dialogic
innovatie • interactie

Gebruik van en toezicht op AI-toepassingen in telecominfrastructuren

Advies aan de toezichthouder over inrichting van risico-gebaseerd AI-toezicht

In opdracht van:
Agentschap Telecom

Project:
2019.166

Publicatienummer:
2019.166.2004 v1.2.6

Datum:
Utrecht, 16 juni 2020

Auteurs:
ir. Tommy van der Vorst
ir. Nick Jelicic
ir. Jan van Rees
prof. dr. ir. ing. Rudi Bekkers
ir. ing. Reg Brennenraedts MBA
Roma Bakhyshov MSc

Managementsamenvatting

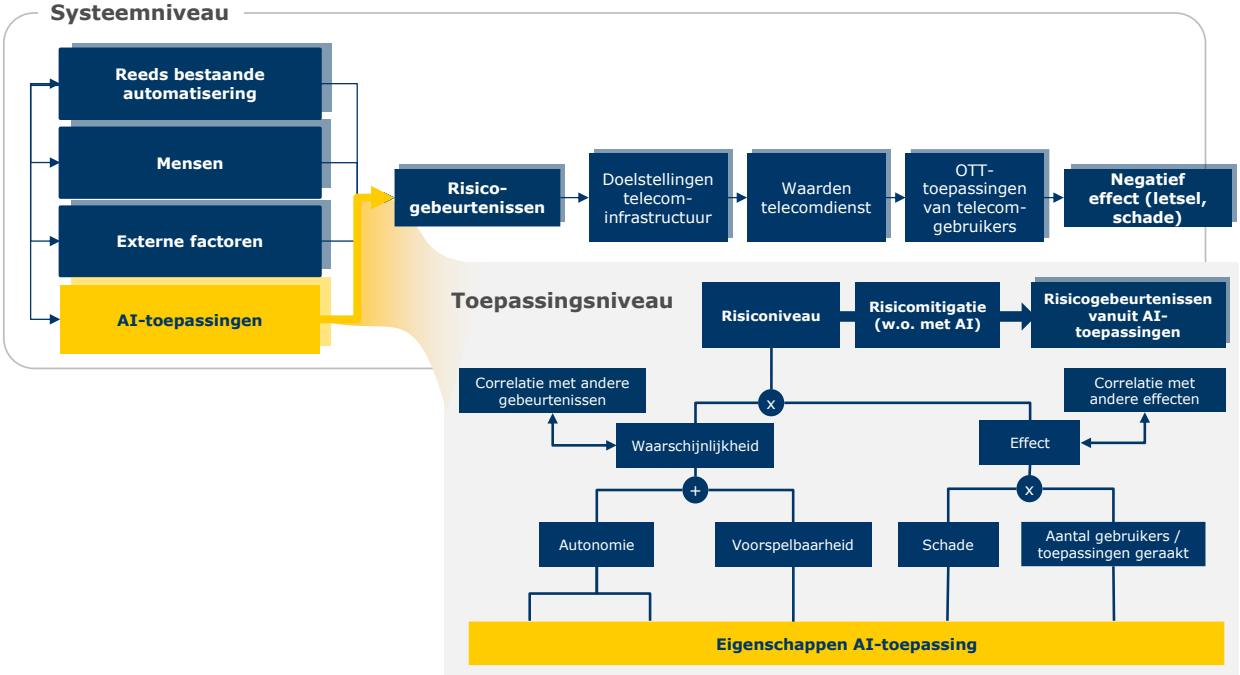
Telecominfrastructuren zijn van vitaal maatschappelijk belang. Steeds meer toepassingen zijn afhankelijk van goed werkende, betrouwbare en altijd beschikbare telecomvoorzieningen. In Nederland houdt het Agentschap Telecom hier toezicht op. Door de opkomst van AI-toepassingen in telecominfrastructuren veranderen het karakter en de risico's van de telecomsector fundamenteel. Om de juiste werking van telecominfrastructuur te borgen is een aanpassing van de relevante kennis, aanpak van het toezicht en het toezichtbeleid noodzakelijk. Dit rapport geeft inzichten in de veranderingen als gevolg van toepassing van AI, de risico's en mogelijke aanpak voor Agentschap Telecom om maatschappelijk vertrouwen in de telecom infrastructuur te behouden. Er zijn drie onderzoeksmethoden ingezet: literatuuronderzoek, interviews en sessies met experts.

Beantwoording onderzoeksvragen

Wat zijn risico's van huidige en toekomstige inzet van AI in de telecomsector?

De relevante definitie van AI in dit kader is, gezien de ontwikkelingen, het gebruiken van algoritmes op basis van deep learning, getraind met behulp van grote hoeveelheden data, om taken te automatiseren die voorheen alleen (goed) door een mens zouden kunnen worden uitgevoerd. AI zal naar verwachting een steeds centralere rol innemen in telecomnetwerken.

AI-toepassingen hebben *specifieke eigenschappen* welke risico's kunnen opleveren bij gebruik in telecominfrastructuren (het *toepassingsniveau*). AI-toepassingen interacteren met andere AI-toepassingen, mensen, 'gewone' automatisering, en mogelijk de buitenwereld. Het is daarom van belang om de toepassing van AI in de telecomsector te beoordelen op *systeemniveau;* dat wil zeggen, kijkend naar de effecten en risico's voor de gehele keten in plaats van AI-toepassingen in isolatie. Daaronder vallen nadrukkelijk de uiteindelijke toepassingen die worden gerealiseerd op basis van de telecominfrastructuren.



Op toepassingsniveau zijn de mate van autonoom leren en handelen, de mate van onvoorspelbaarheid, het handelingskader en de invloedssfeer van de AI-toepassing bepalend voor

de waarschijnlijkheid en de impact van de additionele risico's. Hier bovenop bestaan in de volledige levenscyclus van een AI-toepassing (planning, dataverzameling, training, testen en validatie en operatie) onverminderd de *gangbare* risico's ten aanzien van informatiebeveiliging.

Hoewel AI-toepassingen nieuwe (soorten) risico's introduceren kan het tot slot nadrukkelijk waarde toevoegen bij het *mitigeren* van risico's.

Hoe kan Agentschap Telecom als toezichthouder en uitvoeringsorganisatie deze risico's beperken?

We bevelen aan om te starten met instrumenten op *systeemniveau*. Met voorlichting en bewustwording, het vereisen van transparantie, het faciliteren van risicoanalyse en -mitigatie, het ontwikkelen van criteria en het stellen van procesvereisten kan Agentschap Telecom de risico's van AI-toepassingen in de telecomsector beperken. Op het *toepassingsniveau* kunnen eventueel specifiekere instrumenten worden ingezet. Daarbij kunnen certificering, auditing en handhaving op specifieke soorten of aspecten van AI-toepassingen een rol spelen. Tot slot zal in bredere zin een maatschappelijke discussie moeten plaatsvinden over het gewenste dienstenniveau van telecominfrastructuren.

Hoe ziet het huidige gebruik van AI er nu en in de komende vijf jaar uit in de telecomsector en sectoren die gebruik maken van digitale connectiviteit?

Op dit moment zien we dat de meeste toepassingen van AI in telecominfrastructuren betrekking hebben op optimalisatie van specifieke parameters. Het betreffen sterk afgebakende toepassingen, zoals het optimaliseren van parameters van een radiosignaal, power management, of het routeren van verkeer door een netwerk.

Kijken we naar de komende vijf jaar, dan zien we AI-toepassingen steeds geavanceerder worden. Een eindvisie, die wordt gedeeld door een aantal leveranciers van telecomapparatuur, is dat een groot deel van de netwerkfuncties in telecomnetwerken wordt bestuurd door een AI. Hoewel het de vraag of dit al (volledig) binnen 5 jaar wordt geïmplementeerd is deze eindvisie er wel degelijk een waar op geanticipeerd dient te worden.

Hoe kunnen de risico's voor de verschillende aspecten relatief ten opzichte van elkaar worden gewogen in een risicomodel voor digitale connectiviteit?

AI-toepassingen kunnen bepaalde eigenschappen bezitten die leiden tot aanvullende risico's voor telecominfrastructuren. De eigenschappen hebben te maken met de volgende aspecten van AI:

- **De mate van autonoom leren en handelen van de AI-toepassing.** Wanneer hier in hoge mate sprake van is, neemt de waarschijnlijkheid van risicogebeurtenissen toe. Een belangrijke parameter is of de toepassing wordt gecontroleerd door mensen of regels.
- **De mate van voorspelbaarheid van de AI-toepassing.** Wanneer de modellen niet-deterministisch of sterk niet-lineair zijn, is het lastiger om te valideren of een AI-toepassing onder alle omstandigheden goed werkt. Een factor die daarbij meespeelt is welke data wordt gebruikt en of die manipuleerbaar is.
- **Het handelingskader van de AI-toepassing.** Wanneer de AI-toepassing een sterk beperkte invloed heeft op telecominfrastructuren is het effect van een risicogebeurtenis beperkter. Een toepassing met een breed handelingskader leidt in potentie tot grotere effecten.
- **De invloedssfeer van de AI-toepassing**. Een toepassing die op centraal niveau werkt en een telecominfrastructuur bestuurt is risicovoller dan een toepassing die op laag niveau een specifieke parameter optimaliseert.

Onderstaand schema geeft een overzicht van de relevante aspecten en een weging daartussen. De scores kunnen worden gecombineerd volgens het schema onder "toepassingsniveau" hierboven.

