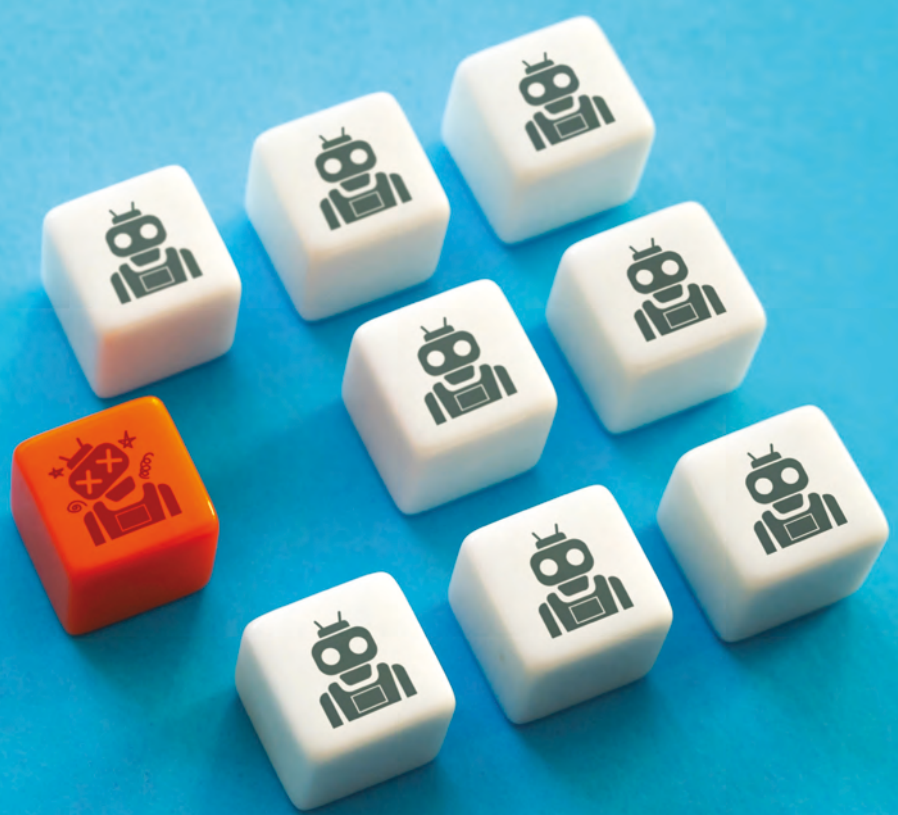
Auteur: Sabine Steenwinkel-den Daas, Consultant Responsible AI bij Highberg, zij is bereikbaar via e-mail: sabine.dendaas@highberg.com

Verantwoord inzetten van algoritmen en AI

AI beïnvloedt ook het informatiebeveiligingsvakgebied. Het geeft enerzijds meer mogelijkheden om bedreigingen op te sporen, anderzijds geeft het cybercriminelen geavanceerde aanvalstechnieken in handen. Zoals alle IT-systemen hebben ook AI-systemen beveiligingskwetsbaarheden en zijn ze hierdoor potentiële doelwitten van cyberaanvallen. Er zijn zowel kansen als risico's. Hoe kun je als organisatie verantwoord omgaan met AI?



Afwegingskaders.

Verantwoorde inzet van AI valt binnen de 'sweetspot' van wat wettelijk mag, wat technisch kan en wat ethisch wenselijk is.

1. **Wat wettelijk mag:** AI-systemen moeten voldoen aan bestaande en toekomstige wet- en regelgeving. Dit houdt bijvoorbeeld in dat ook AI verantwoord met persoonsgegevens moet omgaan en dat informatie gebruikt door AI voldoende veilig is, zoals de AVG en de NIS2 voorschrijven. Specifiek voor AI-systemen gelden aanvullende productvereisten met de AI Act.
2. **Wat technisch kan:** hoewel de mogelijkheden van AI enorm zijn, is het cruciaal om realistisch te blijven over wat technisch haalbaar is en vooral hoe realistisch de adoptie ervan is. Bovendien wordt een AI-systeem kwalitatief nooit beter dan de kwaliteit van de informatie die het toegediend krijgt: 'garbage in = garbage out'.
3. **Wat ethisch wenselijk is:** zelfs als iets technisch mogelijk en wettelijk toegestaan is, betekent dit niet dat we de inzet van een AI-systeem moreel acceptabel vinden. Het streven naar verantwoorde AI-inzet houdt in dat men ook rekening houdt met waarden zoals transparantie, duurzaamheid en eerlijkheid. Dit zorgt ervoor dat AI niet alleen effectief, maar ook oprecht bijdraagt aan de waarden die voor een organisatie belangrijk zijn.

