



## Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)

05-11-2025:

Onderzoek naar inbraak op Citrix-systemen Openbaar Ministerie nog niet gestart

Twee maanden na de inbraak op de Citrix-systemen van het Openbaar Ministerie (OM) is het onderzoek door een commissie van cyberexperts nog niet begonnen. Het kabinet had eerder aangekondigd een onderzoekscommissie in te stellen om de incidenten te onderzoeken. De inbraak leidde tot het loskoppelen van de interne systemen van het OM, naar aanleiding van waarschuwingen van het Nationaal Cyber Security Centrum (NCSC). De gevolgen van het incident zijn groot voor de strafrechtketen. Een eerste technisch en forensisch onderzoek heeft geen aanwijzingen opgeleverd dat er data is gestolen of aangepast. De minister van Justitie laat weten dat het onderzoek door de commissie zich zal richten op de reactie van het OM op het incident, de genomen maatregelen en de versterking van de IT-weerbaarheid. Het is nog niet duidelijk wanneer het onderzoek zal beginnen en wanneer de commissie wordt aangesteld.

VS sanctioneren Noord-Koreaanse bankiers voor witwassen van gestolen cryptocurrency

De Verenigde Staten hebben sancties opgelegd aan een groep Noord-Koreaanse bankiers en financiële instellingen die beschuldigd worden van het witwassen van geld afkomstig van cybercriminaliteit. Het ministerie van Financiën meldt dat de betrokkenen verantwoordelijk zijn voor het witwassen van fondsen die zijn verkregen door malware-aanvallen en fraude met cryptocurrency. Dit geld zou gebruikt worden om het nucleaire wapenprogramma van Noord-Korea te financieren. De sancties zijn gericht op acht individuen en twee bedrijven, waaronder Noord-Koreaanse bankiers Jang Kuk Chol en Ho Jong Son. Tussen 2022 en 2025 zou Noord-Korea via cyberaanvallen meer dan 3 miljard dollar hebben gestolen, voornamelijk in digitale activa. De VS heeft eerder gewaarschuwd voor Noord-Koreaanse hackers die zich voordoen als IT-werkers om toegang te krijgen tot financiële netwerken en zo sancties te omzeilen.

Russische hackers misbruiken Hyper-V voor malwareverberging in Linux VM's

De Russische hacker-groep Curly COMrades misbruikt Microsoft Hyper-V om malware in een virtuele Linux-machine te verbergen, waarmee ze traditionele



## Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)

endpoint detectie omzeilen. Door een Alpine Linux-gebaseerde VM te creëren, konden ze hun malware veilig uitvoeren zonder detectie door beveiligingssystemen. De malware bestaat uit twee tools: CurlyShell, die commando's uitvoert via HTTPS, en CurlCat, die wordt gebruikt voor het tunnelen van netwerkverkeer. Deze technieken maken gebruik van de Hyper-V virtualisatietechnologie in Windows om netwerkverkeer via de hostmachine te laten verlopen, zodat het lijkt alsof de communicatie van een legitiem systeem komt. Deze aanvallen zijn specifiek gericht op cyberespionage en sluiten aan bij de geopolitieke belangen van Rusland. De groep heeft eerder aanvallen uitgevoerd op overheids- en energiebedrijven in landen als Georgië en Moldavië.

### DDoS-aanval door NoName op meerdere Belgische websites

NoName, een hacktivistische groep, heeft meerdere websites in België aangevallen, waaronder die van verschillende regionale en lokale overheden. De doelwitten omvatten onder andere de websites van Wallonië, de Brusselse lokale autoriteiten, de gemeente Burg-Reuland, de Provincie Luik en de Provincie Vlaams-Brabant. De aanval is uitgevoerd via een gedistribueerde denial-of-service (DDoS)-aanval, wat resulteerde in tijdelijke onbereikbaarheid van deze websites. NoName heeft de verantwoordelijkheid voor de aanval opgeëist, en het incident onderstreept de groeiende dreiging van dergelijke cyberaanvallen gericht op lokale overheidsinstellingen in België.