Autonomie			Voorspelbaarheid			Schade		Scope		
Leren	Validatie	Handelen	Transparantie	Deterministisch & lineair	I/O	Handelingskader	Gebruik	Reikwijdte	Redundantie & variatie	
Continu lerend	Ongevalideerd	Closed loop	Intransparant, 'black box'	Niet deterministisch en niet lineair	Unconstrained/ untrusted I/O	Breed handelingskader	Closed loop	Centraal	Niet-redundant element	10
	Model niet in te zien of testbaar	AI in closed loop	Model niet in te zien en niet gecertificeerd	Stochastische AI-algoritmes	Gebruik van onbeperkte set gegevens	Inrichting netwerk	AI in closed loop	Network orchestrator		
		Constrained closed loop	Hoog aantal parameters	Volgorde-gevoelig (RNN's)	Gebruik gegevens derden	Besturing netwerk-element	Constrained closed loop	Edge	Redundant element	
Online lerend										
	Enkele scenario's getest	Human in closed loop	Trainings-data onbekend	Non-lineaire elementen	Gebruik meetgegevens eigen netwerk	Besturing verkeer	Human in closed loop	Basisstation POP, MDF		
	Alle input-combinaties getest							CPE		
Eenmalig getraind		AI in open loop	Gecertificeerd	Lineaire regressie	Alleen gegenereerde data	Optimalisatie van parameter	AI in open loop	Handset, terminal	Redundant gevarieerd element	
Offline lerend	Gevalideerd	Open loop	Verklaarbaar, 'white box'	Deterministisch en lineair	I/O constrained en trusted	Beperkt handelingskader	Open loop	Decentraal, geïsoleerd	Dubbel redundant, gevarieerd element	1

Score risicomodel

Eigenschappen AI-toepassing

Uitsluitend kijken naar de risico's van AI-toepassingen in isolatie geeft echter een te beperkt beeld van de maatschappelijke risico's (en overigens voordelen) van de inzet van AI-toepassingen in telecominfrastructuren. Op systeemniveau beïnvloeden de volgende factoren de risico's:

- **Interactie tussen AI-toepassingen en andere systemen.**
- **Vervanging van een mens door een AI.** Het door een mens laten uitvoeren van een taak brengt risico's met zich mee, en die kunnen hoger of lager zijn dan bij een AI-toepassing. In dit onderzoek worden de risico's van menselijk handelen in telecominfrastructuren niet in kaart gebracht. Het gepresenteerde model kan worden gebruikt om de risico's van de vervangende AI-toepassing te bepalen om zo de afweging bij het inzetten van een mens-vervangende AI-toepassing te kunnen maken.
- **Inzet van AI-toepassingen voor risicomitigatie.** Op systeemniveau kunnen AI-toepassingen bijdragen aan het verlagen van het risiconiveau, bijvoorbeeld door het sneller detecteren van problemen of aanvallen, en het ondersteunen bij het vinden van oorzaken en oplossingen.
- **Cyber(on)veiligheid van AI-toepassingen.** Uiteraard zijn ook AI-toepassingen onderhevig aan cyberdreigingen en bijbehorende veiligheidsrisico's. Deze risico's worden mogelijk vergroot, omdat voor het trainen van AI-toepassingen grote hoeveelheden (soms gevoelige) data bijeen worden gebracht.

Inhoudsopgave

Citeren als Dialogic, van der Vorst, T., Jelicic, N., et al. (2020). *Gebruik van en toezicht op AI-toepassingen in telecominfrastructuren. Advies aan de toezichthouder over inrichting van risico-gebaseerd AI-toezicht.* Agentschap Telecom, Groningen.

1 Introductie

1.1 Aanleiding

Sinds circa 2010 is kunstmatige intelligentie (Artificial Intelligence; AI) met een opmars bezig. Hoewel het concept al sinds de jaren '40 van de 20e eeuw bekend is en wordt toegepast, zijn door de toename van rekenkracht, opslagcapaciteit en telecommunicatie in de afgelopen decennia nieuwe mogelijkheden ontstaan. Nieuwe algoritmes op basis van *deep learning* maken gebruik van data om taken te leren en uit te voeren die voorheen slechts door mensen konden worden uitgevoerd. Omdat de hoeveelheden data die dergelijke geautomatiseerde systemen verwerken vele orden hoger liggen dan wat een mens kan verwerken, ontstaan er nieuwe mogelijkheden, maar ook moeilijkheden: de systemen zijn lastiger te analyseren en te 'begrijpen' door mensen. De toepassing van AI brengt daarom diverse maatschappelijke en ethische kwesties met zich mee.

AI wordt reeds toegepast in een groot aantal sectoren, waaronder de telecomsector. Telecominfrastructuren zijn van vitaal maatschappelijk belang. Steeds meer toepassingen zijn afhankelijk van goed werkende, betrouwbare en altijd beschikbare telecomvoorzieningen. Dit onderzoek beschouwt (mogelijke) risico's die kleven aan het gebruik van AI in telecominfrastructuren.

1.2 Onderzoeksvragen

In dit onderzoek wordt de volgende onderzoeksvraag beantwoord:

> **Wat zijn risico's van huidige en toekomstige inzet van AI in de telecomsector, en hoe kan Agentschap Telecom deze risico's beperken?**

Vanuit de opdrachtgever zijn daarbij de volgende deelvragen gedefinieerd:

1. Hoe ziet het huidige gebruik van AI eruit in de telecomsector en sectoren die gebruik maken van digitale connectiviteit?
2. Welke ontwikkelingen worden de komende 5 jaar[1] voorzien voor het gebruik van AI bij het verzorgen en gebruiken van digitale connectiviteit?
3. Welke risico's ten aanzien van beschikbaarheid, authenticiteit, integriteit, vertrouwelijkheid, transparantie en voorspelbaarheid ontstaan er in de diverse sectoren als gevolg van het huidige en toekomstige gebruik van AI? Hoe kunnen de risico's voor de verschillende aspecten relatief ten opzichte van elkaar worden gewogen in een risicomodel voor digitale connectiviteit?
4. Hoe kan Agentschap Telecom als toezichthouder en uitvoeringsorganisatie deze risico's beperken?

1.3 Aanpak

Om antwoord te geven op de onderzoeksvragen zijn drie onderzoeksmethoden ingezet: literatuuronderzoek, interviews en validatiesessies met experts. Ter beantwoording van de

[1] Een tijdshorizon van vijf jaar lijkt wellicht kort voor een verkenning als deze. Desondanks stellen we vast dat ontwikkelingen op het vlak van AI, binnen en buiten telecom, in volle gang zijn. Kijken we vijf jaar *terug*, dan was het op dat moment al een behoorlijke uitdaging geweest om de stand van AI van vandaag de dag te kunnen voorspellen.

onderzoeksvragen 2 en 3 is met name gebruik gemaakt van literatuur. Om toekomstige AI-ontwikkelingen in de telecomsector te verkennen is gekeken naar wetenschappelijk onderzoek naar AI in de telecomsector. Op basis hiervan is bepaald welke onderzoeksrichtingen zich later zullen vertalen naar toegepast onderzoek door leveranciers. Aanvullend is gekeken naar whitepapers en roadmaps van leveranciers op het gebied van telecom en AI om toekomstige innovaties te identificeren. Mitigatiestrategieën specifiek ten aanzien van AI-risico's zijn eveneens zo geïdentificeerd.

Interviews zijn ingezet om aanvullende informatie op te halen gerelateerd aan alle onderzoeksvragen. Literatuuronderzoek kan mogelijke richtingen geven in welke AI-toepassingen mogelijk zijn, maar dit betekent nog niet dat deze toepassingen ook daadwerkelijk in de Nederlandse telecomsector worden gebruikt. Door interviews te houden met telecomoperators is het beter mogelijk om te ontdekken welke AI-toepassingen nu en in de nabije toekomst concreet zullen worden ingezet in telecomnetwerken in Nederland. Om te identificeren welke richting onderzoek en ontwikkeling op kan gaan spraken wij met leveranciers van netwerkapparatuur. Zo zijn zowel ontwikkelingen aan de vraag- als de aanbodkant in kaart gebracht.

1.4 Leeswijzer

In hoofdstuk 2 beschrijven we allereerst de ontwikkeling die we op dit moment zien ten aanzien van AI in de telecomsector. We gaan in op wat AI, op dit moment en in de komende vijf jaar, betekent in de context van de telecomsector: welke AI-toepassingen zien en verwachten we in de telecomsector, welke kansen brengt dit met zich mee, en wat zijn in het algemeen risico's van AI-toepassingen? In hoofdstuk 3 presenteren we een model waarmee risico's van het gebruik van AI-toepassingen specifiek in telecominfrastructuren kan worden ingeschat. In hoofdstuk 4 gaan we in op de rol die Agentschap Telecom zou kunnen vervullen ten aanzien van het mitigeren van risico's van AI-toepassingen in telecominfrastructuren. Tot slot beantwoorden we kort de onderzoeksvragen in hoofdstuk 5.

2 De opmars van AI in telecom

In dit hoofdstuk gaan we allereerst in op het begrip AI: wat is het, en waarom is het relevant om ernaar te kijken in de context van risico's? Vervolgens kijken we naar huidige en toekomstige toepassingen van AI in telecominfrastructuren.

2.1 Wat is AI?

De term AI bestaat al sinds de jaren '40. Het is evident dat de term over de jaren een andere betekenis heeft gekregen. Voor dit onderzoek hanteren we de volgende werkdefinitie:

Kunstmatige intelligentie of AI is het gebruiken van algoritmes op basis van deep learning, getraind met behulp van grote hoeveelheden data, om taken te automatiseren die voorheen alleen (goed) door een mens (of in beperkte mate) door traditionele automatisering zouden kunnen worden uitgevoerd.

Deze werkdefinitie is gekozen om dit onderzoek op een passende manier in te kaderen, en niet bedoeld als normatieve definitie. Zouden we in brede zin kijken naar menselijke taken die door een machine worden overgenomen, dan vallen veel zaken, zoals een zakrekenmachine, ook onder "kunstmatige intelligentie". De onderzoeksvraag gaat echter specifiek in op een (gepercipieerde) nieuwe "golf" van AI-toepassingen in de telecomsector. Deze toepassingen vormen een nieuwe generatie van AI, waarbij het gebruik van deep learning algoritmes, grote rekenkracht, en grote hoeveelheden data kenmerkend zijn. Het is deze combinatie van ingrediënten die het noodzakelijk maakt om te kijken naar risico's: zoals we verderop toelichten leidt de inzet ervan (onder andere) tot systemen die lastiger te begrijpen en te controleren zijn.[2] In onderstaande paragraaf onderbouwen we deze definitie in de historische context.

