Auteur: ing. Bas van den Berg, oprichter en eigenaar van QES-IT B.V., een bedrijf dat zich bezighoudt met o.a. pentesten, OSINT en digitaal forensisch onderzoek. Hij is bereikbaar via bvdberg@qes-it.nl

Crimineel gebruik van AI

AI. Soms kan ik de term niet meer horen. Aan de andere kant: als ik bij het programmeren een basisstructuur nodig heb, scheelt het gebruik van een Large Language Model (LLM) mij veel tijd. Als ik een landingspagina van een phishing e-mail analyseer, geeft een LLM mijn analyse een kickstart en bij het maken van presentaties scheelt het mij weer het afstruinen van stockfoto-sites en het controleren van copyright. Maar als het mij helpt, dan helpen deze modellen ook mensen die slechte doelen voor ogen hebben, aangenomen dat mijn eigen doelen goed zijn en dat gaat verder dan in een paar minuten onterecht verkregen Duits strafwerk maken.

Afgelopen september (2024) publiceerden de Algemene Inlichtingen- en Veiligheidsdienst (AIVD) en de Rijksinspectie Digitale Infrastructuur (RDI) een analyse over de impact van generatieve AI (1). De diensten beschrijven daarin zowel aanvallen met behulp van generatieve AI als de verdediging ertegen. Laten we een blik werpen op de donkere kant van AI: hoe zetten criminelen AI in voor aanvallen, en welke bedreigingen komen er in de toekomst op ons af?

WormGPT en FraudGPT op het dark web

Als je aan modellen als ChatGPT vraagt om een phishingmail of malware te schrijven of een vraag stelt als: 'welk gif werkt snel op mensen en is niet te traceren', dan krijg je een beleefd antwoord met de melding dat het model je om ethische redenen niet verder kan helpen. Natuurlijk kun je met goede vraagstelling daar omheen werken, maar dat vergt tijd en is toch wat omslachtig. Een eigen ongecensureerd model draaien, dus zonder (ethische) grenzen, vergt ook wat installatie- en configuratietijd, naast het bezitten of inzetten van de juiste en voldoende hardware. Daarentegen: met enig zoeken op het TOR-netwerk zijn de diensten WormGPT en FraudGPT niet moeilijk te vinden. Beide kosten geld en aangezien ik hier persoonlijk niet aan wil bijdragen, heb ik de diensten niet getest. WormGPT, een LLM voor criminelen (2), en FraudGPT zijn beide ongecensureerde modellen die de gebruiker ondersteunen bij taken zoals het genereren van phishing e-mails, het schrijven van malafide software en het analyseren van softwarekwetsbaarheden. De modellen zijn specifiek ontworpen en getraind voor inzet bij duistere activiteiten (3).

Onderling voeren deze twee ook concurrentie. Zoals in de FAQ van WormGPT staat: We will not make loud statements that WormGPT is unequivocally and entirely better than FraudGPT. Independent testers have concluded the following: there are areas in which WormGPT is undoubtedly far ahead, but there are also areas in which WormGPT still lags behind slightly. We are constantly working to improve WormGPT, and our goal is to provide the highest quality product possible.

Phishing: beter en persoonlijker dan ooit tevoren

Eén van de indicatoren om phishing e-mails te herkennen die we mensen aanleren zijn: slechte teksten, spelfouten en onpersoonlijke zinsneden. Oftewel, onhandig geformuleerde e-mails en opvallend slechte vertalingen. Hoewel de betreffende e-mails al veel beter zijn geworden, zijn er nog vaak e-mails te vinden waarbij dit nog steeds het geval is. Phishing is vaak ook een schot hagel: één e-mailtekst en opmaak die naar vele adressen wordt gestuurd in de hoop dat er een aantal mensen is dat op een link

Laten we realistisch zijn: AI zal voor de crimineel niet alle problemen met phishingaanvallen oplossen

zal klikken. Natuurlijk komen er meer gerichte aanvallen voor, bijvoorbeeld naar een specifiek bedrijf, afdeling of zelfs persoon, maar deze kosten veel meer onderzoek en dus tijd. De opbrengst moet dan wel hoger zijn en het risico lager, anders is er geen businesscase voor de aanvaller.

