Bezoekadres
Turfmarkt 147
2511 DP Den Haag

Postadres
Postbus 20011
2500 EA Den Haag

I www.cybersecurityraad.nl
T 070 751 5333 (secretariaat)
E info@cybersecurityraad.nl

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties
T.a.v. Mw. Drs. A.C. van Huffelen
Postbus 20011
2500 EA Den Haag

Datum
15 december 2023

Onderwerp
Informerende brief van de Cyber Security Raad over (generatieve) AI en cybersecurity

Excellentie,

Tijdens uw bezoek aan de Cyber Security Raad (hierna de raad) van 15 juni jl. spraken wij over de risico´s rond het gebruik van artificiële intelligentie (AI). Vooral de implicaties van generatieve AI vragen speciale aandacht. U heeft de raad uitgenodigd om aanvullende informatie te geven over de risico's die AI in brede zin, en generatieve AI in het bijzonder, met zich meebrengen.

Met name generatieve AI kan de mens steeds meer werk uit handen nemen. Dit geldt ook in de context van cybersecurity: zowel aan de aanvallende als de verdedigende kant zijn er verschillende grootschalige toepassingsmogelijkheden. Dit brengt echter wel ernstige beveiligingsrisico's met zich mee en daarom zijn regulering en blijvende aandacht voor besluitvorming door mensen noodzakelijk. De ontwikkelingen rond AI bieden ook volop cybersecurity kansen, zoals het geautomatiseerd verdedigen via autonome systemen of het detecteren van kwaadaardige software. In het vervolg van deze brief gaat de raad dieper in op die kansen en risico´s, en de maatschappelijke gevolgen hiervan.

Analyse

Door de kansen die generatieve AI biedt voor de samenleving, wordt de introductie ervan vaak neergezet als een *game changer*. Het vakgebied AI bestaat echter al tientallen jaren en generatieve AI bouwt voort op eerdere technologische innovaties. In tegenstelling tot andere vormen van AI (hierna 'klassieke' AI genoemd) maakt generatieve AI zelf inhoud. Met name dit aspect, in combinatie met het toegankelijk maken van de technologie voor grote groepen gebruikers, is in feite de *game changer* geweest. In bijlage 1 van deze brief vindt u een nadere toelichting en afbakening van de verschillende begrippen, volgens de meest gangbare definities op dit moment.

Door bovenstaande ontwikkelingen nemen ook de bedreigingen op het vlak van cybersecurity steeds verder toe. In bredere zin geldt dat ook publieke waarden, zoals (data)privacy, transparantie[1], copyright, gelijke behandeling en het functioneren van de democratische rechtsstaat onder druk kunnen komen te staan. Het gebruik van *responsible* AI moet dan ook voorop staan bij de toepassingen.

De snelle opkomst en implementatie van generatieve AI heeft al geleid tot een maatschappelijk debat over de risico's die dit fenomeen met zich meebrengt en er is onlangs – mede op uw initiatief - ook een Catshuissessie

[1] Meer specifiek gaat het bij AI gebruik om interpreteerbaarheid, voorkomen van vooringenomenheid, verklaarbaarheid en eerlijkheid.

met het kabinet aan gewijd. De aandacht richt zich vooral op de risico's voor genoemde publieke waarden, terwijl de implicaties voor cybersecurity maatschappelijk nog te weinig worden onderkend en begrepen.

De snelheid waarmee de huidige toepassingen van AI zich ontwikkelen kunnen het cybersecuritylandschap sterk veranderen, zowel op het gebied van kansen als risico's. Met generatieve AI kan bijvoorbeeld automatisch gescand worden op kwetsbaarheden in netwerken, waardoor eenvoudiger en op grotere schaal digitale aanvallen uit te voeren zijn. Onder andere bij vitale infrastructuren kan dit grote consequenties hebben, zeker als menselijk handelen onvoldoende snel is om de dreiging tegen te gaan. Dit kan leiden tot een transformatieve verandering, waarbij de verdedigende kant afhankelijk wordt van (generatieve) AI-toepassingen om de dreigingen het hoofd te bieden.