- Wallonia
- Bruxelles Pouvoirs Locaux
- Burg-Reuland Municipal Administration
- Province of Liège
- Provincie Vlaams-Brabant

### Nieuwe hacktivistische alliantie tussen NoName en Perun Swaroga

De hacktivistische groepen NoName en Perun Swaroga hebben officieel een nieuwe alliantie aangekondigd. Deze samenwerking is significant voor de cyberdreigingslandschap, aangezien beide groepen bekend staan om hun betrokkenheid bij politieke en sociale doelstellingen via cyberaanvallen. De alliantie



## Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)

kan leiden tot een intensivering van aanvallen gericht op overheidsinstellingen, bedrijven en infrastructuur, met mogelijk bredere gevolgen voor de cybersecurity van landen die hun activiteiten als doelwit beschouwen. Deze ontwikkeling ondersteunt de groeiende trend van samenwerkingen binnen hacktivistische groepen die zich steeds vaker organiseren om gezamenlijke doelstellingen te behalen.

### DDoS-aanvallen gericht op Belgische websites door NoName-groep

Op 4 november 2025 heeft de hacker-groep NoName bekendgemaakt meerdere websites in België te hebben aangevallen. De getroffen doelwitten bevinden zich in verschillende regio's van het land, waaronder de provincie Limburg, de gemeentelijke administratie van Waimes, de provincie Brabant Wallon, de provincie Luxemburg en de stad Antwerpen. Deze DDoS-aanvallen (Distributed Denial of Service) verstoren de toegang tot de betreffende websites door ze te overspoelen met verkeer, wat kan leiden tot langdurige onbeschikbaarheid van online diensten. Het is nog onduidelijk welke verdere impact deze aanvallen hebben, maar het lijkt een gecoördineerde cyberoperatie te zijn gericht op publieke en administratieve instellingen. De Belgische autoriteiten zijn op de hoogte van de situatie en onderzoeken de zaak verder.

- Province of Limburg
- Waimes Municipal Administration
- Province du Brabant wallon
- Province de Luxembourg
- City of Antwerp

### Nieuwe hacktivistenalliantie gevormd door NoName en Heaven of the Slavs

De hacktivistische groepen NoName en Heaven of the Slavs hebben een officiële alliantie aangekondigd. Deze samenwerking tussen de twee groepen benadrukt hun gezamenlijke focus op het uitvoeren van cyberaanvallen. De groepen, die al verantwoordelijk zijn voor verschillende gerichte aanvallen, bundelen hun middelen en expertise in een poging om hun impact te vergroten. De alliantie kan mogelijk leiden tot grotere en complexere aanvallen, met als doel om politieke en maatschappelijke systemen te verstoren. De samenwerking is een voorbeeld van



## Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)

hoe hacktivistische groepen zich steeds verder ontwikkelen, met nieuwe allianties die de dreiging in cyberspace versterken.

Kritieke kwetsbaarheid in Post SMTP plug-in bedreigt duizenden WordPress-sites

Er is een kritieke kwetsbaarheid ontdekt in de Post SMTP-plug-in voor WordPress, die actief wordt misbruikt door aanvallers. De kwetsbaarheid, aangeduid als CVE-2025-11833, stelt ongeauthenticeerde aanvallers in staat om op afstand e-mails te lezen die naar de logbestanden van websites worden gestuurd, inclusief wachtwoordreset-links. Dit stelt hen in staat om wachtwoorden van beheerdersaccounts te resetten en de volledige controle over de website te verkrijgen. De impact van de kwetsbaarheid is beoordeeld met een 9,8 op een schaal van 10. Hoewel er sinds 29 oktober een beveiligingsupdate beschikbaar is, hebben nog ongeveer 200.000 websites de update niet geïnstalleerd. Sinds 1 november zijn aanvallers begonnen met het misbruiken van de kwetsbaarheid. De plug-in wordt door meer dan 400.000 websites gebruikt.

Google patcht kritiek lek in Android dat aanvalscodes mogelijk maakt

Google heeft beveiligingsupdates uitgebracht voor Android om een ernstige kwetsbaarheid te verhelpen, aangeduid als CVE-2025-48593. Het probleem bevindt zich in het systeemcomponent van Android 13, 14, 15 en 16, en stelt aanvallers in staat om op afstand code uit te voeren zonder extra rechten. Deze kwetsbaarheid kan ernstige gevolgen hebben voor Android-apparaten, omdat kwaadwillenden toegang kunnen krijgen zonder de toestemming van de gebruiker. Naast deze kritieke kwetsbaarheid heeft Google ook een probleem gepatcht waarbij een aanvaller met toegang tot een apparaat de rechten kan verhogen. De updates zijn beschikbaar voor toestellen met Android 13 tot 16, maar niet alle apparaten zullen automatisch de updates ontvangen, vooral niet voor oudere modellen of toestellen die geen ondersteuning meer krijgen. Google heeft fabrikanten tijdig ingelicht zodat zij de benodigde updates konden ontwikkelen.