2.1.1 Geschiedenis van AI

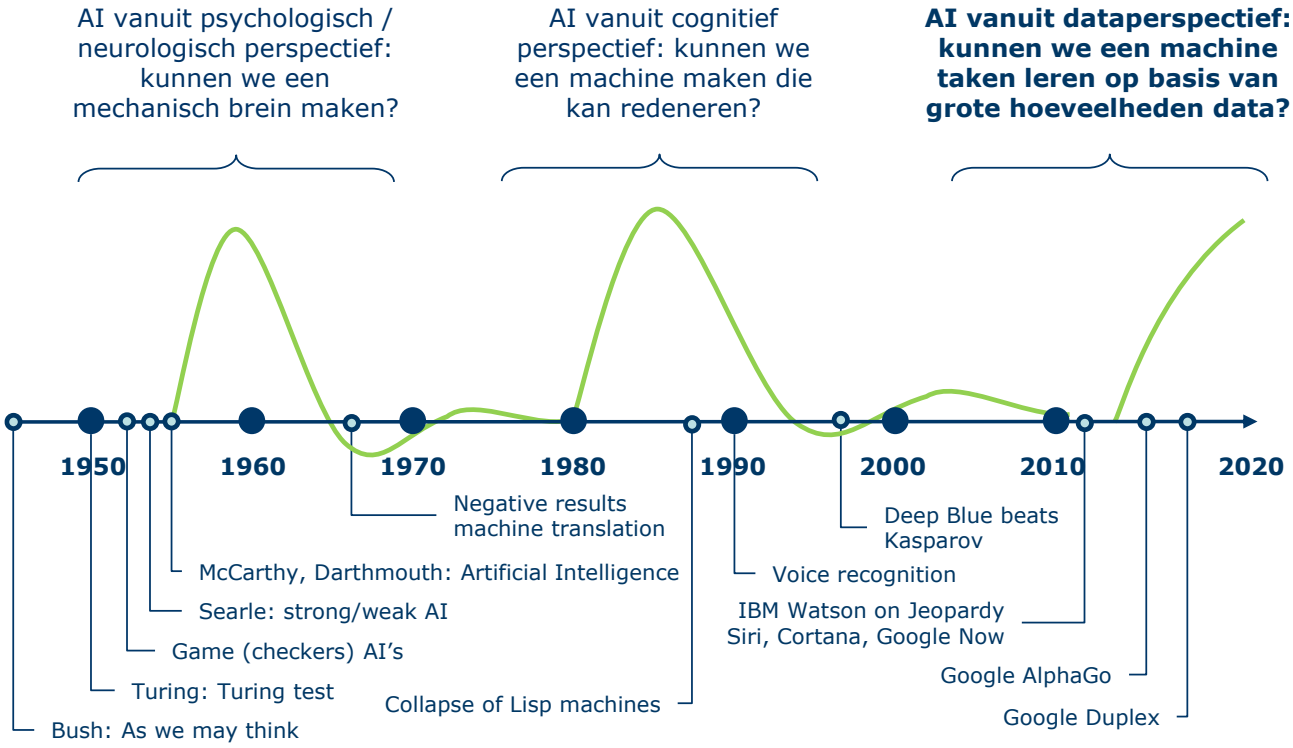
De geschiedenis van AI gaat verder terug dan menigeen zou vermoeden: de droom van het automatiseren van menselijk gedrag en redeneervermogen is terug te herleiden tot de oudheid. In Hellenistische mythes zien we bronzen automaten zoals Talos, die Kreta beschermde tegen piraten. Ook in de Middeleeuwen zien we ideeën van AI terug in de verhalen rondom de Golem van Praag. [1] Moderne ideeën rondom AI gaan terug naar de opkomst van de computer met sleutelfiguren zoals Turing, Walters en Minsky.

De term "kunstmatige intelligentie" werd in 1956 geïntroduceerd door wetenschapper John McCarthy. [2] Het wetenschappelijke veld heeft sindsdien een aantal cycli gekend; hoogtepunten waarin AI sterk werd gehypet, gevolgd door teleurstelling en kritiek (een zogenaamde 'AI-winter'). Het AI-veld heeft tot nu toe drie van dergelijke grote oplevingen en twee terugvallen gehad. De eerste opleving speelde rond de jaren 50 en 60, rond de periode dat de term in leven werd geroepen, gedreven door pioniers bij MIT en Stanford. In de jaren 70 werd er echter flink in de onderzoeksbudgetten gesneden, omdat AI in de praktijk niet in staat bleek om (bijvoorbeeld) teksten vanuit het Russisch naar het Engels te vertalen – destijds een toepassing waar vraag naar was en waarvan werd verwacht dat AI deze zou

[2] Het is zeer denkbaar dat bevindingen in dit onderzoek over AI-toepassingen bínnen de werkdefinitie ook (deels) gelden voor AI-toepassingen die (net) búiten de definitie vallen. Een aantal van de in hoofdstuk 3 gevonden risico's zijn bijvoorbeeld van toepassing op zelflerende systemen, ook wanneer dat leren niet werkt op basis van deep learning.

kunnen vervullen. Het klassieke voorbeeld hierbij is de vertaling van "*De geest is gewillig, maar het vlees is zwak*" in het Russisch, naar "*de wodka is goed maar het vlees is bedorven*" in het Engels, door een AI-toepassing. [3]

In de jaren tachtig zien we Japan sterk inzetten op AI om haar industrie vooruit te helpen. De Verenigde Staten en het Verenigd Koninkrijk volgen snel. In deze periode ligt de nadruk op expertsystemen. Deze systemen *emuleren* bepaalde specifieke taken/handelingen van mensen. Bij deze systemen was de intelligentie nog compleet met de hand geprogrammeerd en kon het systeem nog geen nieuwe taken aanleren zonder dat een mens de nieuwe regels programmeerde. De systemen zijn het best te beschrijven als een voorgeprogrammeerde 'beslisboom' die systematisch in software wordt doorlopen. Figuur 2 toont een voorbeeld van een dergelijke beslisboom – een expertsysteem zou uiteraard, om complexere beslissingen te faciliteren, bestaan uit een veel grotere 'boom'.

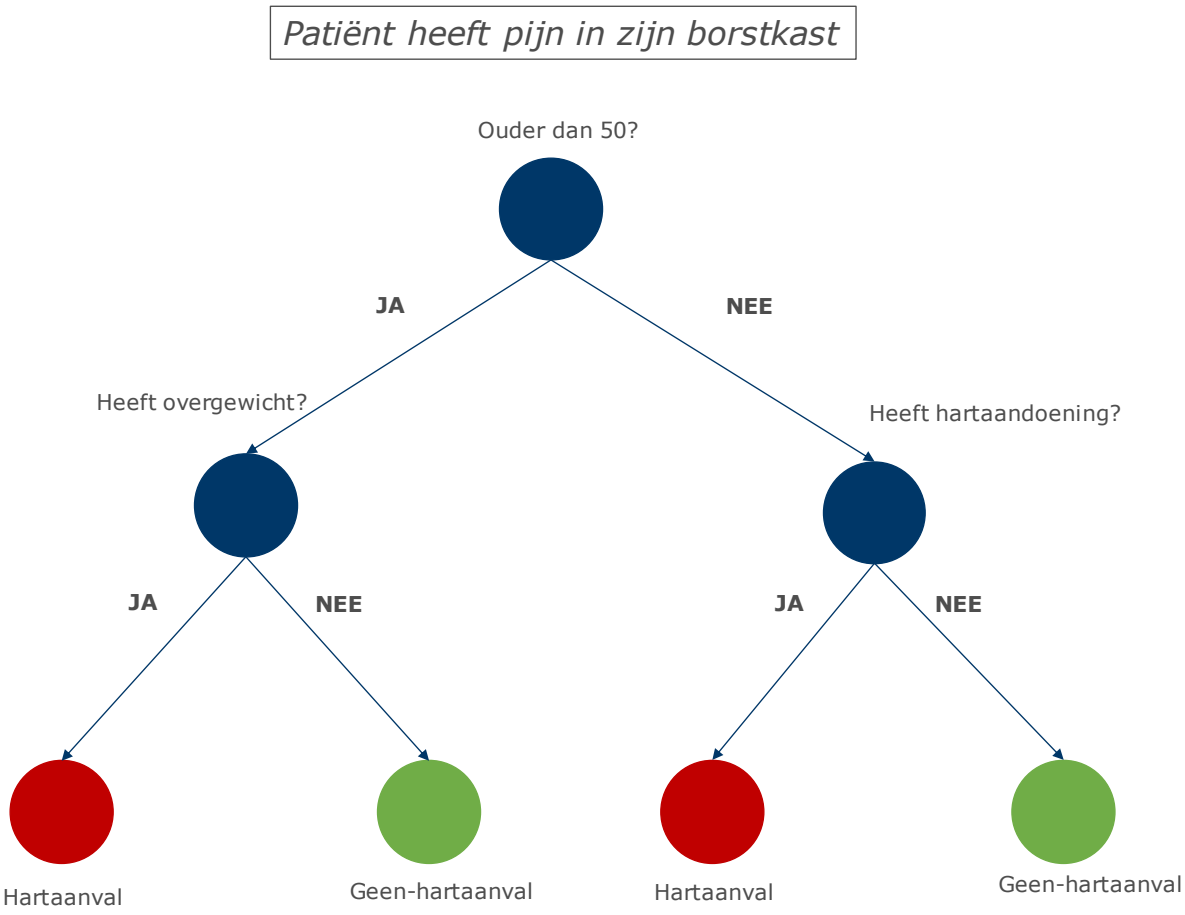


Figuur 1 Overzicht van de geschiedenis van AI en de drie 'waves' die daarin te onderscheiden zijn [3]

Expertsystemen zijn doorgaans ingewikkeld om te maken. Zo moet domeinkennis en programmeerkennis worden gecombineerd om zo krachtige software te ontwikkelen die 'menselijke' besluiten kan nemen. Hoewel de hype rondom expertsystemen groot is, worden de verwachtingen uiteindelijk niet waargemaakt. In de jaren 90 neemt de aandacht voor AI dan ook wederom sterk af. Dit neemt niet weg dat expertsystemen in meer of mindere mate werden geadopteerd en tot op heden succesvol worden gebruikt om beslissingen te automatiseren waaronder in de telecomsector [4].

