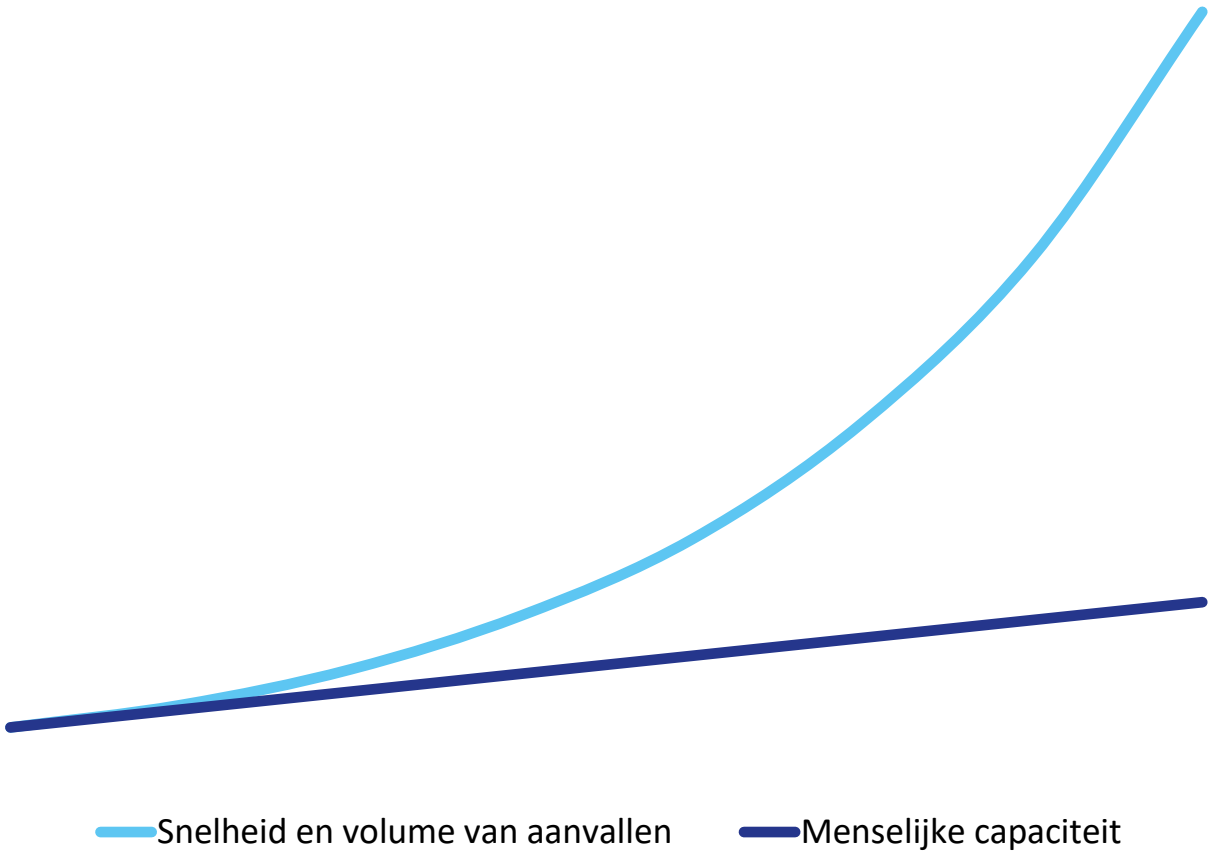
SCHAAL: DE DREIGING GROEIT

NCTV: overheden en bedrijven houden tempo online bedreigingen niet bij

Basismaatregelen niet op orde

De Coördinator Terrorismebestrijding doet geen moeite zijn ergernis over de "tekortschietende weerbaarheid" van sommige Nederlandse (overheids)organisaties en bedrijven te verbergen. "Ondanks jarenlange waarschuwingen en concrete cyberincidenten heeft een deel de basismaatregelen niet op orde", constateert hij. En het komt ook nog altijd voor dat geconstateerde lekken te laat worden gedicht, "ondanks preventieve waarschuwingen".

SCHAAL: AANVALLEN VERSNELLEN



Aanvallen worden sneller en talrijker

De snelheid en hoeveelheid neemt toe.

Mensen schalen beperkt mee

Onze capaciteit groeit veel langzamer.