Apple brengt beveiligingsupdates uit voor privacylekken in iOS en macOS

Apple heeft beveiligingsupdates uitgebracht voor iOS en macOS, gericht op meerdere kwetsbaarheden, voornamelijk privacygerelateerd. Deze lekken stellen



## Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)

apps in staat om toegang te krijgen tot gevoelige gebruikersdata, al geeft Apple geen details over de aard van deze gegevens. Ook zijn er kwetsbaarheden verholpen die apps in staat stelden om gebruikers te volgen of te fingerprinten. Een ander probleem betrof het camerabeeld, dat kon worden bekeken voordat de camerapermissie werd goedgekeurd. Fysieke toegang tot een apparaat stelde aanvallers in staat om "restricted content" of "sensitive user information" op een vergrendeld scherm te bekijken. Daarnaast werd een lek in de Mail-app gepatcht, waardoor trackingpixels konden worden geladen, zelfs als deze optie was uitgeschakeld. Safari had een kwetsbaarheid waardoor apps privacy-instellingen konden omzeilen. Gebruikers van macOS en iOS kunnen de updates installeren via de laatste versie van hun besturingssysteem.

Kritieke kwetsbaarheid in JobMonster WordPress-thema uitgenut door hackers

Hackers maken gebruik van een kritieke kwetsbaarheid in het JobMonster WordPress-thema, die hen in staat stelt beheerdersaccounts over te nemen. De kwetsbaarheid, geïdentificeerd als CVE-2025-5397, wordt veroorzaakt doordat de functie `check_login()` de identiteit van gebruikers niet goed verifieert, wat aanvallers in staat stelt de standaard authenticatie te omzeilen. Dit probleem beïnvloedt alle versies van het thema tot versie 4.8.1 en heeft een ernstige score van 9.8 op de CVSS-schaal. Het lek heeft alleen impact als de social login-functie is ingeschakeld, wat de mogelijkheid biedt voor aanvallers om zich voor te doen als beheerders zonder geldige inloggegevens. De kwetsbaarheid is inmiddels opgelost in versie 4.8.2 van JobMonster, en het wordt aanbevolen deze versie onmiddellijk te installeren. Gebruikers wordt ook aangeraden social login uit te schakelen, tweefactorauthenticatie in te schakelen en toeganglogs te controleren op verdachte activiteiten.

Kritieke React Native CLI-kwetsbaarheid stelt miljoenen ontwikkelaars bloot aan externe aanvallen

Een ernstige kwetsbaarheid in het populaire React Native CLI-pakket stelde aanvallers in staat om op afstand besturingssysteemcommando's uit te voeren. De kwetsbaarheid, aangeduid als CVE-2025-11953, heeft een CVSS-score van 9,8, wat wijst op de hoge ernst ervan. De kwetsbaarheid deed zich voor in het `@react-native-community/cli`-pakket, dat wereldwijd miljoenen ontwikkelaars ondersteunt bij het



## Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)

bouwen van mobiele applicaties. De kwetsbaarheid ontstaat doordat de Metro-ontwikkelservers verbinding maakt met externe interfaces in plaats van localhost en een endpoint "/open-url" heeft dat kwetsbaar is voor besturingssysteemcommando-injectie. Aanvallers konden via een speciaal gemaakte POST-aanvraag willekeurige commando's uitvoeren. Meta heeft inmiddels een patch uitgebracht in versie 20.0.0. Ontwikkelaars die gebruikmaken van andere frameworks die niet de Metro-server gebruiken, werden niet getroffen door deze kwetsbaarheid.