Rond 2011 krijgt AI een nieuwe opleving door inbreng van onderzoekers als Andrew Ng [5], Geoffrey Hinton [6] en Yann LeCun [7]. Zij ontwikkelden *deep learning* – technieken waarmee een sprong kon worden gemaakt in de intelligentie van algoritmes. AI-toepassingen die tot dan toe onmogelijk werden geacht, bleken ineens haalbaar. Een voorbeeld is het door Google ontwikkelde *AlphaGo,* dat in 2016 de wereldkampioen Lee Sedol in het spel Go versloeg, terwijl tot dan toe werd gedacht dat alleen met menselijke intelligentie (en intuïtie) het spel Go op hoog niveau kon worden gespeeld. Het spel Go kent 10^{170} mogelijke geldige

bordposities [8] en per beurt tot wel 361 mogelijke zetten - een veelvoud van het aantal bordposities en zetten dat mogelijk is bij schaken, en te veel om met traditionele methodes (bijv. met een *minimax tree search*[3]) de meest optimale strategie te bepalen.



Figuur 2 De toepassing van een beslisboom op de eerste hulp: heeft iemand een hartaanval bij pijn in de borst?

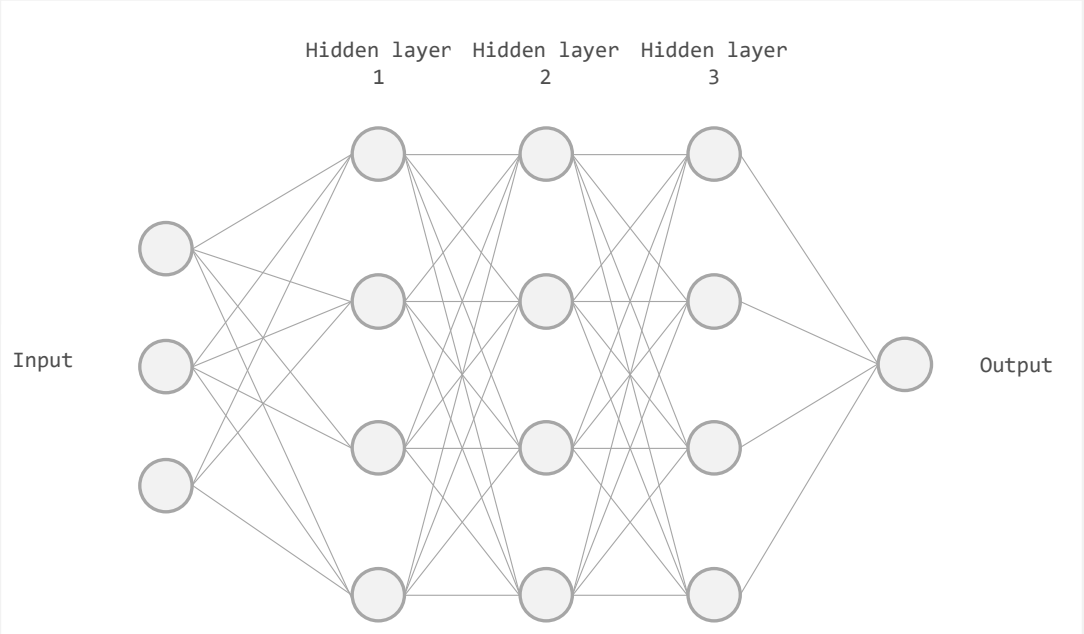
[3] Een minimax is een beslisregel die stelt dat de beste keuze de keuze is die een worst-case scenario voorkomt. In een schaakcomputer betekent dit dat de beste zet de zet is met de kleinste kans op het verliezen van een stuk. De effectiviteit van een minimax treesearch is afhankelijk van het aantal toekomstige stappen dat in acht wordt genomen. Hoewel minimax een eenvoudige regel is werd het door IBM met succes ingezet om Gary Kasparov te verslaan [47].

Wat is machine learning?

AI kan op verschillende manieren worden gerealiseerd. In de tijd van de expertsystemen was de intelligentie van het systeem met de hand geprogrammeerd. De ontwerper van het expertsysteem moest alle mogelijke handelingen van de AI zelf specificeren. Bijgevolg was de intelligentie van de expertsystemen vrij beperkt: alle situaties waarvoor de ontwerper geen regel had voorzien konden niet door het systeem worden behandeld. Tegenwoordig zijn AI-systemen niet *regelgedreven* maar *datagedreven*: de AI-systemen leren zelf de regels uit de data. De onderliggende algoritmes van deze zelflerende systemen worden doorgaans *machine learning algoritmes* genoemd.

Een speciale klasse van machine learning is *deep learning*. Bij deep learning wordt er iteratief een mapping tussen de input en de output van een model gezocht in de vorm van een serie aan wiskundige transformaties. Deze transformaties zijn geïnspireerd op de manier waarop ons menselijke brein werkt: een zogenaamd 'neuraal netwerk'. In onze hersenen prikkelen onze zintuigen hersencellen – de 'neuronen'. Deze neuronen sturen, afhankelijk van de prikkel, wel of geen signaal naar andere neuronen. Honderden miljarden neuronen leiden samen tot intelligent gedrag.[4]

Een neuraal netwerk bestaat uit verschillende lagen: een input laag, meerdere *hidden layers* en een output laag. De *hidden layers* extraheren eigenschappen uit de data op basis van de input. De eigenschappen van de laatste *hidden layer* worden uiteindelijk gebruikt om de output te construeren. In elke laag van het netwerk vermenigvuldigt elke neuron de input uit de vorige laag, vermenigvuldigt dit met een gewicht, sommeert al deze vermenigvuldigde inputs en past een non-lineariteit toe. Figuur 3 toont een schematische weergave van een neuraal netwerk met 3 *hidden layers*.



Figuur 3 Schematische weergave van een neuraal netwerk

[4] Daarbij staat de relatie tussen het *aantal* neuronen en intelligentie overigens ter discussie. Gedacht wordt bijvoorbeeld dat niet het aantal, maar juist de mate van Interconnectie tussen neuronen bepalend is voor de mate van intelligentie. Mogelijk kan kunstmatige intelligentie hier leiden tot nieuwe inzichten, wanneer de hypothesen op kunstmatige intelligentie getest worden.

Een vuistregel die geldt bij machine learning is dat de prestatie van het model verbetert naar mate er meer data beschikbaar is voor het trainen van het model. Als gevolg van steeds verder gaande digitalisering van processen wordt veel data gegenereerd die hiervoor kan worden ingezet. In combinatie met alsmaar toenemende computerkracht en academische doorbraken [9] [10] groeien de ontwikkelingen op het gebied van machine learning sterk.

Deep learning heeft tal van toepassingen die menselijk presteren benaderen of zelfs voorbijstreven, zeker uitgedrukt in rekenkracht en snelheid. In de medische wereld wordt deep learning gebruikt voor het diagnosticeren van huidkanker en het classificeren van CT-scans. In zelfrijdende auto's wordt deep learning gebruik om op basis van camera's de richting en snelheid van de auto te bepalen. In de strijd tegen nepnieuws zet onder andere Facebook deep learning in om de authenticiteit van teksten te beoordelen.

Hoe ziet AI er over vijftig tot honderd jaar uit? In *science fiction* zien we toekomstbeelden waarin machines zich gedragen en voordoen als mensen, en in veel gevallen zelfs over superieure intelligentie beschikken. Een niveau van AI met een intelligentie gelijk aan die van mensen wordt *artificial general intelligence* (AGI) genoemd; is de AI slimmer dan de mens, dan spreken we van *artificial superintelligence* (ASI). Sommige futurologen zijn van mening dat AGI en met name ASI wel eens de laatste uitvinding van de mens zou kunnen zijn. Een AGI of ASI zou zelfs een bedreiging voor de mensheid kunnen zijn, daar we als mens de superieure intelligentie van een dergelijke AI niet meer kunnen bijhouden [11]. Dat daarbij ethische aspecten onder druk kunnen komen te staan is evident.

Over of (en wanneer) dit ooit zal plaatsvinden zijn de meningen verdeeld. Veel experts zijn het met elkaar eens dat deep learning in zijn huidige vorm waarschijnlijk niet bruikbaar is om menselijke intelligentie te vervaardigen [12] [13].

2.1.2 Toegevoegde waarde van AI

AI biedt in algemene zin een aantal kansen:

- **AI kan sneller, en soms betere, beslissingen nemen dan mensen.** Waar een mens soms enkele seconden tot minuten (afhankelijk van de hoeveelheid informatie die moet worden verwerkt) nodig heeft om een beslissing te nemen kan een machine learning model vaak in een fractie van een seconden duizenden datapunten verwerken. Een fraude detectie AI kan bijvoorbeeld in *real time* duizenden creditcard transacties monitoren en mogelijke frauduleuze transacties blokkeren.