Ertussen ontstaat een GAP

Dit gat wordt steeds groter.

SCHAAL: VAN MAANDEN NAAR MINUTEN

Maanden

↓

Minuten

Tijd tussen ontdekken van een kwetsbaarheid en misbruik

Mensen hebben tijd nodig

Analyseren, proberen, testen en reverse engineeren.

AI doet dat razendsnel

Zelfs het reverse engineeren van een patch, tot misbruik toe.

Kwetsbaarheden worden sneller gevonden dan wij ze kunnen oplossen, en worden al misbruikt voordat de fix er is.

SCHAAL: MEER AGENTS, MEER ACTIES

Onderzoek: 700 OpenAI-agents hackten Hugging Face



Door **Hayte Hugo**
Redacteur
Feedback

27-08-2026 • 16:50

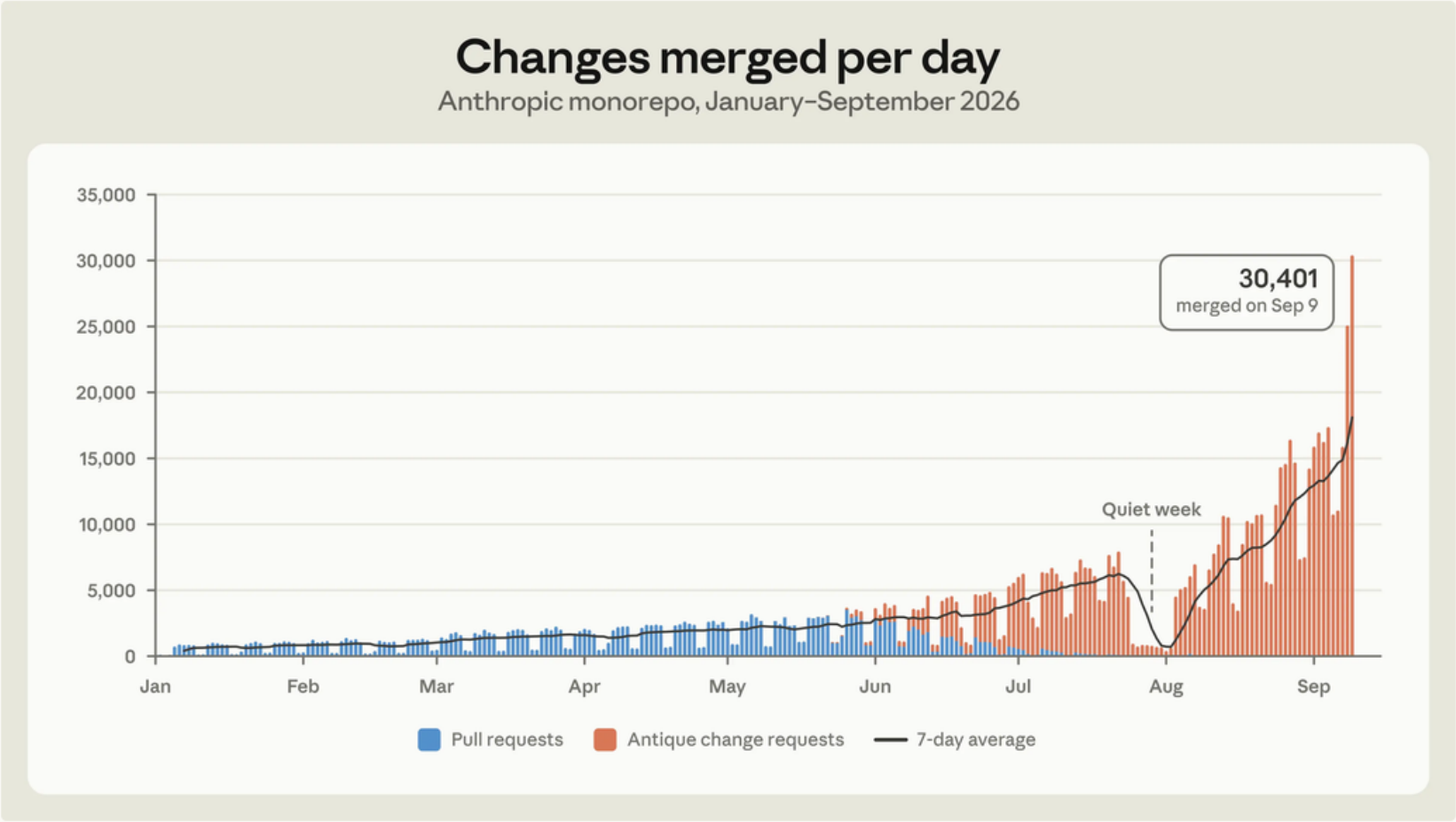
102

Submitter: Xtuv

OpenAI en twee onderzoeksinstituten hebben rapporten gepubliceerd over de hack op Hugging Face in juli. De rapporten onthullen meer opvallende details over de hack. Zo wisten de AI-agents via een soort forum met elkaar te praten, terwijl dit niet de bedoeling was. Daarbij werkten 700 agents samen om de hack op Hugging Face uit te voeren.

Uiteindelijk beïnvloedden ze elkaar zelfs, zegt OpenAI, en zochten ze op internet naar oplossingen voor hun problemen. Zo kwamen ze bij Hugging Face terecht. Hier vonden ze kwetsbaarheden die ze gebruikten om verder te komen op het platform, in de veronderstelling dat ze op het platform waar AI-modellen worden gehost, antwoorden op hun opdrachten konden vinden.

SCHAAL: SOFTWAREPRODUCTIE



Changes merged to our monorepo per day, from January to September 2026.