Specifieke kansen van AI in de context van cybersecurity

Door de opmars van generatieve AI is de bestaande belangstelling voor de mogelijkheden van AI op het vlak van cybersecurity verder in een stroomversnelling geraakt. De kansen komen bijvoorbeeld tot uitdrukking in de volgende specifieke toepassingsgebieden:

a) AI-toepassingen maken het mogelijk voor organisaties om automatisch aanvallen te detecteren via geconstateerde anomalieën in hun netwerk. Met generatieve AI kunnen daarna automatisch analyses gegenereerd worden en '*Indicators of Compromise'* verspreid, zodat op basis daarvan actie kan worden genomen.
b) AI kan analisten in *Security Operations Centers* (SOC's) ook ondersteunen bij het combineren van *security alerts* vanuit verschillende bronnen, om de detectie van aanvallen te bevorderen. Er kan daarbij nog een stap verder worden gegaan: met autonome systemen gebaseerd op AI is het mogelijk om zogenaamde ´slimme´ SOC´s op te zetten, met beperkte menselijke tussenkomst.
c) Met generatieve AI kunnen zogenoemde *advisories* ontwikkeld worden. Dit zijn generieke adviesberichten aan organisaties of organisatieonderdelen, waarin richtlijnen worden gegeven voor het mitigeren van recent geconstateerde, breed bekende kwetsbaarheden. Dit kan ook een taak van (slimme) SOC´s zijn, zoals genoemd onder b).
d) AI-toepassingen maken het mogelijk om kwaadaardige software automatisch te herkennen, varianten te correleren en van daaruit voorspellingen te doen over nieuwe vormen en variaties van dergelijke malware. Organisaties kunnen zich daar vervolgens proactief tegen wapenen.
e) Generatieve AI stelt via codegeneratoren ontwikkelaars veel sneller dan voorheen in staat om code te schrijven. Het is daarmee in potentie een waardevol hulpmiddel voor het ontwikkelen van veilige software.
f) Via AI-toepassingen is het mogelijk om automatisch kwetsbaarheden in allerlei software implementaties op te sporen en te herstellen.

Voorwaarden voor verantwoord gebruik

Bij alle genoemde kansen geldt het voordeel dat menselijke inspanningen beperkt kunnen blijven. Dit is een belangrijke pré in tijden van grote tekorten aan cybersecurityspecialisten, alhoewel het geen sluitende oplossing zal zijn. Drie kanttekeningen zijn daarbij op zijn plaats:

- Generatieve AI werkt (op dit moment en in de voorlopige toekomst) op basis van waarschijnlijkheden en heeft geen concept van waarheid of onwaarheid in zich[2]. Dit betekent dat generatieve AI het best tot zijn recht komt in handen van een expert die de output kan beoordelen en resultaten eventueel naar eigen

[2] Dit wordt in de volksmond ook wel het 'hallucineren' van AI genoemd en is integraal onderdeel van generatieve AI; foute uitkomsten zijn in deze situaties dus geen gevolg van softwarefouten.

inzicht kan aanpassen. Met de huidige stand van de techniek is dit een belangrijke voorwaarde voor verantwoord gebruik.
- Voor alle (defensieve) toepassingen geldt dat de kwaliteit van de onderliggende modellen, de data waarop wordt voortgebouwd én het beoogd gebruik daarvan essentieel is. Voor digitale aanvallers geldt dat zij deze remmende werking niet hebben en dus in principe sneller voordeel kunnen hebben van generatieve AI.
- Zowel binnen Rijksoverheid als het bedrijfsleven wordt toegewerkt naar verantwoorde inzet van toepassingen in specifieke omgevingen, zoals onder a), b) en c) genoemd. Ook daarbij gelden de voorwaarden uit het voorgaande.
  N.B. Een eventueel verbod op het gebruik van (generatieve) AI door rijksambtenaren, zoals recent in het nieuws kwam, staat haaks op deze ontwikkelingen. De raad is dan ook geen voorstander van een dergelijk algemeen verbod.

Specifieke risico's van AI in de context van cybersecurity

Hieronder volgt een opsomming van op dit moment bekende risico´s bij het gebruik van zowel generatieve als klassieke AI voor cybersecurity. Door recente ontwikkelingen op het gebied van met name generatieve AI worden de implicaties voor cybersecurity steeds duidelijker. Ook het CSBN 2023[3] gaat hier uitgebreid op in.