- **Met AI kan (schaarse) expertkennis efficiënter worden ingezet.** Mensen genieten vaak enkele jaren een opleiding voordat ze de arbeidsmarkt betreden. Zelfs enkele jaren daarna zal een persoon nog niet op zijn top zijn. Expertkennis is daarom schaars en lastig op te schalen. Met machine learning kan de kennis van een expert worden gedestilleerd in een model en zo deze kennis breder worden ingezet. Zo heeft Google een AI-systeem ontwikkeld dat net zo goed als een ervaren oncoloog tumoren kan herkennen op CT-scans [14]. Een dergelijke AI-toepassing kan de schaarse expertkennis in bepaalde scenario's leveren en daarmee de 'echte' experts vrijmaken, zodat zij efficiënter ingezet kunnen worden.

- **AI is goed in repetitieve taken.** Dergelijke taken worden door mensen vaak als onbevredigend ervaren. Wanneer de taken goed omkaderd zijn is AI echter bij uitstek geschikt om deze over te nemen. AI hoeft niet te slapen, uit te rusten, of pauzes

te nemen, omdat ze zich niet gaat vervelen of moe wordt. Daarnaast lenen repetitieve taken zich goed voor AI, omdat er (wanneer de taak nu door mensen wordt uitgevoerd) waarschijnlijk voldoende data beschikbaar is om de AI te trainen.

2.2 Risico's bij de toepassing van AI

Ten opzichte van traditionele automatisering en (niet-AI) algoritmes beschikken AI-gebaseerde systemen over een aantal unieke, nieuwe eigenschappen. Deze eigenschappen kunnen, wanneer er onvoldoende rekening mee wordt gehouden, in generieke zin, risico's met zich meebrengen. Deze risico's zijn inherent aan de achterliggende werking van AI, en bestaan onafhankelijk van het toepassingsdomein. In deze paragraaf omschrijven we deze eigenschappen, hoe ze zich vertalen naar risico's, en illustreren we dit specifiek met voorbeelden in het toepassingsdomein telecom.

2.2.1 Gebrek aan verklaarbaarheid leidt tot onzekerheid in verdere besluitvorming

Een AI-systeem vertaalt, in het algemeen, een set invoervariabelen ('inputs') tot een bepaalde uitkomst ('output'). Bij op deep learning gebaseerde AI-systemen is de wijze waarop een bepaalde uitkomst tot stand komt niet evident. Hierdoor is het niet altijd duidelijk hoe een beslissing tot stand is gekomen, is het ingewikkeld om bepaalde handelingen te verifieren, en kunnen fouten onopgemerkt het systeem binnensluipen.[5] Ook de legitimatie van beslissingen die op basis van de AI-toepassing worden genomen, kan in het geding komen.

Voorbeeld in de telecomsector: Covariance shift

Een *anomaly detection*-systeem kan gebruikt worden om verdacht verkeer in een netwerk op te sporen en te blokkeren. Een dergelijk systeem is getraind om afwijkende patronen te herkennen. Stel nu dat er een nieuwe browser op de markt wordt geïntroduceerd, en die past een nieuw protocol toe om HTTPS-verzoeken beter te stroomlijnen. Als gevolg vindt er een *covariance shift*[6] plaats. Het anomaly detection systeem heeft in het verleden, tijdens de training, nog geen verkeer gezien dat via deze browser verloopt. Hierdoor wordt het verkeer van deze browser opgemerkt als 'afwijkend' en geblokkeerd in het netwerk.

2.2.2 Onvoorspelbaarheid leidt tot een gebrek aan vertrouwen

Om AI-systemen bepaalde beslissingen toe te kunnen vertrouwen is het vaak van belang dat het AI-systeem voorspelbaar is in haar gedrag en uitkomsten [15]. Er zijn minstens twee redenen waarom een AI-systeem (ondanks het feit dat het gaat om een stuk software, waarvan de uitvoering volledig kan worden getraceerd) toch onvoorspelbaar gedrag kan vertonen:

- **Het algoritme is te complex, voor een mens, om het gedrag te kunnen begrijpen.** Hoewel het conceptueel voor te stellen is dat een AI-model bestaat uit een groot aantal lagen met functies daartussen, is het exacte gedrag van een model nauwelijks tot niet meer te begrijpen of te traceren bij de zeer grote modellen. Op

[5] Hierbij moet de kritische noot worden geplaatst dat veel telecominfrastructuren zonder AI al een complex systeem zijn, en men zich kan afvragen in welke mate er nu al niet- (of niet-volledig) verklaarbare keuzes worden gemaakt.

[6] Een covariance shift kan plaatsvinden wanneer de aard van de data verandert. Dit houdt in dat de data waar een model in eerste instantie op getraind is niet meer representatief is.

dit moment zien we AI-modellen in gebruik met honderden miljoenen parameters [16]

- **Het AI-algoritme is niet deterministisch.** Veel 'gewone' algoritmes zijn deterministisch: voer je ze twee keer uit met dezelfde invoerparameters, dan komt er twee keer hetzelfde resultaat uit. Dit is echter niet bij alle algoritmes het geval: sommige algoritmes maken gebruik van willekeur, zoals Bayesiaanse methoden [17]. In deze modellen zijn de parameters geen vaste waarden, maar distributies van kansen. Wanneer een voorspelling wordt gemaakt met deze modellen worden er willekeurig steekproeven getrokken uit deze distributies, en is de kans zeer klein dat er twee keer dezelfde steekproef wordt getrokken. AI-algoritmes behoren vaak tot deze laatste categorie, of hebben vergelijkbare, niet-deterministische eigenschappen. Wanneer het algoritme - en daarmee het resultaat - niet deterministisch is, is de toepassing lastiger te begrijpen en te controleren.

- **Het algoritme is volgorde-gevoelig.** Sommige AI-modellen (o.a. *recurrent neural networks* of RNN's) werken op basis van 'streaming' data – er worden continu gegevens ingevoerd, en er volgt doorlopend een uitkomst. Dergelijke modellen zien we bijvoorbeeld veel terug bij het begrijpen van taal. Dat is logisch, aangezien de betekenis van een woord vaak afhangt van de omliggende woorden: er is een *volgorde-effect*. Dankzij dit volgorde effect kan echter een bepaalde invoer (c.q. een woord) anders worden geïnterpreteerd afhankelijk van de eerdere en latere invoer. De modellen hebben een 'geheugen' welke het resultaat beïnvloedt [18].

Voorbeeld in de telecomsector: Niet-deterministische systemen

Om virtuele netwerkcomponenten op de fysieke infrastructuur te 'embedden' wordt een AI-systeem gebruikt om de beste configuratie te bepalen. Er zijn echter ontelbare configuraties te verzinnen en het is te kostbaar om via *brute force* de beste configuratie te achterhalen. Geïnspireerd door schaakcomputers passen ingenieurs daarom *Monte Carlo Tree Search* toe om de beste configuratie te benaderen [19].

Toepassing van deze methode zal, vanwege het gebruik van sampling, nooit de garantie bieden dat twee keer dezelfde configuratie zal worden gesuggereerd, gegeven dezelfde omgevingsfactoren. Testen kan dus slechts tot een bepaalde hoogte zekerheid geven in hoe het model zich in verschillende situaties zal gedragen. Omdat de beste configuratie op voorhand ook niet bekend is kan er ook niet (door een mens) worden geverifieerd of het AI-systeem inderdaad de beste configuratie wist te bepalen. Er kan uiteraard wel worden vergeleken tussen oplossingen van verschillende systemen of met uitkomsten van meer traditionele optimalisatie-algoritmen om te beoordelen hoe 'goed' of 'slecht' de uitkomst is (de uitkomst van de AI zou bijvoorbeeld terzijde kunnen worden gelegd wanneer deze slechter is dan de uitkomst van een traditioneel algoritme).

2.2.3 Wie is aansprakelijk wanneer het fout gaat?

Een AI-systeem is juridisch gezien geen rechtspersoon, maar zal moeten worden gezien als een gereedschap (net als een computer) waarmee een rechtspersoon een handeling verricht. [20] Het gebruik van AI-systemen neemt de verantwoordelijkheid dan ook niet weg van de persoon die besluit zijn of haar taak uit te besteden. Bij de ontwikkeling van dergelijke systemen zijn doorgaans verschillende actoren betrokken die elk een eigen bijdrage leveren. Hierbij heerst er onduidelijkheid wie exact de verantwoordelijkheid heeft wanneer er taken worden geautomatiseerd.

Ook is het de vraag of de persoon of een rechtspersoon die gebruik maakt van een AI-systeem voldoende weet over het systeem om de verantwoordelijkheid erover te kunnen aanvaarden. Hier ligt waarschijnlijk een regulerende en/of toezichthoudende taak van de overheid: hoewel veel AI-systemen zijn ontworpen om (letterlijk) de mens de handen aan het stuur te laten houden, is het maar zeer de vraag of deze persoon ook in staat is om te herkennen wanneer de AI in de fout gaat, en tijdig te reageren. Hierbij speelt mogelijk ook een effect van gewenning, wanneer een observerend persoon de AI steeds goede keuzes ziet maken.

Voorbeeld in de telecomsector: Waar ligt de schuld?

Een AI-systeem is getraind door een leverancier en wordt ingezet door een operator. De leverancier heeft het AI-systeem getraind op basis van gegevens uit andere telecomnetwerken. Waar ligt de schuld wanneer er, op basis van het AI-systeem, een verkeerde beslissing wordt gemaakt? In sommige gevallen kan dit vastgelegd worden in SLA-overeenkomsten. AI kan echter ook effect hebben op kwaliteitsindicatoren (*key performance indicators:* KPI's) die niet in deze contracten zijn vastgelegd. In deze gevallen kan het onduidelijk zijn wie verantwoordelijk is.