Verantwoorde inzet van AI-systemen vereist dat deze drie elementen in balans worden gehouden, zodat AI-systemen per saldo positieve impact maakt zowel binnen als buiten de organisatie.

Maak de afweging tussen wat kan, wat mag en wat wenselijk is en breng daar balans in aan

De AI act: een risicogebaseerde aanpak

De AI Act is een Europese wet die regels stelt voor de ontwikkeling en het gebruik van AI-systemen. Ook geeft de wet rechten aan burgers die in aanraking komen met AI-systemen. Het doel van de wet is dat AI-systemen die organisaties (binnen de EU) gebruiken, veilig zijn en fundamentele rechten respecteren (1).
De AI Act gaat uit van een risicogebaseerde aanpak. Dit houdt in dat afhankelijk van het domein waar deze ingezet wordt, een bepaalde risicocategorie en daarbij behorende set aan regels gelden. In totaal kent de verordening drie risiconiveaus: onacceptabel, hoog en beperkt. Het uitgangspunt hierbij is dat hoe groter het risico, des te strenger de regels zijn die hiervoor gelden. De risicocategorie van een systeem hangt nauw samen met het toepassingsgebied en ziet er dus op toe dat er strengere regels zijn als de toepassing ervan kritischer is, ofwel meer impact heeft. Daarnaast gelden er specifieke vereisten voor leveranciers van generieke AI-systemen om ook oplossingen met een generieke toepassing te reguleren.

Systemen met een onacceptabel risico zijn door de EU vanaf februari 2025 niet toegestaan (1). Een onacceptabel risico kenmerkt zich door ethische onwenselijke activiteiten. Vaak zijn deze activiteiten al vanuit andere wetgeving verboden zoals etnische profilering.
In de beperkt-risicocategorie vallen AI-systemen die content genereren (generatieve AI), voor deze toepassingen geldt een transparantieverplichting.

Voor AI-systemen in 'hoog risico'-toepassingsgebieden gelden aanvullende verplichtingen, zoals het inrichten van een risico- en kwaliteitsmanagementsysteem in de organisatie van de Gebruiksverantwoordelijke (bijvoorbeeld

ISO42001), uitvoeren van een impact assessment op mensenrechten (bijvoorbeeld IAMA (2)), registreren in een Europees (algoritme)register en eigen administratie; implementeren van loggingsmogelijkheden en bewaartermijnen, inrichten van 'human oversight' en treffen van passende beveiligingsmaatregelen (3).
De risicobenadering uit de AI-act classificeert alleen AI-systemen in specifieke toepassingsgebieden als hoog risico. De bepalingen gelden dus niet voor algoritmen met hoge impact:

- ook buiten deze domeinen kunnen AI-systemen grote impact hebben op betrokkenen;
- ook algoritmen, die niet onder de definitie van AI vallen, kunnen impact hebben op betrokkenen.

De publicatiestandaard voor het algoritmeregister (4) van Nederlandse Overheden bevat een hulpmiddel voor de selectie voor publicatie van algoritmen. Het is te adviseren om ook voor de algoritmen met impact op betrokkenen (categorie B) eveneens de afweging te maken dit ook te integreren in je risico- en kwaliteitsmanagementsysteem om passende aanvullende maatregelen te nemen zoals het uitvoeren van een IAMA en/of adequate menselijke tussenkomst.

Risicoclassificatie en risico-assessments bij AI-systemen

Bij een risicoclassificatie voor een AI-systeem zul je dus al snel beginnen met het inschatten van de risicocategorie volgens de AI-Act en – afhankelijk van het beleid van je organisatie – ook concluderen of er sprake is van een impactvol algoritme. Het gaat hierbij dus vooral om hoe het algoritme wordt ingezet.
Het uitvoeren van risico-assessments voor AI verschilt op belangrijke punten van de standaard IT-risicoanalyses. Dit is te verklaren doordat traditionele IT-systemen anders werken dan AI-systemen. Traditionele IT-systemen volgen vaak vastgestelde parameters en deterministische processen, waar AI extra complexiteit en onvoorspelbaarheid introduceert doordat het juist ontwikkeld is om zelf patronen te herkennen. De parameters staan dus niet meer vast. Hierdoor gaan andere risico's een rol spelen. Denk hierbij aan:

Zelflerende en adaptieve systemen

Doordat AI-systemen zelf patronen herkennen en hierdoor de gevonden patronen dynamisch kunnen aanpassen op basis van nieuwe gegevens, kan het voorkomen dat deze systemen bijvoorbeeld onbedoeld onjuiste of zelfs schadelijke patronen aanleren.