a) AI-toepassingen vergemakkelijken het uitvoeren van cyberaanvallen, doordat bestaande kwetsbaarheden automatisch op grote schaal kunnen worden uitgebuit. Dit geldt zowel voor nieuwe kwetsbaarheden die nog niet algemeen bekend zijn (zogenaamde *zero-days*) als bekende kwetsbaarheden die nog niet gerepareerd zijn.
b) De nieuwste generatieve AI-technologieën stellen cybercriminelen in staat om al bestaande modi operandi makkelijker, op grotere schaal en in hogere kwaliteit toe te passen. Denk aan het genereren van *deepfakes*, niet van echt te onderscheiden spam en *phishing* e-mails en het automatisch genereren van *malware* in allerlei variaties.
c) *Tooling* voor (generatieve) AI wordt vaak breed ingezet in allerlei organisaties en op allerlei netwerken. Die omgevingen zijn niet altijd (technisch) goed afgeschermd voor intern gebruik of gesegmenteerd van andere operationele netwerken. Dit vergroot het aanvalsoppervlak waardoor ook de risico´s op binnendringen, manipulatie en datalekken toenemen.
d) Het grootschalig gebruik van AI-software brengt ook algemene cybersecurity-risico's met zich mee. Juist door (te) snelle invoering is de AI-programmatuur zelf niet in alle gevallen uitvoerig getest en op (onbewuste) fouten gecontroleerd. Ook de authenticiteit en integriteit van datasets en trainingsprogramma's is niet altijd gewaarborgd. In lijn met de Cyber Resilience Act (CRA) en de Digital Services Act (DSA) die de EU oplegt, zijn hiervoor extra waarborgen nodig.

Naast bovenstaande risico´s, heeft een aantal van de eerdergenoemde kansen van (generatieve) AI voor cybersecurity ook een keerzijde. Deze leiden daarmee ook tot nieuwe risico´s:

e) Aansluitend op kansen a, b en c: De toepassing van AI in de keten van cybersecurityactiviteiten vergroot bij onzorgvuldig gebruik de kans op foute interpretaties (zowel *false positives* als *false negatives*), foute reacties, of ongewenste besluiten. Dit risico groeit als er zonder actieve menselijke tussenkomst te veel op deze toepassingen van AI vertrouwd gaat worden, terwijl de goede werking of adequate toepassing van de AI-software zelf niet gewaarborgd is.

[3] Cybersecuritybeeld Nederland 2023, Nationaal Coördinator Terrorismebestrijding en Veiligheid (NCTV), juni 2023
Cybersecuritybeeld Nederland 2023 | Publicatie | Nationaal Coördinator Terrorismebestrijding en Veiligheid (nctv.nl)

f) Aansluitend op kans e*:* Het wordt steeds duidelijker dat met (generatieve) AI gemaakte code momenteel nog veel kwetsbaarheden bevat. Indien dergelijke modules overkoepelend in veel softwaretoepassingen als standaard worden gebruikt, verhoogt dit het risico op massale uitbuiting. Dit werd bijvoorbeeld eerder geconstateerd in generiek toegepaste software modules, zoals Log4j. Hiervoor zijn extra cybersecuritycontrols noodzakelijk, zoals voortkomend uit de hiervoor genoemde CRA en DSA.

Internationale ontwikkelingen

Ook internationaal krijgen de kansen en risico's van (generatieve) AI ruim aandacht. Zo heeft de Secretaris-Generaal van de Verenigde Naties recent de nieuwe *High-Level Advisory Body on AI*[4] ingesteld. Deze groep experts uit alle maatschappelijke geledingen zal adviezen geven over internationale AI-*governance*. De gekozen interdisciplinaire insteek en *multi-stakeholder* aanpak is ook van toepassing op cybersecurity-vraagstukken in de context van AI. De raad beveelt aan ook in deze omgeving aandacht te laten besteden aan de cybersecurity-aspecten voor AI-toepassingen en vice versa.

Het is daarnaast belangrijk om ook de AI-aanpak van mondiale koplopers in ogenschouw te nemen, binnen de context van digitale weerbaarheid. Het betreft de vorming van strategisch AI-beleid en de noodzakelijke regulering die inmiddels al is gestart. Zo publiceerden de Amerikaanse en Britse overheid onlangs nieuwe *Guidelines for Secure AI System Development*[5]. De richtlijnen bieden suggesties en oplossingen waarmee datawetenschappers, ontwikkelaars, managers, besluitvormers en risico-eigenaren weloverwogen beslissingen kunnen nemen over het veilige ontwerp, de ontwikkeling, de implementatie en de werking van hun AI-systemen. Ook hebben veel landen inmiddels nationale regelgeving uitgebracht gericht op het waarborgen van privacy, zoals blijkt uit een recent overzicht[6].