2.2.4 Autonome systemen kunnen worden misbruikt

De uitkomsten van een AI-systeem zijn afhankelijk van de invoerwaarden. Wanneer het mogelijk is deze input te variëren kan het mogelijk zijn om de uitkomst te veranderen. In een deterministisch algoritme kan worden beredeneerd hoe een bepaalde verandering in de input doorwerkt in de output. Bij AI-systemen is dat, zoals hierboven al aangegeven, over het algemeen veel lastiger of zelfs onmogelijk. AI-systemen zijn derhalve kwetsbaar voor zogenaamde *adversarial attacks* [21]. Een adversarial attack is gebaseerd op gemanipuleerde data die voor mensen niet te onderscheiden is van legitieme data. De data kunnen zó worden gemanipuleerd dat een AI een verkeerd besluit neemt, zonder dat mensen kunnen herkennen waar de fout zit. [22]

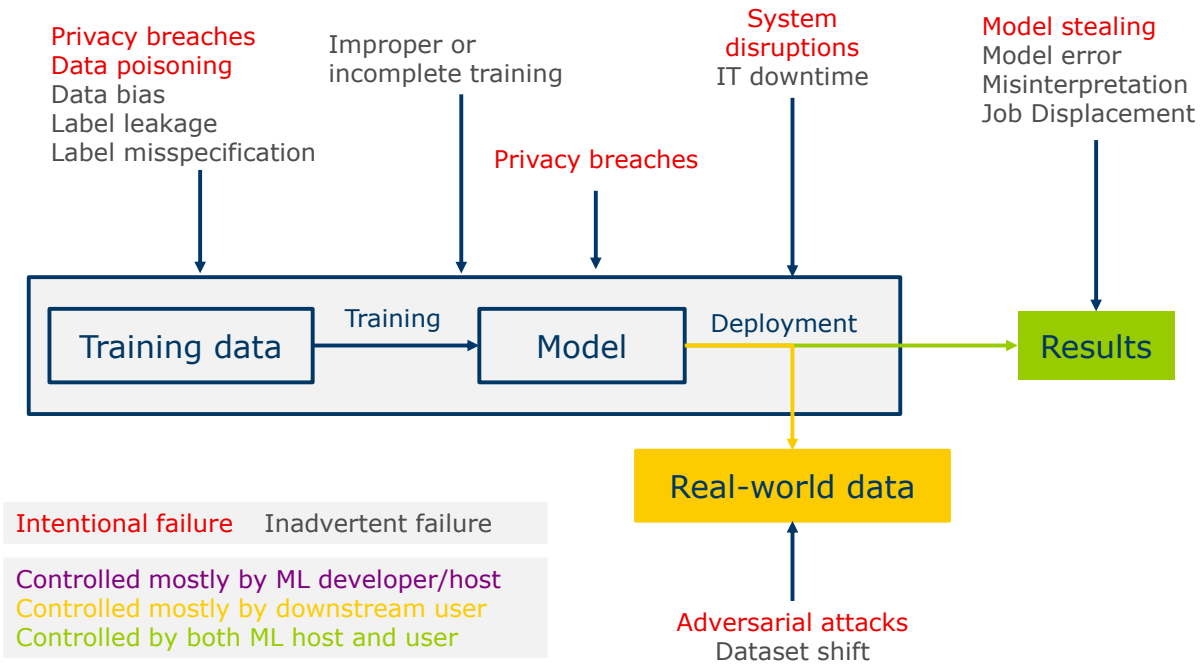
Adversarial attacks worden vaak gegenereerd door andere AI-systemen. Hierdoor kunnen aanvallen op AI-systemen vaak geautomatiseerd plaatsvinden. In theorie zijn er weinig AI-systemen die niet voor de gek te houden zijn door een *threat actor* met genoeg rekenkracht.

Voorbeeld in de telecomsector: Adversarial ransomware

Een anti-malwaresysteem herkent, op basis van AI, bestanden met onveilige inhoud. Een kwaadwillende partij beschikt ook over dit systeem, en heeft het gebruikt om een eigen AI te trainen (de "adversarial AI"), die bestanden kan genereren die onveilig zijn (bijvoorbeeld ransomware bevatten), maar niet door het anti-malwaresysteem worden herkend. Het anti-malwaresysteem pikt het gevaar in deze bestanden niet op, en zo kan de ransomware zich verspreiden binnen het netwerk [23]. Hierdoor raakt een groot deel van de bestanden in de telecomaanbieder versleuteld, en wordt de aanbieder geconfronteerd met afpersing.

Naast het manipuleren van de invoergegevens, is manipulatie ook mogelijk in andere stappen van de AI-levenscyclus. Figuur 4 geeft dit schematisch weer. In de trainingsfase bestaan mogelijkheden om, door manipulatie van de trainingsgegevens, de AI zich (uiteindelijk) anders te laten gedragen. Manipulatie van het model en/of de opbouw daarvan kan leiden tot

blootleggen van gegevens uit de trainingsdata. De eerder besproken *adversarial attacks* bevinden zich in de fase waarin de AI daadwerkelijk in gebruik is ('productiefase').



Figuur 4 Mogelijke aanvallen tegen een AI-systeem in de levenscyclus daarvan. 'ML' verwijst in de afbeelding naar 'machine learning' [24]

2.3 Huidige toepassingen van AI in telecom

In het kader van dit onderzoek is een verkenning uitgevoerd naar huidige en toekomstige toepassingen van AI in telecom. De verkenning is gebaseerd op bureauonderzoek (met name het vinden van casestudies en whitepapers over productaanbod) en gesprekken met telecomoperators en -leveranciers.

Op hoofdlijnen wordt AI in telecominfrastructuren op dit moment ingezet voor (1) configuratie, planning en optimalisatie van (het functioneren van) de netwerken, (2) onderhoud van het netwerk. Hieronder geven we een overzicht van huidige toepassingen van AI in telecominfrastructuren die geïdentificeerd zijn binnen dit onderzoek.

2.3.1 Optimalisatie van telecominfrastructuur

De telecomsector heeft een aantal fasen van automatisering gekend. Waar vroeger nog verbindingen met de hand werden gemaakt door het omschakelen van kabels werd dit werk geautomatiseerd door hardware. Nu zien we dat deze functies niet meer door specifieke hardware wordt vervuld maar virtueel via software worden gedefinieerd.

In het optimaliseren van het netwerk wordt gebruik gemaakt van heuristieken bedacht door mensen; denk aan heuristieken die bepalen hoe belangrijk of dreigend een datapakket is dat moet worden gerouteerd, of het bepalen van de wijze waarop radioresources in een mobiel netwerk worden toegewezen. Algoritmische optimalisatie (van configuratie) van functies in telecomnetwerken vindt eveneens al een geruime tijd plaats. Het gaat daarbij om algoritmes die werken op basis van (onder andere) traditionele wiskundige methoden voor optimalisatie. Sommige van deze technieken zijn al enkele honderden jaren bekend. Zo wordt de Newton-Rhapson-methode (1690) gebruikt om het optimum van een wiskundige functie te bepalen aan de hand van de afgeleide. De kleinste kwadratenmethode ontwikkeld door Gauss (1735)

ligt aan de basis van het oplossen van regressieproblemen. Methoden zoals lineair programmeren (1939) worden gebruikt om een systeem met randvoorwaarden te optimaliseren.

In telecomnetwerken wordt veel data gegenereerd, en deze data komt steeds sneller op één centrale plek beschikbaar voor analyse. Ter illustratie: de Indiase mobiele operator Reliance Jio genereerd dagelijks 4 tot 5 petabyte aan data uit de operatie van het netwerk. [25] Deze data zijn bij uitstek geschikt voor analyse en optimalisatie.

Bij machine learning vinden we een nieuwe generatie algoritmes die gebruikt kunnen worden voor vergelijkbare doeleinden als 'traditionele' optimalisatietechnieken. De beschikbaarheid van grote hoeveelheden data maakt inzet van machine learning in telecominfrastructuren mogelijk. Machine learningmethoden zijn niet gebonden aan veel van de beperkingen van traditionele technieken. Zo kunnen traditionele methoden doorgaans alleen lineaire functies benaderen, terwijl met deep learning theoretisch gezien elke continue functie gemodelleerd kan worden (het *Universal Approximation Theorem* [26]). Daarnaast vereisen traditionele methoden, in tegenstelling tot machine learning, doorgaans aannames over onderliggende distributies (zoals een normaalverdeling) en over de onafhankelijkheid van input parameters. Machine learning systemen kunnen echter zonder deze aannames werken.

De volgende AI-toepassingen worden op dit moment toegepast of ontwikkeld voor telecominfrastructuren:

- *Power management:* Machine learning wordt ingezet om stroombesparingen te realiseren in mobiele netwerken. Op basis van meteorologische data, het aantal gebruikers en de positie van die gebruikers, passen antennes actief hun stralingspatroon, richting en stralingssterkte aan naar de vraag. Dit resulteert in energiebesparing, bijvoorbeeld tijdens nachtelijke uren wanneer de datavraag relatief laag is, en in een efficiënter gebruik van de basisstations, omdat er een groter oppervlak bedient kan worden bij opstelpunten waar vraag naar capaciteit niet uniform is.

- *Radio-optimalisatie:* Op dit moment wordt machine learning gebruikt om de datadoorstroom van en naar een basisstation in een mobiel netwerk te optimaliseren. De afstand tot gebruikers, het aantal verbonden gebruikers, en bepaalde omgevingsfactoren bepalen daarbij de *radioparameters,*[7] en die bepalen dan weer maximale hoeveelheid data die per hoeveelheid spectrum per tijdseenheid verzonden kan worden (met als afweging dat met een mogelijk efficiëntere modulatie, klanten met een zwakker signaal weer niet bediend kunnen worden). Ook interferentie speelt een rol: radioresources kunnen gecoördineerd worden tussen micro- en macrocel. Om de efficiëntie te maximaliseren worden (nu al) algoritmes ingezet om dynamisch te bepalen welk deel van het spectrum voor welke gebruiker met welke parameters ingezet dient te worden. De parameters van deze algoritmes kunnen worden 'getuned' met behulp van een AI [27].