### AMD Zen 5-processors RDSEED-kwetsbaarheid compromitteert integriteit van willekeurige getallen

AMD heeft een kritieke kwetsbaarheid onthuld in de Zen 5-processors, die de betrouwbaarheid van willekeurige getallengeneratie aantast, een essentieel onderdeel van moderne beveiliging. De fout, geïdentificeerd als CVE-2025-62626, betreft de RDSEED-instructie die gebruikt wordt om cryptografisch veilige willekeurige getallen te genereren, noodzakelijk voor encryptie en authenticatie. Door een defect in de implementatie van de RDSEED-instructie kan de instructie een waarde van nul teruggeven, terwijl ten onrechte succes wordt aangegeven via de carry flag. Dit creëert een gevaarlijke situatie waarin software denkt dat een geldig willekeurig getal is ontvangen, terwijl het in werkelijkheid een voorspelbare nulwaarde betreft. De kwetsbaarheid is van invloed op zowel 16-bits als 32-bits versies van de RDSEED-instructie, maar de 64-bits versie blijft onaantast. AMD werkt aan microcode-updates en firmware-aanpassingen, die naar verwachting in november 2025 beschikbaar komen. Tot de updates uitgerold zijn, wordt aangeraden software-aanpassingen door te voeren.

### Meer dan 40 miljoen downloads van kwaadaardige Android-apps op Google Play

Meer dan 40 miljoen downloads van kwaadaardige Android-apps op Google Play werden geregistreerd tussen juni 2024 en mei 2025, volgens een rapport van Zscaler. Het aantal malware-aanvallen op mobiele apparaten groeide in hetzelfde periode met 67% in vergelijking met het jaar ervoor. De meeste dreigingen kwamen van spyware en banktrojans. Cybercriminelen verschuiven van traditionele creditcardfraude naar aanvallen op mobiele betalingen door middel van phishing, smishing, SIM-swapping en betalingsfraude. Bovendien nam het aantal gedetecteerde adware-aanvallen toe, wat nu 69% van alle Android-malware vormt,



## Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)

bijna dubbel zoveel als vorig jaar. De meest getroffen landen waren India, de Verenigde Staten en Canada, maar er werden ook enorme stijgingen waargenomen in Italië en Israël. Drie belangrijke malwarefamilies werden opgemerkt: Anatsa, Android Void (Vo1d) en Xnotice. Zscaler raadt aan om regelmatig beveiligingsupdates te installeren en alleen apps van vertrouwde ontwikkelaars te downloaden.

Microsoft Teams bugs laten aanvallers zich voordoen als collega's en berichten aanpassen zonder dat dit zichtbaar is

Onderzoekers van Check Point hebben vier ernstige kwetsbaarheden in Microsoft Teams onthuld die aanvallers de mogelijkheid bieden om gesprekken te manipuleren, collega's te imiteren en meldingen te misleiden. Deze kwetsbaarheden stellen aanvallers in staat om berichtinhoud te wijzigen zonder dat het label "Bewerkt" wordt weergegeven, en om de afzender van berichten te veranderen, zodat deze afkomstig lijkt van een vertrouwde bron, zoals hooggeplaatste medewerkers. Dit vergemakkelijkt social engineering-aanvallen, waarbij slachtoffers worden misleid om schadelijke berichten te openen of gevoelige informatie te delen. Bovendien kunnen aanvallers de weergavenaam in privégesprekken en tijdens oproepen aanpassen, wat het mogelijk maakt om de identiteit van bellers te vervalsen. Microsoft heeft na de verantwoorde bekendmaking van de kwetsbaarheden in maart 2024 patches uitgerold, maar de gebreken blijven een risico vormen voor de beveiliging van organisaties.

Cybercriminaliteitssamenvoeging: Scattered Spider, LAPSUS\$ en ShinyHunters bundelen krachten

Een fusie van drie prominente cybercriminelen, Scattered Spider, LAPSUS\$ en ShinyHunters, heeft geleid tot de oprichting van een nieuwe cybercrime-groep genaamd "Scattered LAPSUS\$ Hunters" (SLH). Sinds augustus 2025 zijn ze actief op Telegram, waar ze tot wel 16 kanalen hebben gecreëerd en opnieuw gestart, telkens onder variaties van hun naam. Deze kanalen dienen zowel als communicatieplatform als als middel voor het uitvoeren van extortion-aanvallen. De groep biedt 'extortion-as-a-service' aan, waardoor andere groepen zich kunnen aansluiten en profiteren van hun merk en naamsbekendheid om losgeld te eisen. De SLH-groep is verbonden met andere cybercriminaliteitclusters zoals





## **Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)**

CryptoChameleon en Crimson Collective. Hun tactieken omvatten geavanceerde sociale engineering, zoals spear-phishing en vishing, en het gebruik van kwetsbare toegangspunten voor datadiefstal en afpersing. Ze tonen een mix van financieel gemotiveerde criminaliteit en hacktivistische motieven.