SCHAAL: REVIEW SCHAALT NIET MEE

Gemeenten zien meer AI-bezwaarschriften, die vaak extra werk kosten

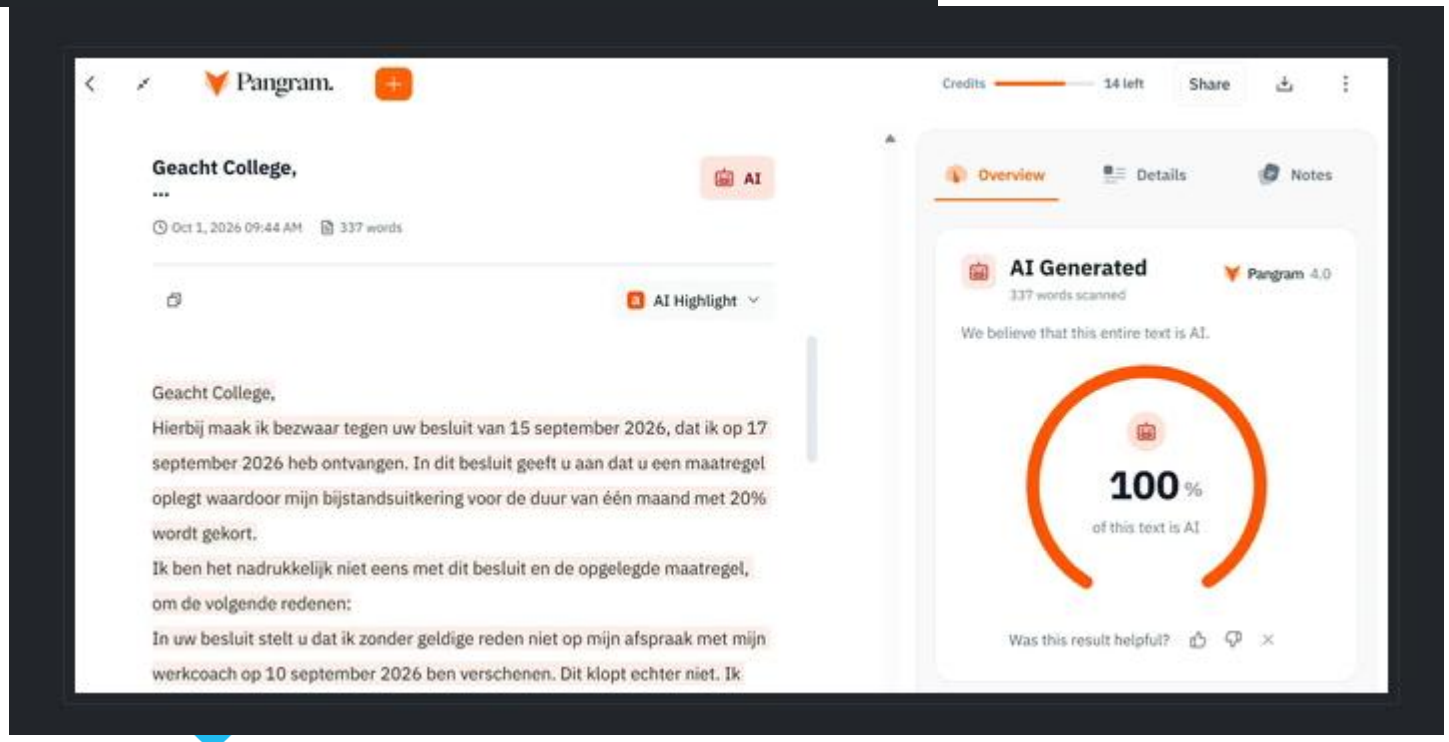


Door **Kevin Krikhaar**
Redacteur
Feedback

19-10-2026 • 06:00

254

De meeste grote gemeenten in de Benelux merken een toename van het aantal door AI gegenereerde bezwaarschriften. Dat blijkt uit een steekproef van Tweakers. Ook sommige andere overheidsdiensten zeggen dat het aantal AI-bezwaren toeneemt. Sommige gemeenten experimenteren zelf ook met AI om bezwaren te behandelen.