- *Quality of Transmission (QoT) estimation*: In optische verbindingen kan het signaal verstoord of onderbroken worden. Machine learning wordt toegepast om op voorhand in te schatten hoe goed de transmissie zal verlopen over een verbinding. Op basis van onder meer de kabellengte, andere signalen binnen de kabel, leeftijd van de apparatuur wordt het beste pad uitgerekend. Op basis van deze inschatting wordt het verkeer gerouteerd. Het is ook denkbaar dat dergelijke algoritmes in draadloze

[7] Onder andere de modulatievorm (in LTE/5G: QPSK, 16-QAM, 64-QAM, 256-QAM, etc.) en de hoeveelheid bits die wordt ingezet voor foutcorrectie.

netwerken worden toegepast en daarbij bijvoorbeeld bepalen hoeveel foutcorrectie of redundantie (bijvoorbeeld hertransmissie) wordt gehanteerd.

- *Optical network signal amplification*: In optische netwerken kan op verschillende punten degradatie van het signaal plaatsvinden. Op dit moment worden verschillende AI-technieken toegepast (zoals QoT estimation) ter identificatie van punten en momenten waarop degradatie kan plaatsvinden. Hier wordt strategisch versterking van het signaal uitgevoerd om ruis tijdens transmissie te minimaliseren.

- *Path computation:* Om de beste route tussen twee nodes in een netwerk te bepalen zijn in het verleden verschillende algoritmes ontwikkeld die bepaalde heuristieken toepassen. Zo worden algoritmes zoals A* of Dijkstra traditioneel gebruikt om het *kortste* pad te bereken. Er wegen echter meer factoren mee bij het bepalen van het meest optimale pad, en die zijn lastig te integreren in traditionele algoritmes. Machine learning kan daarentegen wel rekening houden met zaken zoals congestie en bottlenecks in het netwerk en zo een betere inschatting maken van de optimale routes.

- *Self-organizing networks*: Op basis van beschikbare informatie kunnen netwerkcomponenten zich in beperkte mate automatisch configureren. Zo kan een mobiel basisstation zelf uitzoeken welke andere basisstations in de buurt zijn, en zo automatisch "neighbour relations" bepalen. Dergelijke functionaliteit kan worden gebruikt om snel onderdelen van een netwerk in te richten. In de praktijk zien we dat de functionaliteit daarna echter wordt uitgeschakeld, waarna er met de hand nog "fine tuning" plaatsvindt. De functionaliteit blijkt over het algemeen nog te instabiel om aan te laten staan voor dynamische herconfiguratie.

2.3.2 Onderhoud aan telecominfrastructuur

Ten behoeve van het onderhoud van telecomnetwerken zien we de volgende toepassingen in gebruik:

- *Performance monitoring:* Het monitoren van signalen binnen een optisch transmissienetwerk is essentieel om storingen te kunnen detecteren. Dit wordt doorgaans gedaan door het meten van verschillende parameters, zoals *optical signal to noise ratio (*OSNR), *nonlinearity factors*, *chromatic dispersion* (CD) en *polarization mode dispersion* (PMD). Door het monitoren van deze variabelen kunnen tijdig problemen in het netwerk worden vastgesteld. Met machine learning kan beter worden ingeschat bij welke combinatie van waarden de kans op een storing toeneemt, en wanneer er het best kan worden ingegrepen.

- *Predictive maintenance:* Het uitvallen van apparatuur is zeer kostbaar, zowel in termen van reparatiekosten als in gederfde omzet en claims. Veel factoren zijn van invloed op slijtage van apparatuur en onderdelen, waaronder het weer, de gemiddelde intensiteit van het gebruik, en het type onderdeel. Om netwerkuitval te minimaliseren worden modellen gebruikt op basis van machine learning die kunnen voorspellen wanneer uitval te verwachten is. Er kan vervolgens preventief onderhoud worden uitgevoerd.

2.4 Toekomstige toepassingen van AI in telecom

In het onderzoek zijn (op basis van het eerdergenoemde bureauonderzoek en gesprekken, aangevuld met gesprekken met experts) een aantal toekomstige toepassingen van AI in telecominfrastructuren geïdentificeerd.

2.4.1 Optimalisatie van telecominfrastructuur

In de toekomst worden een aantal nieuwe vormen van optimalisatie voorzien die gebaseerd zijn op machine learning:

- *Smart handovers:* In mobiele netwerken kan het signaal sterk afnemen wanneer de afstand tot het basisstation toeneemt of er fysieke barrières bestaan tussen de ontvanger en het basisstation (*path loss, penetration loss*). In een klassiek mobile netwerk moeten handovers dit probleem oplossen, maar ook die aanpak kent haar beperkingen. Eén van de toekomstige oplossingen om dit probleem beter te mitigeren is *Multi Tower Beamforming (*ook bekend als "coordinated multipoint" of CoMP). Hierbij wordt het signaal gefocust op een apparaat door middel van combinatie van gecoördineerde signalen van meerdere basisstations. Deze modellen kunnen gebaseerd zijn op AI en leveren dan naar verwachting betere prestaties dan modellen die dat niet zijn.

- *Network orchestration*: Door *Software Defined Networking* (SDN) en *Networking Function Virtualisation* (NFV) is het eenvoudiger om op één centrale plek een netwerk te definiëren en data uit verschillende deelsystemen te combineren. Deze data kunnen vervolgen als input fungeren voor machine learning-modellen die de samenwerking tussen de netwerkfuncties kan optimaliseren. Diverse partijen zijn bezig met het vormgeven van een op AI-gebaseerd telecomnetwerk. Hierbij worden uit alle netwerkelementen grote hoeveelheden data verzameld (denk aan meetgegevens, verkeersgegevens, et cetera). Deze gegevens worden in een model verwerkt tot een set met configuratieparameters voor diezelfde netwerkelementen. Op basis van veranderingen in het netwerk of het gebruik daarvan kan de AI het netwerk snel herconfigureren.

- *Optical network nonlinearity mitigation:* In glasvezelnetwerken kan ruis optreden door de non-lineariteiten. Machine learning kan worden ingezet om het signaal op te schonen voor verdere verwerking. Hierdoor wordt een hogere capaciteit behaald.

2.4.2 AI-gebaseerde telecomfuncties

We zien een aantal nieuwe netwerkfuncties die kunnen worden gerealiseerd op basis van AI:

- *Bandwidth slicing / Resource allocation:* In telecomnetwerken maken verschillende toepassingen gebruik van dezelfde infrastructuur en beschikbare bronnen. De eisen van de verschillende toepassingen zijn echter verschillend. Door het efficiënt inzetten van bronnen (denk aan spectrum en 'resource blocks' in mobiele netwerken) kan optimaal tegemoet worden gekomen aan de behoeften van de verschillende toepassingen. Zo kunnen toepassingen die lage latency vereisen beter werken naast toepassingen die hoge capaciteit vragen. Het netwerk 'herkent' de lage latency-eis en behandelt het verkeer vervolgens anders dan het capaciteitsgedreven verkeer. Machine learning wordt ingezet voor zowel het herkennen van verkeersstromen als het optimaliseren van de verdeling van resources.

- *Virtual topologies (VT):* Bij virtuele typologieën zijn niet alleen de functies van het netwerk gevirtualiseerd, maar ook de algehele topologie van het netwerk. Hierdoor kan dynamisch worden bepaald welke transmissiepaden (lichtpaden) moeten worden

gebruikt of waar mogelijk extra capaciteit moet worden aangebracht (bijvoorbeeld in de vorm van meer verbindingen, of in de vorm van een extra lokaal datacenter). Machine learning-algoritmes worden ingezet om de optimale parameters van de topologie te bepalen. Door de hoge snelheid van besluitvorming kunnen mensen deze continue configuratietaak niet overnemen.

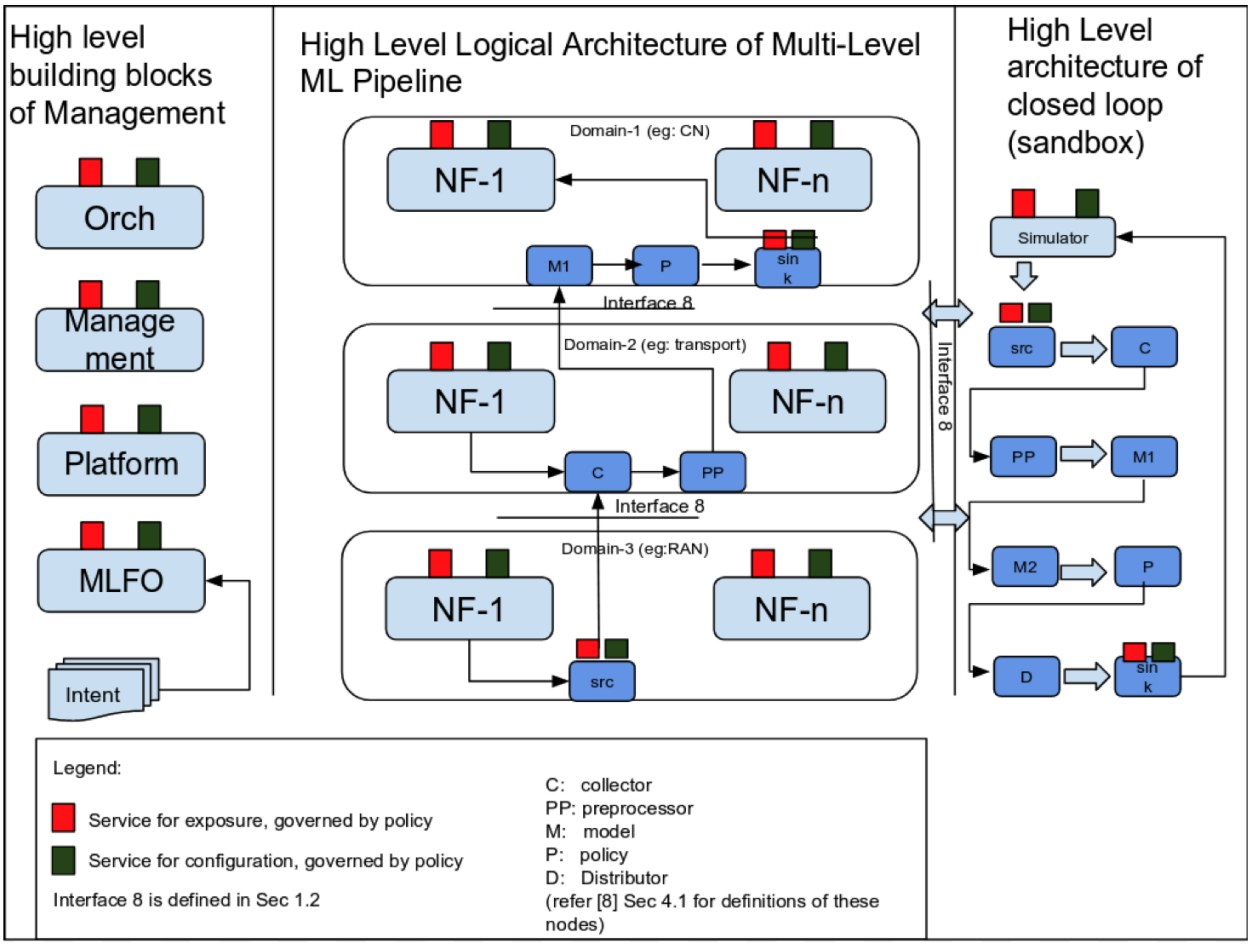
- *Anomaly detection / malicious traffic detection:* hier wordt AI ingezet om normaal gedrag te modelleren binnen telecominfrastructuren. Vervolgens worden voorspellingen van het gedrag van de infrastructuur vergeleken met de realiteit: wanneer de afstand tussen de echte waarde en de voorspelde waarde groot is, dan wordt een gebeurtenis aangemerkt als afwijkend. Door anomaly detection te trainen op grote logbestanden kan deze technologie in vrijwel alle punten in het netwerk worden toegepast om afwijkende gebeurtenissen te identificeren.