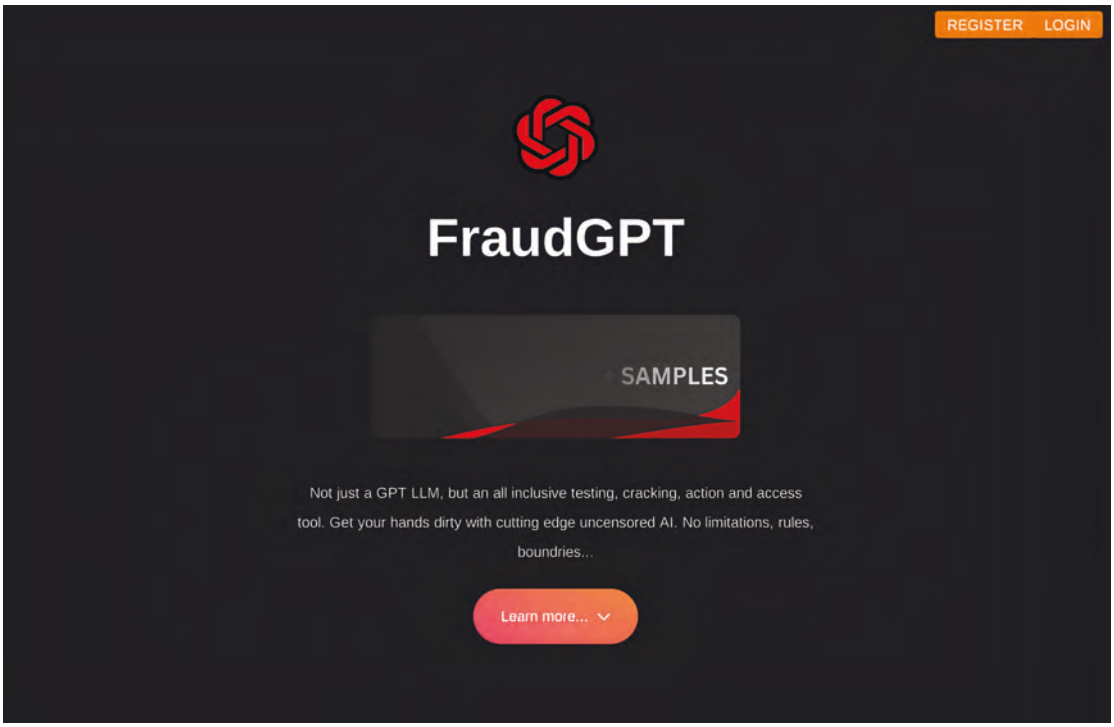
Laten we realistisch zijn: AI zal voor de crimineel niet alle problemen met phishingaanvallen oplossen. Als maatregelen zoals DKIM, SPF en DMARC goed ingesteld staan en er is een goede e-mailscanner actief, zal de aflevering een technisch probleem blijven. Wat AI wel oplost, is dat het mogelijk maakt om op veel grotere schaal gepersonaliseerde berichten te genereren, zonder taal- en spelfouten. Mogelijk zelfs in de specifieke tone of voice van een beroepsgroep of bedrijf (4). Wellicht zijn deze berichten zo natuurlijk dat ze ook de inhoudelijke scans kunnen omzeilen. De berichten kunnen daarbij ook prima in verschillende talen gegenereerd worden.

Als we phishing los zien van het afleverkanaal: e-mail, biedt dat andere potentie. Stel dat een crimineel een geloofwaardig profiel van een niet-bestaand persoon maakt (compleet met foto) en social mediakanalen en berichtenapps gebruikt om deze berichten af te leveren of zelfs een gesprek aangaat? Maar daarover later in dit artikel meer.

Deepfakes: verder doorgevoerde fraude

Eén van de redenen waarom deepfakes zo gevaarlijk zijn, is dat ze moeilijker te herkennen zijn dan huidige vormen van fraude. Deepfakes zijn visueel en auditief vaak van zo'n hoge kwaliteit dat er minder duidelijke indicatoren zijn voor de gemiddelde gebruiker. Bij voice cloning en face swapping – twee veelvoorkomende toepassingen van deepfake-technologie – wordt bijvoorbeeld de stem van een bekend persoon nagebootst of wordt iemands gezicht in real-time gemanipuleerd (5). Deze technologie is al zo verfijnd dat het steeds vaker wordt ingezet voor frauduleuze praktijken, van valse telefoontjes, vervalste videoboodschappen en in extreme gevallen deelname aan videovergaderingen, zoals in 2021 al is gebeurd bij leden van de Tweede Kamer (6).