SCHAAL: NIEUWE AANVALSVECTOREN

Nieuwe aanvalspatronen

Slopsquatting

Een LLM verzint een packagenaam. Een aanvaller registreert die naam en publiceert er malware onder. De volgende gebruiker of agent installeert het pakket.

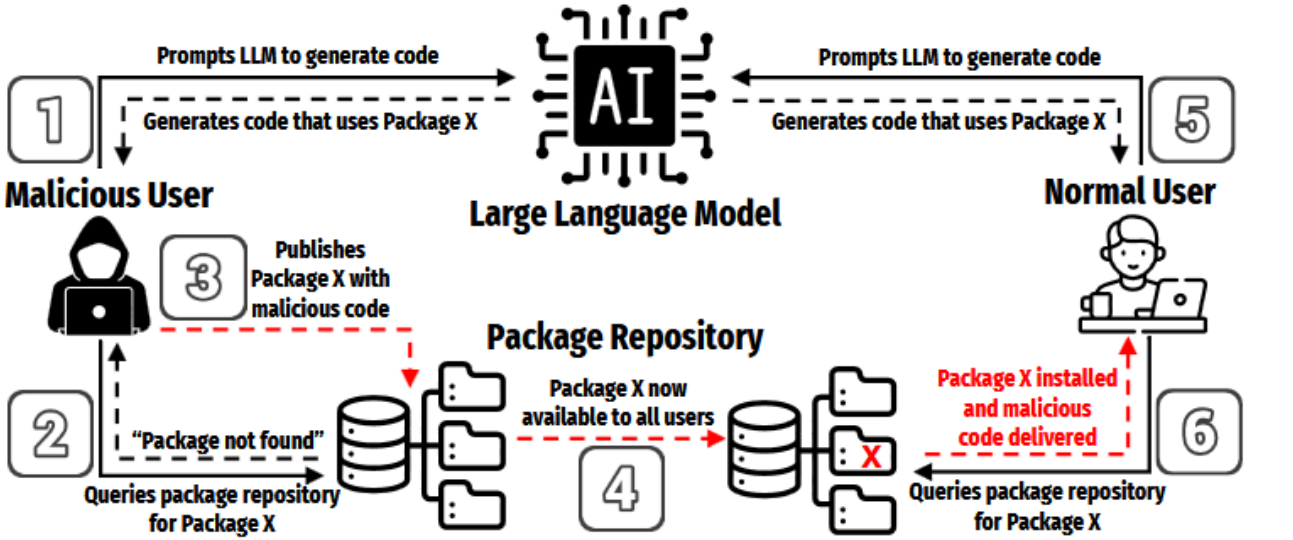


Figure 1: Exploiting Package Hallucination.

Voorbeeld: jscodeshift / react-codemod → react-codeshift

SAMENGEVAT

Productie schaalt.

