Steeds betere AI-modellen ontwrichten de wereld van cybersecurity. ‘De verdediging staat eigenlijk met 4-0 achter’

Jorit Verkerk

De alarmistische manier waarop AI-bedrijven vaak over hun nieuwste modellen praten, leidt steevast tot scepsis. Toen Anthropic begin april bekendmaakte dat zijn nieuwste AI-model zó goed is in het vinden en exploiteren van kwetsbaarheden in software dat het te gevaarlijk zou zijn om het voor iedereen beschikbaar te maken, haalden sceptici hun schouders op. Het is alsof een snoepwinkel waarschuwt dat de nieuwste lolly’s zó lekker zijn, dat je er maar beter mee kunt oppassen, omdat je er anders verslaafd aan kan raken. *Sure*.

Maar het model in kwestie, Claude Mythos, zet de wereld van cybersecurity wel degelijk op scherp. „De afgelopen weken ben ik alleen maar bezig geweest onze organisatie wakker te schudden”, zegt Frank Breedijk, hoofd digitale veiligheid bij IT-bedrijf Schuberg Philis (meer dan vijfhonderd arbeidsplaatsen). Hij is ook crisismanager bij de DIVD, een Nederlandse vrijwilligersorganisatie van cybersecurity-experts.

Ook onderzoeker Jeroen van der Ham-De Vos van de Universiteit Twente vreest ontwrichting van de digitale infrastructuur die onmisbaar is voor reizen, communiceren en betalen. En uit het bankwezen komen waarschuwingen van gelijke aard: de Nederlandsche Bank waarschuwde deze week dat het betalingsverkeer kan uitvallen door AI-gedreven cyberaanvallen. „AI dwingt ons fundamenteel anders naar cybersecurity te kijken”, zegt Van der Ham-De Vos.

Cybersecurity is altijd al te vergelijken geweest met een klassiek kat-en-muisspel – en het belang ervan groeit naarmate ieders leven dieper in het digitale domein verankerd raakt. Je hebt verdedigers en aanvallers, software is hun strijdtoneel. Software is een complex bouwwerk, opgemaakt uit code die bijna nooit waterdicht is. Doordat AI-modellen zo goed kunnen coderen, zijn ze ook goed in het vinden van kwetsbaarheden in software – ook al zijn ze nooit met dat doel gebouwd. En er zijn nogal wat kwetsbaarheden. Wie lang genoeg zoekt, vindt in vrijwel alle software een zwakke plek, een *bug*. Zowel verdedigers als aanvallers zijn daar continu actief naar op zoek.

Wanneer cybersecuritybedrijven kwetsbaarheden aantreffen, melden ze die bij de softwaremakers zodat die een *patch* (pleister) kunnen maken. Zodra die pleister klaar is, wordt hij met de wereld gedeeld – samen met informatie over de zwakke plek zelf. Die informatie is voor iedereen toegankelijk: voor systeembeheerders die moeten weten hoe urgent een update is, maar ook voor cybercriminelen. Bedrijven of individuen die de kwetsbare software gebruiken, moeten na de patchpublicatie hun computer updaten, of een app bijwerken. Pas dan is de bug, en daarmee het risico op een hack, verleden tijd.

Exploitatietijd

Tien jaar geleden werden zwakke plekken in kwetsbare software minder vaak en minder snel afgestraft. *Bugs* bestonden soms decennialang zonder te worden opgemerkt. „We hebben daardoor heel lang losjes over cybersecurity nagedacht”, zegt onderzoeker Van der Ham-De Vos. „De kans dat misbruik gemaakt werd van *bugs* in je software, was best klein”, zegt Breedijk.

Die tijd is voorbij, laat een grafiek in een recent rapport van non-profitorganisatie Cloud Security Alliance zien. De grafiek geeft de *time-to-exploit* weer: de gemiddelde tijd tussen publiek maken van een zwakke plek en het eerste bewezen misbruik ervan. In 2018 had een organisatie gemiddeld ruim twee jaar om een bekende kwetsbaarheid te repareren voordat iemand er misbruik van maakte. In 2022, het jaar dat ChatGPT werd gelanceerd, duurde het al veel korter: gemiddeld ruim acht maanden. En dit jaar hebben cybercriminelen gemiddeld ruim anderhalve dag nodig om een bekende zwakke plek te exploiteren. De verwachting is dat deze trend zich blijft doorzetten. Wat nu een dag duurt, kan straks in een uur.

Vooraanstaande AI-bedrijven als Anthropic, OpenAI en Alphabet zijn de drijvende krachten achter deze revolutie. De AI-taalmodellen die de bedrijven maken, zijn inmiddels misschien wel beter en ongetwijfeld sneller in het maken van software dan de beste programmeurs, zegt IT-beveiliger Breedijk. Hij begon het telefonisch interview voor dit artikel met de woorden „AI, AI, AI, AI, AI”, op zijn Engels uitgesproken. Om daarna uit te leggen: „Ik heb zelden meegemaakt dat iets zo ontzettend mijn dagtaak veranderd heeft.”

Met behulp van de nieuwste generatie AI-modellen, waar Claude Mythos deel van uitmaakt, kunnen *bugs* in een mum van tijd worden ontdekt. De ontwikkelaar Anthropic zegt bijvoorbeeld dat Mythos een 27 jaar oude *bug* in de software van OpenBSD vond, een besturingssysteem dat er juist om bekendstaat vrijwel ondoordringbaar te zijn. Anthropics onderzoekers hadden Mythos net zo lang verkenningsrondjes door de software laten lopen totdat ze de kwetsbaarheid tegenkwamen. Op 22 mei had Mythos volgens Anthropic meer dan duizend ernstige kwetsbaarheden gevonden in opensourceprojecten.