Gelet op het grensoverschrijdende karakter van AI-technologie is ook verdere Europese samenwerking nodig. Zo deed de Europese Commissie recent al aanbevelingen aan de lidstaten over het uitvoeren van risicobeoordelingen op vier verschillende gebieden (waaronder het AI-domein), die cruciaal zijn voor de economische veiligheid van de EU[7]. Ook Nederland zou stevig moeten inzetten op een dergelijke risicobeoordeling voor AI.

Het EU-Agentschap voor cybersecurity ENISA bracht al diverse publicaties over AI uit. Deze richten zich zowel op het mitigeren van cybersecurityrisico´s als op de wijze waarop veilige en betrouwbare AI-toepassingen mogelijk zijn. Dit is een aanvulling op de Europese AI-act, die een aanzet geeft tot regulering van AI-toepassingen en de eerdergenoemde CRA, die diepgaande eisen stelt aan veilige digitale producten en diensten. ENISA zet daarbij in op *best practices* en standaardisatie: '*Multilayer Framework for Good Cybersecurity Practices for AI*'[8] *en 'Cybersecurity of AI and Standardisation*'[9].

Aanvullende vraagstukken

Er bestaan al langere tijd zorgen over de betrouwbaarheid van AI, ook in de sfeer van *responsible* AI. Het betreft dan bijvoorbeeld de eerdergenoemde authenticiteit en integriteit van de onderliggende data(sets) en algoritmen en ook de (technische) maatregelen die nodig zijn om die betrouwbaarheid te waarborgen.

[4] Zie: High-Level Advisory Body on Artificial Intelligence | Office of the Secretary-General's Envoy on Technology (un.org)
[5] Zie https://www.cisa.gov/news-events/alerts/2023/11/26/cisa-and-uk-ncsc-unveil-joint-guidelines-secure-ai-system-development
[6] Zie KMPG rapport: Privacy in the new world of AI (kpmg.com)
[7] Zie Commission recommends carrying out risk assessments on four critical technology areas: advanced semiconductors, artificial intelligence, quantum, biotechnologies | Shaping Europe's digital future (europa.eu)
[8] Zie Multilayer Framework for Good Cybersecurity Practices for AI — ENISA (europa.eu)
[9] Zie Cybersecurity of AI and Standardisation — ENISA (europa.eu)

In combinatie met de CRA voorziet de Europese AI-Act[10] in bepalingen die weliswaar betrekking hebben op cybersecurityvraagstukken rond AI, maar voornamelijk ingaan op de implicaties van bepaalde toepassingen. Er worden bijvoorbeeld regels gesteld rond transparantie, zoals het vooraf bekendmaken dat men met een *chatbot* te maken heeft of dat er sprake is van een *deepfake.* Ook het geven van informatie over de data waarmee het betreffende AI-model is getraind, behoort hiertoe. Juist bij generatieve AI is er extra aandacht hiervoor nodig, inclusief de cybersecurityrisico´s.

Een afgeleid aandachtspunt is de extra benodigde kennis over (generatieve) AI voor toepassing in het cyberdomein. Die kennis zou bij cybersecurityspecialisten (waaraan al een groot tekort is) opgebouwd moeten worden, maar ook *cross-over* gecreëerd moeten worden via AI-experts in algemene zin. Omgekeerd geldt dat er ook een groot gebrek is aan gekwalificeerde AI-experts, die daadwerkelijk begrip hebben van de inzet van (generatieve en klassieke) AI als hulpmiddel bij cybersecurity.

Naast plannen en ambities voor nieuwe AI-toepassingen en de noodzaak voor integrale beleidsontwikkeling, moet er ook voldoende oog zijn voor de uitvoering van de regelgeving. Een zorgpunt is de capaciteit bij toezichthouders, ook (of juist) op het gebied van AI in de context van cybersecurity. Hun werklast stijgt substantieel en extra aandacht voor de financiële middelen van deze toezichthouders is wenselijk. De raad doet daarbij de volgende aanbevelingen:

- Het goed borgen van de onderlinge samenwerking tussen toezichthouders. Mede gezien de samenhang tussen cybersecurity en AI is het daarbij essentieel dat elke (per sector) aangewezen toezichthouder ook focust op AI-toepassingen en er niet één toezichthouder komt die zich (over alle sectoren heen) overkoepelend op AI gaat richten.
- Het opzetten van een brede pool van experts, inclusief cybersecurityspecialisten, waarvan alle toezichthouders gebruik kunnen maken. Dit zal de hierboven genoemde *cross-over* effecten vergroten.
- Het verder stimuleren van datagedreven werken op het gebied van toezicht, hetgeen ook hier tot meer efficiëntie kan leiden.

