NL SECURE [ID]

» AI-systemen nemen taken over van mensen, en nemen zelfstandig beslissingen. Dat biedt een enorme potentie, maar introduceert ook geheel eigen, specifieke risico's. De beveiliging van AI verdient dan ook veel meer aandacht dan het nu doorgaans krijgt, stelt Henk-Jan van der Molen. Als docent aan de Security Academy pleit hij voor een meerlaagse aanpak. "Maar uiteindelijk is de inzet van AI altijd risicovol en een menselijke afweging."

De impact van AI **is een twee-snijdend zwaard**

Henk-Jan van der Molen
Docent, Security Academy

AI is in opkomst. Chipfabrikanten gebruiken kunstmatige intelligentie bijvoorbeeld om te zoeken naar nog slimmere manieren om transistoren te plaatsen. Daarmee neemt de rekenkracht verder toe, en in het kielzog ook de kracht en de mogelijkheden van AI-systemen. Het is daarmee een zichzelf versterkende katalysator voor de wet van Moore. "Uiteindelijk gaan we naar een AI-systeem dat de Turing-test doorstaat en dus kan communiceren als een mens", voorspelt Van der Molen.

AI-systemen spelen door die toenemende rekenkracht en mogelijkheden een steeds belangrijkere rol in ons dagelijks leven. Deze systemen hebben een waardevol voordeel ten opzichte van reguliere IT: ze kunnen in nieuwe situaties autonoom beslissingen nemen, zonder dat ze expliciet daarvoor zijn geprogrammeerd. Dat is een welkome eigenschap voor tal van toepassingen. Denk aan een zelfrijdende auto die zelfstandig een route bepaalt in een drukke ochtendspits. Of denk aan een HR-systeem dat uit duizenden vacatures de tien meest geschikte kandidaten voor een nieuwe vacature destilleert.

Die autonomie maakt AI waardevol, maar ook risicovol. We moeten erop kunnen vertrouwen dat zo'n systeem integer handelt, zonder manipulatie van kwaadwillenden. Dat kan volgens Van der Molen namelijk onvoorziene, mogelijk rampzalige gevolgen hebben. "Als bij een regulier IT-systeem de data gemanipuleerd is, dan is er nog niet per se veel aan de hand. Maar stel je voor dat een zelfrijdende auto door een hack denkt dat het 100 meter achter een andere auto zit, terwijl dat in werkelijkheid 10 centimeter is. Dat kan leiden tot levensgevaarlijke situaties. Dan is het wel belangrijk dat er een bepaalde controle is die de scope van die autonomie toetst."

kpn Security

Autonomie vraagt om beveiliging
Het autonome karakter van AI-systemen vraagt dus om beveiligingsmaatregelen. Die beveiliging wijkt af van die van traditionele systemen, legt Van der Molen uit: "Bij traditionele IT draait security om het beschermen van processen tegen falen. Je kunt je afvragen of dat uitgangspunt ook opgaat bij AI-systemen. Het gaat dan niet alleen om falen, maar vooral ook om het voorkomen van ongewenst handelen."
Een belangrijke factor die dat handelen bepaalt, zijn de trainingsdata. "Bij traditionele systemen is de logica grotendeels voorgeprogrammeerd. Bij AI-systemen is dat anders. Daarbij wordt de logica grotendeels bepaald door de gegevens waarmee het algoritme getraind wordt. Bij spellen zoals schaken en Go gaat dat goed, AI-systemen zijn hierin heer en meester. Maar bij veel andere toepassingen is het verschil tussen succesvolle inzet en levensgevaarlijke situaties flinterdun. Daarom is het belangrijk dat er voldoende controle is over de authenticiteit van die trainingsdata."

Die controle is volgens Van der Molen cruciaal. "Aanvallers kunnen door het manipuleren van trainingsdata de werking van het systeem beïnvloeden. Een AI-systeem gebruikt trainingsdata om bijvoorbeeld patronen te leren herkennen, en zo te bepalen op basis van welke condities het beslissingen neemt. Veranderen die data, bijvoorbeeld doordat een hacker deze manipuleert, dan verandert uiteindelijk de werking van het systeem. Daarnaast kan een AI-systeem ook veel privacygevoelige persoonsgegevens bevatten. Deze verdienen natuurlijk sowieso bescherming tegen datalekken."

"Autonomie maakt AI waardevol, maar ook risicovol."

Henk-Jan van der Molen

Databescherming cruciaal
Het beveiligen van (trainings)data is dan ook cruciaal voor een betrouwbare werking van AI-systemen. Volgens Van der Molen bestaan daarvoor al jaren goede strategieën. "Een aantal modellen voor informatiebeveiliging is heel bruikbaar, zoals Bell-LaPadula en Biba (zie kader). Deze methoden zijn weliswaar al tientallen jaren oud, maar nog altijd heel waardevol. Deze voorkomen dat gegevens naar buiten lekken, of worden gemanipuleerd door onbevoegden."

ISO-normeringen kunnen verdere houvast bieden. "Zonder objectieve standaarden is het erg lastig om verbeterpunten in de beveiliging te identificeren. ISO-normeringen die specifiek bruikbaar zijn voor AI-beveiliging zijn nog schaars, maar volop in ontwikkeling. De gepubliceerde normen dekken nog maar een beperkt deel van het AI-vakgebied af. Wel zitten er nog 22 normen in de pijplijn, waarmee ISO uiteindelijk de hele Plan Do Check Act (PDCA)-cyclus voor AI-beveiliging wil afdekken. Voor het zover is, is het voor eigenaren van AI-systemen raadzaam regelmatig te controleren of er nieuwe normen zijn uitgevaardigd."