Met behulp van AI kunnen kwetsbaarheden in software niet alleen sneller worden ontdekt, criminelen kunnen *bugs* ook makkelijker misbruiken om systemen binnen te dringen. Cybercriminelen houden de lijsten met *patches* die softwarebedrijven als Microsoft regelmatig online publiceren nauwlettend in de gaten. Op basis van de vrijgegeven informatie – waar zit de kwetsbaarheid, hoe wordt die gerepareerd – kunnen ze gericht op die code cyberaanvallen uitvoeren. Mensen die nog niet braaf hebben geüpdatet en dus geen *patch* hebben, zijn dan kwetsbaar. „Daar heb je niet eens per se hele goede AI-modellen voor nodig", zegt onderzoeker Van der Ham-De Vos. „Je kunt gewoon tegen je AI-model zeggen: ga daar zoeken en vind het probleem."

'Mooie marketingcampagne'

Die dreiging leidt tot grote zorgen. Onder de naam 'Project Glasswing' heeft Anthropic zijn Claude Mythos begin vorige maand dan ook beperkt beschikbaar gesteld aan een selectie (tech)bedrijven, waaronder Microsoft, Google, Crowdstrike, JPMorgan Chase en de Linux Foundation. De hoop is dat ze die voorsprong gebruiken om kritische kwetsbaarheden in hun software te repareren, voordat kwaadwillenden er misbruik van maken. De komende tijd komen er meer bedrijven bij, zegt Anthropic.

Maar niet alleen Anthropic heeft die ontwikkeling doorgemaakt, ook de concurrentie. „Er wordt ineens moord en brand geschreeuwd over Anthropic", zegt Van der Ham-De Vos, omdat het bedrijf „met een hele mooie marketingcampagne" op de gevolgen van AI voor cybersecurity heeft gewezen, terwijl die gevaren deels al bekend waren en Mythos bovendien niet als enige in staat is om snel zwakke plekken in software te vinden. Dat herkent ook Breedijk, die zijn presentatie over de gevolgen van AI de veelzeggende titel '*This presentation is not about Mythos*' gaf.

Anthropic heeft allang geen monopolie meer op AI-modellen die goed kunnen coderen, en hoewel Mythos absoluut nauwkeuriger, beter en langduriger zelfstandig taken kan uitvoeren dan zijn voorganger, is de sprong in kwaliteit niet zo groot dat je kunt spreken van een revolutionaire ontwikkeling. Uit tests van het Britse AI Security Institute blijkt dat Mythos slechts in beperkte mate beter is in cybersecuritytaken dan AI-modellen van OpenAI – die al breed zijn uitgerold.

Hoe ziet de strijd tegen met AI bewapende hackers er nu uit? Het risico is dat met hulp van AI in korte tijd zoveel zwakke plekken in software worden gevonden dat de tijd ontbreekt om ze allemaal te repareren. Breedijk maakt de vergelijking met de Industriële Revolutie. „Die heeft ons een hoop gebracht. Maar je zal maar arbeider zijn geweest tijdens de Industriële Revolutie … Ik denk dat cyberverdedigers de arbeiders zijn in de AI-revolutie. En we staan als verdediging eigenlijk met 4-0 achter."

Het spel tussen aanvallers en verdedigers is per definitie een ongelijke strijd: de verdedigers moeten alles veilig stellen, terwijl de aanvallers aan één zwakke plek genoeg hebben. „Daarnaast is er gewoon heel veel software in gebruik waar serieuze kwetsbaarheden in zitten", zegt Breedijk. Van der Ham-De Vos: „Iedereen is ook lekker aan het vibecoden [programmeren met behulp van AI, zonder de code echt te begrijpen] en daar komt allemaal brakke software uit. Dat brengt nog meer risico's met zich mee." Hij verwacht snel een „gigantische berg updates" van softwarebedrijven van hun software, omdat zij die nu met behulp van AI hebben doorgelicht.

Daarmee zijn de problemen nog niet opgelost. „We weten namelijk dat veel bedrijven hun patchprocedures niet op orde hebben", zegt Van der Ham-De Vos, die hier onderzoek naar doet. „Het kost ook tijd om softwarereparaties te installeren. Je moet testen of je systemen nog wel blijven werken met alle nieuwe software die je gaat draaien. Het komt eigenlijk nooit goed uit."

Toch is het van groot belang dat bedrijven hun IT-processen snel serieus gaan nemen, zegt hij. Anders verwacht hij een golf aan ransomwareaanvallen en datalekken, nog veel meer dan nu, terwijl bij de hacks bij Clinical Diagnostics, Odido en, recenter, studie-app Canvas de privégegevens van miljoenen Nederlanders al handelswaar werden. De enige manier waarop bedrijven AI-gedreven cybercriminelen kunnen tegenhouden, denkt hij, is door zelf ook AI-modellen in te zetten die zijn toegespitst op cybersecurity. Wie dat niet doet, is straks de klos. Breedijk: „Voor iedere organisatie die moeite heeft om kwetsbaarheden weg te werken, geldt: *your survival is optional*."

Lees ook

Ernst Kuipers na lek bij Clinical Diagnostics: ‘Hackers schenden niet alleen privacy, ze kunnen een heel ziekenhuis platleggen’