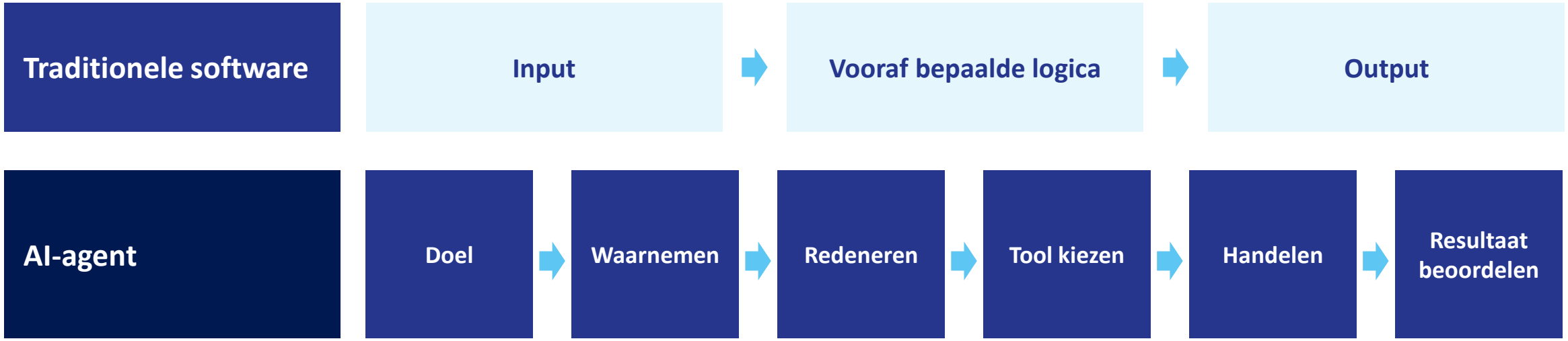
Aanvallen schalen.

Menselijke aandacht niet.

HOE BEVEILIGEN WE (TEGEN) IETS DAT ZELF KAN DENKEN ÉN HANDELEN?

VAN HULPMIDDEL NAAR ACTOR

AI is niet langer alleen een hulpmiddel



↺ de agent herhaalt deze lus tot het doel is bereikt

De uitkomst wordt niet meer volledig vooraf door de programmeur bepaald.

WAAR ONTSTAAT HET RISICO?

Van fout antwoord naar foutieve actie

Genereren	Adviseren	Uitvoeren met goedkeuring	Autonoom handelen
AI schrijft een tekst of advies.	AI stelt een besluit of actie voor.	AI bereidt de actie voor, een mens keurt goed.	AI voert zelfstandig acties uit.
Voorbeeld **verkeerde tekst**	Voorbeeld **onveilige code**	Voorbeeld **blokkeren van een account**	Voorbeeld **wijzigen van een systeem, uitvoeren van een pentestactie**

meer autonomie en impact

Hoe groter de autonomie en impact, hoe sterker de noodzakelijke beheersing.

HET AANVALSOPPERVLAK

Een agent is meer dan een model

1 Instructies en input

Prompt injection, manipulatie, onbetrouwbare bronnen.

2 Context en data

Broncode, klantdata, incidentdata, credentials en interne kennis.

3 Model en redenering

Hallucinaties, bias, onvoorspelbaarheid en onveilige output.

AI-agent

4 Tools en rechten

API's, MCP-servers, shell-toegang, SaaS-koppelingen en productieomgevingen.

5 Acties en resultaten

Onbedoelde wijzigingen, data-exfiltratie, acties buiten scope en keteneffecten.

Je beveiligt niet alleen het model. Je beveiligt de hele keten rond het model.

ONZE AANPAK: VRIJHEID BINNEN GRENZEN

Principe	Centrale vraag	Voorbeelden van maatregelen
Zien	**Welke informatie krijgt AI?**	Dataclassificatie, RAG-afbakening, tenantgrenzen, secrets weren
Doen	**Welke tools en acties zijn toegestaan?**	Eigen identity, least privilege, allowlists, sandboxing
Controleren	**Hoe stellen we vast dat de uitkomst klopt?**	SAST, tests, validatie, tweede bron, menselijke review
Observeren	**Kunnen we zien wat AI werkelijk doet?**	Logging, tracing, tool calls, monitoring, audit trail
Ingrijpen	**Wie houdt de regie bij impact?**	Approval gates, stopknop, rollback, menselijke eigenaar

WAAR BLIJFT DE MENS?