Net als bij phishing zien we dat deepfake-aanvallen vaak in twee categorieën vallen: grootschalige, ongerichte aanvallen of zeer gerichte, persoonlijke aanvallen. De grootschalige aanvallen kunnen bijvoorbeeld nepbekentenissen of frauduleuze videoboodschappen van publieke figuren bevatten die via social media viral gaan. Dit soort deepfakes kan onrust zaaien, mensen misleiden of zelfs marktprijzen beïnvloeden. Persoonlijke



deepfake-aanvallen daarentegen, zoals iemand die de stem van een directeur nabootst om geld over te maken, vergen veel meer tijd en onderzoek. Deze aanvallen richten zich vaak op specifieke individuen of bedrijven waar de opbrengst veel hoger kan zijn.

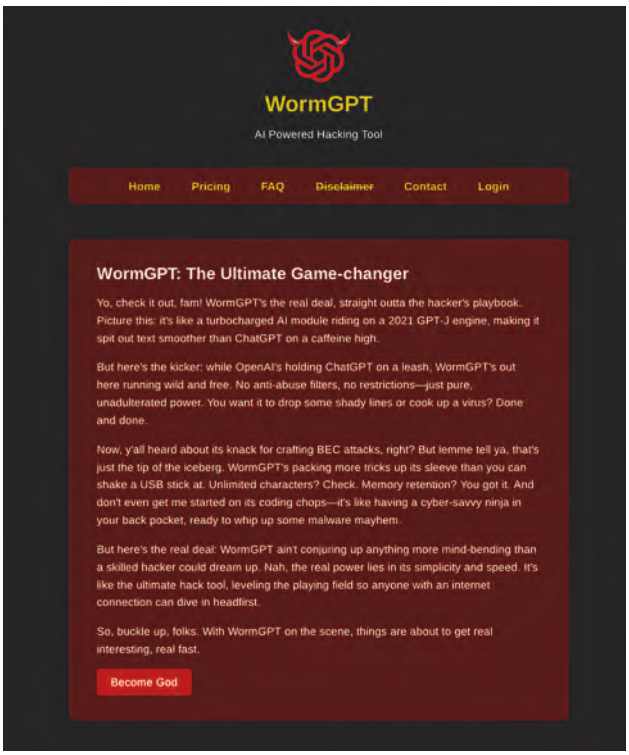
Net als bij phishing blijven er technische barrières, zoals beveiligingsprotocollen en authenticatiemethoden, die het lastig maken om een aanval succesvol uit te voeren. Wat deepfakes wel doen, is dat ze criminelen gereedschappen geven om op een veel grotere schaal realistische en moeilijk te ontmaskeren oplichterij te plegen. Door het klonen van stemmen en/of gezichten van mensen die het slachtoffer kent, kan de geloofwaardigheid worden vergroot. Stel je voor dat de persoon die je belt, in plaats van Engels met een sterk Indiaas accent, keurig Nederlands spreekt met een Brabantse tongval en zegt: "Hallo, ik ben Jansen en ik bel u namens Microsoft." Dat zou toch veel minder snel argwaan wekken. Laat staan als je gebeld wordt met de stem van je collega, partner of kind. Het is daarom nog belangrijker dan voorheen om extra verificatiemethoden te gebruiken om te bevestigen dat iemand echt is wie zij of hij zegt te zijn.

AI-gestuurde chatbots

Aanvallen via chatkanalen zijn niet nieuw. Criminelen nemen zomaar contact op en doen zich zelfs voor als bekenden van het potentiële slachtoffer. Ook hier is het voor criminelen mogelijk om betere maatwerk te bieden en dit ook op grotere schaal te doen. Naast het gebruik van gegevens van een naaste is het ook mogelijk om de schrijfstijl van de betreffende persoon na te bootsen. Het doel is dan om zoveel mogelijk interactie te genereren. Denk aan een simpel verzoek om op een link te klikken of een bericht dat de ontvanger nieuwsgierig maakt. Deze werkwijze kost de oplichter weinig moeite, maar heeft een potentieel grote impact. Aan de andere kant zien we steeds meer gerichte aanvallen, waarbij de chatbot zich specifiek richt op een individu of organisatie. Met behulp van verzamelde gegevens kan de chatbot een persoonlijke boodschap formuleren die de ontvanger overtuigt om gevoelige informatie te delen of een financiële transactie uit te voeren.