### **Hackers Scannen Internet om XWiki RCE-kwetsbaarheid te Exploiteren**

Hackers richten zich actief op een kritieke kwetsbaarheid in XWiki's SolrSearch-component, die een remote code execution (RCE) aanval mogelijk maakt. Deze kwetsbaarheid stelt aanvallers in staat om willekeurige commando's uit te voeren op kwetsbare systemen, wat een groot beveiligingsrisico vormt voor organisaties die deze open-source wiki-software gebruiken. De kwetsbaarheid kan worden misbruikt met minimale gastrechten, wat het voor vrijwel elke gebruiker met basis toegang mogelijk maakt om het systeem te compromitteren. XWiki bracht in februari een beveiligingsupdate uit, maar de exploitatie van de kwetsbaarheid kwam pas onlangs op gang, nadat proof-of-concept code eerder werd vrijgegeven. De aanvallers sturen speciaal geprepareerde GET-aanvragen naar de kwetsbare XWiki-eindpunten, waarbij Groovy-scriptcommando's worden ingesloten om shellcommando's uit te voeren. Organisaties wordt aangeraden om de patch te implementeren en verdachte verzoeken naar SolrSearch te monitoren om aanvallen te voorkomen.

### **Man krijgt jaren cel en contactverbod voor langdurige stalking van vrouw**

Een man is veroordeeld tot drie jaar gevangenisstraf voor het langdurig stalken van een vrouw, zowel online als offline. Hij volgde haar, fotografeerde haar woning, stuurde honderden berichten met bedreigende en seksuele inhoud, en maakte nepaccounts met haar naam en foto's. Ook plaatste hij seksadvertenties waardoor vreemden bij haar aan de deur verschenen. Ondanks een eerder opgelegd contactverbod ging hij door met zijn gedrag. Volgens de rechter veroorzaakte hij ernstige angst en aantasting van de privacy van het slachtoffer. De man kreeg een celstraf van drie jaar, waarvan 510 dagen voorwaardelijk, een proeftijd van drie jaar en een direct uitvoerbaar contactverbod van vijf jaar. Daarnaast moet hij het slachtoffer 12.500 euro schadevergoeding betalen en worden zijn telefoons, die hij gebruikte voor de stalking, in beslag genomen.





## Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)

Eurojust pakt cryptocurrency-oplichting aan van 600 miljoen euro

Eurojust heeft een gecoördineerde actie geleid tegen een netwerk van cryptocurrency-oplichters die slachtoffers voor meer dan 600 miljoen euro hadden opgelicht. Negen verdachten werden gearresteerd in drie landen, waaronder Spanje, Cyprus en Duitsland. Het netwerk creëerde valse cryptocurrency-investeringsplatformen die hoge rendementen beloofden. Slachtoffers werden via social media, koude telefoontjes en nepnieuwsartikelen misleid om crypto's over te maken, die vervolgens niet konden worden teruggehaald. De criminele opbrengsten werden gewassen met behulp van blockchain-technologie. De gezamenlijke operatie werd uitgevoerd door autoriteiten uit Frankrijk, België, Cyprus, Duitsland en Spanje, onder leiding van Eurojust, waarbij aanzienlijke geldbedragen in bankrekeningen en cryptocurrencies in beslag werden genomen.

Drie nepagenten op heterdaad aangehouden in Kerkdriel

Op zondagavond 2 november heeft de politie drie verdachten gearresteerd die zich voordeden als agenten om een vrouw op te lichten. De vrouw, 84 jaar oud, ontving een telefoontje van een persoon die zich als agent uitgaf en vertelde dat er criminelen in de wijk actief waren. Verdacht van de situatie, belde de vrouw 112 om de echtheid van het telefoontje te verifiëren. De politie ging ter plaatse en observeerde de woning. Kort daarna stond een van de nepagenten voor de deur, waarna hij werd gearresteerd. Twee andere verdachten, een minderjarige en twee mannen uit Amsterdam en Purmerend, werden eveneens aangehouden in een nabijgelegen voertuig. De verdachten worden verhoord en zitten vast. Het onderzoek naar hun betrokkenheid gaat door.