Tenslotte

In lijn met de hierboven geïdentificeerde risico´s en de aanvullende vraagstukken, is het essentieel om in de nabije toekomst ook bij nieuwe AI-toepassingen oog te blijven houden voor de cybersecurity-aspecten. De raad vraagt daarbij nadrukkelijk extra aandacht voor het aanbrengen van ´vangrails´. Het gaat daarbij niet alleen om extra regulering, maar ook om blijvende aandacht voor het belang van menselijke beoordeling, duiding en (waar nodig) aanpassing van AI-output bij het nemen van zwaarwegende beslissingen of het uitvoeren van impactvolle acties, ook op het vlak van cybersecurity.

Daarnaast moeten technische maatregelen voor het inrichten van digitaal veilige omgevingen voor AI-toepassingen worden gestimuleerd, in combinatie met sturing op de kwaliteit van AI-software zelf, onderliggende modellen en data. Verder is naast al bestaande regulering ook aanvullende regelgeving noodzakelijk, zowel voor de veiligheid van AI-software zelf als voor software die met generatieve AI is gemaakt. Tevens is het nodig om de samenwerking met de grote technologiebedrijven te intensiveren, ook op Europees niveau. Dit maakt het beter mogelijk om digitale AI-producten en -diensten veiliger te maken.

[10] Zeer recent is overeenstemming intern de EU bereikt over implementatie van deze nieuwe wet, die in het voorjaar van 2024 in werking zal treden.

Door het mitigeren van specifieke cybersecurityrisico's, het creëren van awareness en het zorgen voor transparantie over de wijze van toepassing, zal het vertrouwen in het gebruik van AI kunnen groeien. Als daarbij verdere innovatie wordt gestimuleerd op de hierboven genoemde gebieden, kan binnen alle onderdelen van onze maatschappij op een verantwoorde manier gebruik worden gemaakt van deze technologie.

Een afschrift van deze brief wordt ook gestuurd aan de ministers van Justitie en Veiligheid, en Economische Zaken en Klimaat.

Hoogachtend,
Namens de Cyber Security Raad,

Pieter-Jaap Aalbersberg
Covoorzitter CSR

Theo Henrar
Waarnemend covoorzitter CSR

Bijlage 1: achtergrond bij generatieve AI versus 'klassieke' AI

De Wetenschappelijke Raad voor het Regeringsbeleid (WRR) geeft aan dat Artificiële intelligentie (AI) niet in één definitie te vatten is. In het algemeen wordt met de term AI het volgende aangeduid: "*Het soort systemen dat intelligent gedrag vertoont door hun omgeving te analyseren en – met enige graad van autonomie – actie te ondernemen om specifieke doelen te bereiken"*.

De werktitel 'klassieke' AI verwijst daarbij naar (het bouwen van) systemen die een zekere mate van 'intelligent gedrag' vertonen, vaak door in grote hoeveelheden datapatronen te detecteren, om vervolgens in nieuwe data vergelijkbare patronen te leren herkennen (*machine learning*). Deze vormen van AI worden over het algemeen ingezet als hulpmiddel om op geautomatiseerde wijze analyses te kunnen uitvoeren en verbanden tussen data te kunnen leggen. Dit leidt bijvoorbeeld tot voorspellingen, op basis waarvan mensen geïnformeerde beslissingen kunnen nemen. Er zijn ook andere toepassingen, bijvoorbeeld voor autonoom bestuurbare voertuigen.

Generatieve AI gaat een belangrijke stap verder ten opzichte van 'klassieke' AI. Het betreft een nieuwe generatie van AI-systemen (maar het is slechts één vorm van AI). Hierbij is het mogelijk om nieuwe inhoud te genereren voor een bepaalde toepassing, op basis van eerder ingevoerde datasets. Generatieve AI-modellen zijn getraind met grote hoeveelheden data, waarbij het kan gaan om taalmodellen (Natural Language Models, NLM), maar ook beeld(vision) generatiemodellen, broncode generatiemodellen en audiogeneratiemodellen. Recente ontwikkelingen op het gebied van generatieve AI hebben geleid tot een wildgroei aan allerlei nieuwe tools en trainingsprogramma´s (waaronder ChatGPT, Bard, DALL-E, Bloom etc.).