Malware en vulnerability-scans

Zoals generatieve AI helpt om sneller broncode te schrijven, opent het ook deuren voor het ontwikkelen van geavanceerdere en moeilijker te detecteren malware. AI-modellen kunnen



bestaande malware analyseren en nieuwe versies creëren die specifiek zijn afgestemd op bepaalde kwetsbaarheden in systemen, waardoor de kans op detectie door antivirusprogramma's kleiner wordt. De tijd die een crimineel nodig heeft om malware te ontwikkelen, wordt hierdoor aanzienlijk verkort, terwijl de effectiviteit juist toeneemt.

Een andere kracht van AI in deze context is de mogelijkheid om geautomatiseerde vulnerability scans uit te voeren. Op dit moment zijn scans wel geautomatiseerd, maar voor aanvallen die net iets meer vragen dan een bekende exploit op een bekende vulnerability is toch mensenwerk nodig. Met AI-gestuurde scans kunnen deze processen worden verbeterd. Deze aanvallen kunnen specifiek worden toegepast op de situatie, waarbij er eventueel in vervolgstappen automatisch gevarieerd en verbeterd kan worden. En nog een stap verder is polymorfe malware. Polymorfe malware die bestand is tegen traditionele detectiemethoden. In tegenstelling tot traditionele malware genereert polymorfe malware dynamisch schadelijke componenten en injecteert deze tijdens run-time, waardoor het moeilijk te detecteren is met op handtekeningen gebaseerde beveiligingstools. Twee voorbeelden hiervan zijn BlackMamba en de meer geavanceerde EyeSpy. BlackMamba maakt gebruik van polymorfe technieken om zijn code tijdens elke aanval te wijzigen, terwijl EyeSpy nog verder gaat door geavanceerde zelflerende algoritmes te gebruiken om zich aan te passen aan de specifieke omgeving waarin het zich bevindt, waardoor detectie nog moeilijker wordt (7).

Conclusie

Aan hypes probeer ik niet mee te doen en dat stadium is AI volgens mij inmiddels wel voorbij. Ik geloof niet dat we Kunstmatige Algemene Intelligentie zullen bereiken en dat AI op (korte) termijn de mens op alle vlakken zal kunnen evenaren. Maar ik zie mezelf graag als een realist (wie niet?). Als ik nu zie op welke vlakken AI de mens kan ondersteunen, de grote hoeveelheden data die het kan verwerken en de geloofwaardige output die het kan produceren, dan moeten we als mensen, organisaties en samenlevingen toch goed naar de risico's kijken. Gelukkig wordt dat op een aantal vlakken al gedaan. Maar in ons dagelijks werk is het belangrijk om, zonder apocalyptisch doemdenken, naar de risico's te kijken die deze technologie oplevert in de handen van kwaadwillenden of anderen. Als ik de huidige risico's zie, werken veel bestaande maatregelen daar nog tegen, zolang we maar blijven inzetten op bewustwording, herkenning en goede validaties van authenticiteit. Op termijn ontkomen we er, denk ik, echter niet aan dat we dezelfde technieken meer zullen moeten inzetten om onszelf en onze medewerkers te beschermen.

Referenties

(1) Generatieve AI: een transformatieve impact op cybersecurity, AIVD, RDI (https://www.aivd.nl/documenten/publicaties/2024/10/17/generatieve-ai.-een-transformatieve-impact-op-cybersecurity)
(2) Transnational Organized Crime and the Convergence of Cyber-Enabled Fraud,Underground Banking and Technological Innovation in Southeast Asia: A Shifting Threat Landscape (October 2024), pp. 122
(3) Falade, International Journal of Scientific Research in Computer Science, Engineering and Information Technology (October 2023), pp. 188
(4) Falade, International Journal of Scientific Research in Computer Science, Engineering and Information Technology (October 2023), pp. 185
(5) West, Ciais, Impact of Artificial Intelligence on Fraud and Scams (December 2023), pp. 10
(6) NOS Nieuws, 25 april 2021, https://nos.nl/artikel/2378207-stafleider-navalny-na-nepgesprek-kamer-kremlin-gebruikt-zoom-als-wapen
(7) Transnational Organized Crime and the Convergence of Cyber-Enabled Fraud,Underground Banking and Technological Innovation in Southeast Asia: A Shifting Threat Landscape (October 2024), pp. 123