Uitlegbaarheid

Het is vaak moeilijk om precies te verklaren hoe een AI-systeem tot een bepaalde beslissing komt. In een traditionele IT-context is het over het algemeen eenvoudiger om beslissingen goed te herleiden. Voor elke context zal afgewogen moeten worden of beslissingen van het systeem wel voldoende uitlegbaar zijn.

Gevoeligheid voor bias (5)

AI leert patronen op basis van data. Hierdoor kan bias ontstaan als deze data (trainingsdata) niet representatief zijn of zelfs expliciete of impliciete vooroordelen bevatten. Dit kan leiden tot ongewenste effecten, zoals ongelijke behandeling.

Impact van autonome beslissingen

Wanneer de resultaten van een AI-systeem gebruikt worden om autonoom beslissingen te nemen, zoals het automatisch blokkeren van een IP-adres bij de detectie van verdachte activiteiten, kan het zijn dat er onterecht verdachte activiteiten zijn gedetecteerd. Dit komt doordat het een inschatting is en geen zekerheid. Een fout in de besluitvorming kan directe operationele gevolgen hebben. Er moet een afweging gemaakt worden wat enerzijds de foutmarge is, die geaccepteerd wordt, en wat anderzijds de gevolgen zijn van het missen van verdachte activiteiten.
Om algoritmerisico's te identificeren en beheersen ontkom je er niet aan om ook voor algoritmen risico- en kwaliteitsmanagement te borgen in de organisatie. Begin vandaag nog met de uitvoering en verbeter gaandeweg, in plaats van alles vooraf tot in detail uit te werken. Door te doen, ga je leren wat wel werkt en ook wat niet. Zo ervaar je samen wat nog verder, duidelijker of anders uitgewerkt moet worden om een 'AI-volwassen'-organisatie te worden. Wil je dit gestructureerd aanpakken en continue blijven verbeteren? Start met het inrichten en implementeren van een AI-governance.

AI-volwassenorganisatie met een AI-governance

AI-governance kun je zien als het totaalpakket van (beleids)afspraken, beleidskaders, processen, procedures, taken, rollen en verantwoordelijkheden; om enerzijds de risico's van AI-systemen te beheersen en anderzijds de kansen van AI maximaal en op verantwoorde en efficiënte

Onderdelen van AI-governance (6).

wijze te benutten. AI-governance is vaak een verlengde of onderdeel van een data- of IT-governance.
AI-governance bestaat dus uit een overzichtelijke set richtlijnen waarin jouw organisatie haar visie op AI, het beheersen van risico's en de verdeling van verantwoordelijkheden vastlegt. Denk hierbij aan een AI-visie, ethisch kader, risicomanagementprocessen en een duidelijke beschreven rolverdeling.

Door in een AI-governance vast te leggen wat afspraken zijn, processen en procedures te beschrijven en duidelijk toe te lichten wie wat moet doen hierin, ontstaat duidelijkheid over ieders rol om verantwoord met AI-systemen om te gaan. Ook beschrijf je tijdens het opstellen van een AI-governance wat er allemaal nodig is aan kennis, capaciteit en middelen om de ambities in de governance te realiseren. Dit faciliteert het organiseren van de benodigde capaciteit, middelen en budget voor succesvolle implementatie van verantwoorde AI-systemen.
Als organisatie bepaal je zelf de scope van deze governance. Bijvoorbeeld wanneer er voor informatiebeveiliging of privacy al veel risico's afgedekt worden. Of om als overheidsorganisatie in lijn te komen met de publicatiestandaard voor het algoritmeregister door te kiezen voor een bredere scope dan AI-systemen, een algoritmegovernance.

Is het mogelijk om als organisatie verantwoord om te gaan met AI?

Het belangrijkste punt hierbij is dat de afweging tussen wat kan, wat mag en wat wenselijk is in balans gebracht kan worden. Bijvoorbeeld doordat de governance rondom AI ingericht is en er gedegen risicomanagement plaatsvindt. Zo zorg je als organisatie dat je in control bent en blijft van AI-systemen in gebruik.

Referenties

(1) https://www.linkedin.com/pulse/power-trio-how-ai-quantum-computing-cybersecurity-our-lokhandwala-hlxmf
(2) https://www.ibm.com/case-studies/cargills-bank-ltd
(3) https://www.splunk.com/en_us/customers/success-stories/tide.html
(4) https://www.trellix.com/assets/case-studies/trellix-election-casestudy.pdf
(5) https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
(6) https://spectrum.ieee.org/real-time-deepfakes
(7) https://drlee.io/ai-security-model-hacking-with-model-inversion-attacks-techniques-examples-and-real-world-a23b5fff272a
(8) https://neptune.ai/blog/adversarial-machine-learning-defense-strategies