AP: kredietverstrekker mag niet zomaar complete bankafschriften eisen

De Autoriteit Persoonsgegevens (AP) heeft aangegeven dat kredietverstrekkers niet zonder meer volledige bankafschriften mogen eisen van consumenten. Dit vormt volgens de AP een te grote inbreuk op de privacy. Het advies van de toezichthouder komt na de beoordeling van een wetsvoorstel over leningen, waarin onder andere strengere regels worden voorgesteld voor kleine leningen en 'koop nu, betaal later'-diensten. De AP vindt dat consumenten meer controle moeten krijgen over de gegevens die kredietverstrekkers mogen inzien. Het verplicht verstrekken van bankafschriften kan namelijk een volledig beeld geven van iemands persoonlijke



## Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)

leven en voorkeuren, wat te ver gaat. De toezichthouder pleit voor regels die kredietverstrekkers in staat stellen om kredietwaardigheid te beoordelen met minder ingrijpende gegevens.

### Brede steun voor Europese bewaarplicht onder EU-landen

Minister Van Oosten van Justitie en Veiligheid meldt dat er brede steun is voor de invoering van een Europese bewaarplicht. De discussie hierover werd recent gevoerd tijdens de Raad Justitie en Binnenlandse Zaken (JBZ). De voorgestelde bewaarplicht volgt op de ongeldigverklaring van de dataretentie-richtlijn door het Europese Hof van Justitie elf jaar geleden. De Europese Commissie onderzoekt of een nieuwe, geharmoniseerde bewaarplicht mogelijk is, mede naar aanleiding van een advies van de High-Level Group 'Going Dark'. Deze groep pleit voor toegang van opsporingsdiensten tot versleutelde data en de invoering van encryptie-backdoors. Het demissionaire kabinet steunt het plan en wijst op de rechtsongelijkheid die door de huidige versnipperde wetgeving is ontstaan. Er is brede steun onder lidstaten, maar er is ook felle kritiek van Europese burgers via een online impactbeoordeling. De Commissie verwacht begin 2026 de resultaten van de impactbeoordeling te presenteren.

### Xi Jinping maakt grap over achterdeurtjes in Xiaomi-smartphones

Tijdens een ontmoeting met de Zuid-Koreaanse president Lee Jae-myung op zaterdag maakte de Chinese president Xi Jinping een grap over Xiaomi-smartphones, waarbij hij suggereerde dat de toestellen mogelijk achterdeurtjes bevatten. De opmerking werd gemaakt nadat Lee, die een Go-bord cadeau gaf aan Xi, vroeg of de smartphones die aan hem werden gegeven, veilig waren. Xi's reactie was dat Lee moest controleren op achterdeurtjes, wat resulteerde in een lach van beide leiders. De opmerking is opmerkelijk, aangezien achterdeurtjes in technologie vaak worden geassocieerd met inbreuk op de privacy en staatsbespionage, wat een veelvoorkomende zorg is rondom Chinese technologiebedrijven zoals Huawei en ZTE. De opmerking van Xi lijkt te spelen met de bezorgdheden die bestaan over de beveiliging van Chinese apparaten, terwijl Xiaomi als een belangrijke speler op de mondiale smartphonemarkt wordt gezien.



## **Cybercrimeinfo - cyberdreigingsanalyse (Openbare versie)**

Experts waarschuwen voor tekortkomingen in AI-testen voor veiligheid en effectiviteit

Deskundigen hebben honderden tests beoordeeld die AI-modellen evalueren op veiligheid en effectiviteit. Uit hun onderzoek blijkt dat vrijwel alle tests zwakke punten bevatten die de geldigheid van de claims over deze AI-modellen kunnen ondermijnen. De onderzoekers, waaronder experts van het AI Security Institute en universiteiten zoals Oxford, Berkeley en Stanford, ontdekten dat veel gebruikte benchmarks – belangrijke maatstaven voor AI-modellen – op verschillende gebieden tekortschieten. Slechts 16% van de benchmarks gebruikt statistische tests die de accuraatheid van de resultaten ondersteunen. Dit probleem heeft ernstige implicaties voor de betrouwbaarheid van AI-modellen die door technologiebedrijven worden gepromoot. Experts roepen op tot de ontwikkeling van uniforme normen voor het testen van AI-modellen. Het onderzoek benadrukt ook de gevaren die gepaard gaan met het gebruik van AI, zoals valse beschuldigingen of zelfs tragische incidenten, wat de noodzaak van robuuste normen verder onderstreept.