Situational awareness
Ondanks een zo goed mogelijke beveiliging van de data kan het alsnog in de praktijk misgaan. "De beveiliging van AI is uiterst complex. Naast de manipulatie van trainingsdata kunnen hackers ook knoeien met het algoritme, of met de data die het systeem verwerkt. Daarnaast is ook nog simpelweg een Denial of Service (DoS)-aanval mogelijk." Een van de voorwaarden voor het gebruik van AI-systemen zou volgens Van der Molen dan ook een soort 'situational awareness voor het eigen functioneren' moeten zijn. "Als een AI-systeem zijn werk niet meer betrouwbaar kan uitvoeren, dan zou het systeem zijn eigen ambities moeten beperken om erger te voorkomen."

"Neem bijvoorbeeld de zelfrijdende auto", vervolgt Van der Molen. "Die zou zichzelf tijdens een cyberaanval direct moeten parkeren langs de kant van de weg, in plaats van met 100 kilometer per uur door te rijden. Daar is een soort kunstmatig bewustzijn voor nodig dat continu checkt op verdachte of afwijkende omstandigheden. Zodat wanneer het misgaat, het zelfstandig kan ingrijpen en zichzelf desnoods kan uitschakelen."

Bell-LaPadula: bewaken van vertrouwelijkheid
Een bruikbaar model voor de beveiliging van AI-systemen is het model Bell-LaPadula. Deze methode voorkomt datalekken met het credo 'no read up, no write down'. In dit model is het voor gebruikers onmogelijk om informatie te raadplegen of te wijzigen met een hogere kwalificatie dan zijn of haar screening. Tegelijkertijd kan iemand met de juiste screening geen geheime informatie kopiëren naar lagen met een lager classificatieniveau.

Biba: bewaken van integriteit
Een ander nuttig model voor de beveiliging van AI-systemen is Biba. Dit model hanteert het credo 'no write up, no read down'. Dit model voorkomt ongewenste wijzigingen van gegevens, en daarmee datavervuiling. Iemand met toegang tot laagwaardige informatie mag deze niet kopiëren naar locaties voor hoogwaardige informatie. Tegelijkertijd mag een uitvraag van hoogwaardige informatie geen laagwaardige data bevatten.

Tweesnijdend zwaard
AI heeft een interessante eigenschap: deze technologie kan helpen bij zijn eigen verdediging. Veel fabrikanten hebben kunstmatige intelligentie op de een of andere manier verwerkt in hun securityproducten. Denk aan detectie-oplossingen die uit grote hoeveelheden dataverkeer verdachte onregelmatigheden kunnen opsporen die mogelijk wijzen op indringers. Ook is AI waardevol voor het opsporen van kwetsbaarheden. “Veelbelovend is de technologie om uit grote hoeveelheden regels code volautomatisch te speuren naar bugs. Die vormen namelijk dankbare ingangen voor cybercriminelen, die deze kunnen misbruiken voor een hack.”

AI heeft voor zijn eigen verdediging nog veel meer in zijn mars. “Mits je voldoende flexibele logica hebt, kun je bijvoorbeeld met AI een systeem ontwikkelen dat trainingsdata kan controleren op verdachte afwijkingen die mogelijk kunnen wijzen op manipulatie.”

Wel maakt Van der Molen een scherpe kanttekening: het gebruik van AI binnen de cybersecurity is helaas een tweesnijdend zwaard. “Ook aanvallers kunnen misbruik maken van kunstmatige intelligentie. Ze kunnen het inzetten als wapen. AI in 2008 ontwikkelde de Carnegie Mellon-universiteit een proof of concept systeem dat volautomatisch malware kon ontwikkelen aan de hand van een software-update.” Een recenter voorbeeld is volgens Van der Molen het algoritme Mayhem, dat in 2016 het levenslicht zag. “Tijdens een DARPA Grand Cyber Challenge wist dit algoritme volledig automatisch de meeste kwetsbaarheden te ontdekken in aangeboden software.”

Volautomatische, AI-gedreven cyberwapens zijn problematisch. “Het cyberdomein is ook zonder AI al een asymmetrisch speelveld”, benadrukt Van der Molen. “Aanvallers hoeven voor een succesvolle hack maar één kwetsbaarheid te vinden en te misbruiken. Terwijl de verdedigende partij alle mogelijke kwetsbaarheden moet dichten. AI maakt dat speelveld nog meer asymmetrisch, bijvoorbeeld omdat een AI-cyberwapen een succesvolle aanval kan uitvoeren voordat een mens kan reageren. Verdedigers die daartegen een AI-firewall inzetten die zich automatisch kan afkoppelen het internet, moeten rekening houden met de impact van valse alarmen.”

“Bij AI-systemen draait beveiliging niet langer om de vraag: hoe beschermen we processen tegen falen? Het doel moet zijn: hoe beschermen we mensen tegen het falen van AI-systemen? Door de grote risico’s van AI is een goede risicoafweging dan ook altijd noodzakelijk. Uiteindelijk kunnen en moeten alleen mensen bepalen of, waar en hoe een AI-systeem zijn werk mag doen.”

“Hoe beschermen we mensen tegen het falen van AI-systemen?”

Henk-Jan van der Molen

Henk-Jan van der Molen
is docent aan de Security Academy.