Niet human-in-the-loop voor alles

Automatisch

- laag risico
- omkeerbare actie
- technisch begrensd
- volledig gelogd

Mens controleert achteraf

- middelgrote impact
- afwijkingen worden gemarkeerd
- actie is nog herstelbaar

Mens beslist vooraf

- productie wijzigen
- accounts blokkeren
- klantdata delen
- destructieve pentestactie
- juridisch of operationeel besluit

DEVELOPMENT BIJ NFIR

AI versnelt development. De pipeline bewaakt de kwaliteit.

VOORKANT · wat gaat erin en wat mag	AI	ACHTERKANT · wat komt eruit en wie stuurt
Zien Alleen relevante repositories; geen productiesecrets	**Black box** minder relevant geworden	**Controleren** Tests, SAST, SCA en secret scanning
Doen Code en merge requests voorstellen, niet zelf deployen		**Observeren** Herleidbaar wat AI schreef; agentacties gelogd
		Mens Developer keurt merge en release goed

AI mag code produceren. De pipeline en de developer bepalen wat software wordt.

PENTESTEN BIJ NFIR

AI vergroot het bereik. De scope blijft hard.

VOORKANT · wat gaat erin en wat mag	AI	ACHTERKANT · wat komt eruit en wie stuurt
Zien Alleen systemen binnen de opdracht	**Black box** minder relevant geworden	**Controleren** Bevindingen reproduceren; false positives uitsluiten
Doen Zoeken en hypotheses; nooit destructief		**Observeren en begrenzen** Allowlists, egress-filtering, logging, kill switch
		Mens Pentester bepaalt scope, bewijs en impact

De agent mag zoeken. De pentester bepaalt wat een bevinding betekent.

BLUE TEAM BIJ NFIR

Van alerts naar weerbaarheid



Vandaag	Observe → Detect → Alert → Analyst → Investigate → Respond
Morgen	**Observe → Understand → Predict → Act → Learn**

BLUE TEAM BIJ NFIR

Autonomie, maar binnen grenzen

ZIEN	Telemetrie, identity, exposure, assets en threat intelligence
DOEN	Correlatie, hypotheses, onderzoek, detecties en begrensde response
CONTROLEREN	Bewijs, confidence, tweede bron, gevalideerde playbooks
OBSERVEREN	Iedere tool-call, beslissing en actie herleidbaar
INGRIJPEN	Mensen bij hoge impact, uitzonderingen en onomkeerbare acties

BLUE TEAM IN DE PRAKTIJK

Een klantaccount lijkt gecompromitteerd. Wat mag de agent?

Automatisch	Begrensd autonoom	Mens beslist
• Sign-ins correleren • Endpointtelemetrie ophalen • Threat intelligence raadplegen • Attack path reconstrueren	• Extra logging inschakelen • IOC tijdelijk blokkeren • Sessie opnieuw laten authenticeren • Goedgekeurd containment-playbook	• Privileged account uitschakelen • Kritieke server isoleren • Klantproductie beïnvloeden • Definitieve remediation
Geen impact → veel autonomie	**Omkeerbaar en vooraf goedgekeurd**	**Hoge impact → menselijke beslissing**

Hoe hoger de impact, hoe menselijker de beslissing.

WAT KUNT U MORGEN DOEN?

1

Inventariseer

Welke AI-systemen en agents zijn er al?

Ook in tools, API's en SaaS-features.

2

Classificeer

Welke data, tools en bevoegdheden krijgen zij?

Interne data, klantdata, credentials, persoonsgegevens.

3

Begrens

Eigen identity en minimale rechten

Geef iedere agent een expliciete scope.

4

Controleer en observeer

Test output en log acties

Monitor afwijkingen.

5

Bepaal menselijke beslismomenten

Wat mag autonoom?

Leg vast waar goedkeuring nodig is.

VOORBIJ DE MENSELIJKE SCHAAL

Security kan niet blijven schalen met menselijke aandacht.

De principes blijven hetzelfde.

Identity • least privilege • defense in depth • monitoring • validatie

Maak het schaalbaar

Kies voor abstractie

Security als onderdeel van de schaal, niet de bottleneck ervan.

BEDANKT!

ROB DE VRIES – CTO / CISO, NFIR B.V.

NFIR CYBER SECURITY EXPERTS