- *Dynamische spectrumtoewijzing: Vergunningsplichtig* spectrum wordt op dit moment voornamelijk statisch verdeeld tussen operators. In systemen als het Amerikaanse CBRS[8] is dynamischere, geautomatiseerde toewijzing mogelijk. Gebruikers 'luisteren' eerst welke signalen zij kunnen ontvangen (*sensing*), en/of raadplegen database (welke frequentieblokken zijn gereserveerd?). Zij kunnen vervolgens voor vrije delen (geautomatiseerd) een aanvraag tot spectrumgebruik doen. Op basis van algoritmes zou spectrum dynamischer en tussen operators (en andere gebruikers) worden toegewezen. Tijdens de *DARPA spectrum collaboration challenge* is deze technologie met succes uitgetest [28]. Bij de uitgifte van (vergunningen tot gebruik van) frequenties zou deze dynamische spectrumtoewijzing kunnen worden toegepast.

Vandaag de dag wordt AI nog voornamelijk toepast voor micro-optimalisaties binnen specifieke componenten of functies. In de toekomst zal, als de huidige ontwikkelingen zich blijven doorzetten, AI naar verwachting een meer centrale rol innemen in telecomnetwerken. Deze ontwikkeling sluit aan bij de meer algemene trend van (netwerk)virtualisatie. [29] Door netwerken centraal te besturen, en de onderliggende infrastructuur 'weg te abstraheren', ontstaat een hogere mate van flexibiliteit ten aanzien van de inrichting van het netwerk. AI kan worden toegepast om deze centrale aansturing optimaal uit te voeren, en krijgt daarmee een meer *holistische, besturende* rol binnen telecominfrastructuren.

Figuur 5 hieronder toont de architectuur van een gevirtualiseerd netwerk dat door een AI kan worden bestuurd, zoals gedefinieerd door een werkgroep van ITU rond dit thema. [30] In het netwerk bevinden zich diverse netwerkfuncties ('NF'). Deze functies genereren data en maken deze beschikbaar voor het trainen van een AI (op basis van machine learning, hier afgekort als 'ML'). De netwerkfuncties ondersteunen daarnaast het automatisch aansturen door een AI.

[8] Citizens Broadband Radio Service. Zie o.a. [wikipedia.org] voor een korte toelichting.

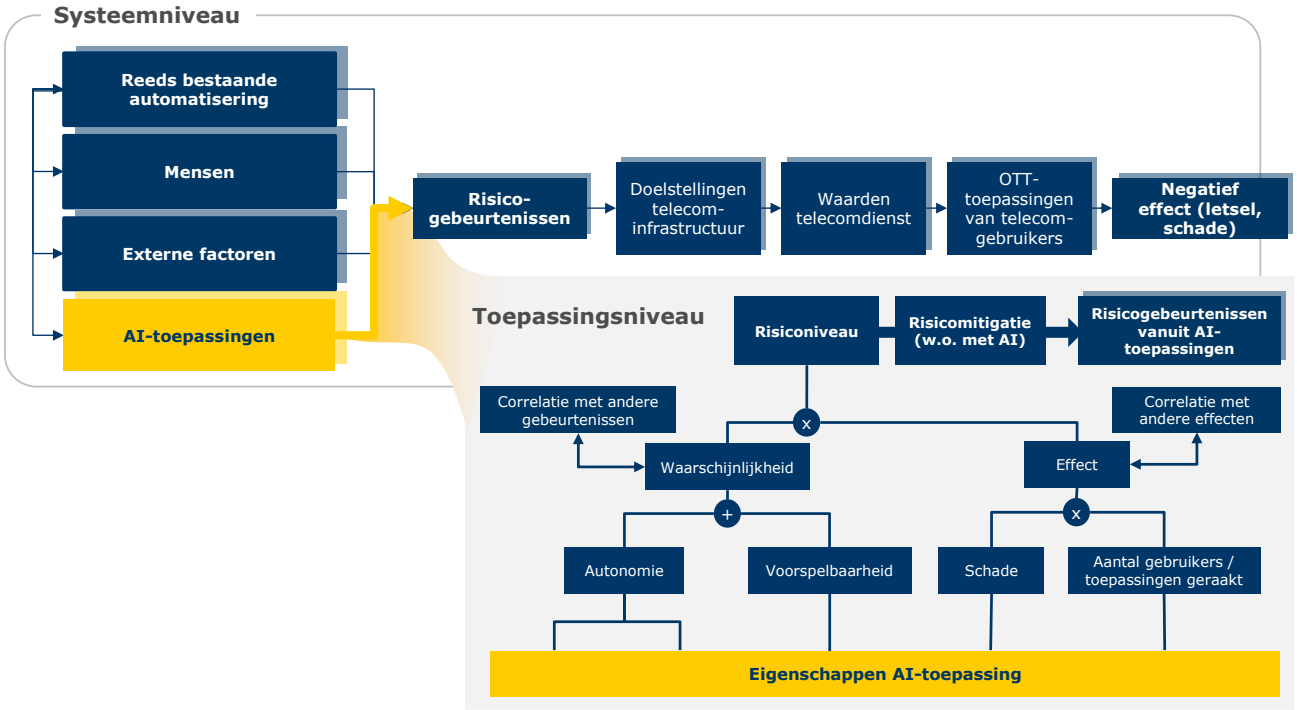


Figuur 5 Architectuur voor een AI-gebaseerd telecomnetwerk uit een rapport van de ITU-focusgroep "machine learning for future networks including 5G". [30, p. 16 Fig. 2]

3 Risico's van AI-toepassingen in telecominfrastructuren

In dit hoofdstuk presenteren we een model waarmee de risico's van het gebruik van AI-toepassingen in de telecomsector kunnen worden geanalyseerd. Dat wil zeggen dat het model een kader biedt waarmee, op basis van specifieke eigenschappen van de toepassing, een kwalitatief risiconiveau kan worden vastgesteld. De toezichthouder kan op basis van deze inschatting bepalen of het risico aanvaardbaar is, of dat er mitigatiemaatregelen moeten worden getroffen.

Bij het inschatten van risico's van AI in telecominfrastructuur maken we onderscheid tussen het *systeemniveau* (risico dat een telecominfrastructuur als geheel loopt) en het *toepassingsniveau* (risico's van een individuele AI-toepassing binnen een specifiek deel van de infrastructuur). Figuur 6 toont het in dit onderzoek ontwikkelde risicomodel. Bovenaan wordt het 'systeemniveau' geschetst: mensen, externe factoren en toepassingen (waaronder AI-toepassingen) kunnen leiden tot risicogebeurtenissen, die (uiteindelijk) leiden tot negatieve effecten en, bijgevolg, het verdwijnen van vertrouwen van bedrijven en burgers in telecominfrastructuur en daarop gebaseerde toepassingen. In dit onderzoek kijken we specifiek naar AI-toepassingen. Op systeemniveau gaat het daarbij om de inbedding van AI-toepassingen. Op het toepassingsniveau kijken we naar de relatie tussen de eigenschappen van deze toepassingen en de kans en impact van risicogebeurtenissen. De beide niveaus worden in de paragrafen verderop in dit hoofdstuk nader uitgewerkt.



Figuur 6 Modellering van risico's van AI in telecom: systeemniveau en toepassingsniveau

3.1 Systeemniveau

3.1.1 Theoretisch kader

Een risico is, in brede zin, een negatieve gebeurtenis die met een bepaalde waarschijnlijkheid kan plaatsvinden. Hoewel het intuïtief duidelijk is (of lijkt) wat risico's zijn, is er allesbehalve een eenduidige definitie te geven voor de telecomsector. Het is zelfs de vraag of risico's objectief zijn in te schatten, of dat er noodzakelijkerwijs subjectieve aannames en keuzes aan gekoppeld zijn [31].

Risico's kunnen, tot op zekere hoogte, op voorhand worden ingeschat. Daarmee is niet gezegd dat zo'n inschatting altijd correct, *volledig en objectief* is, of zelfs zou kúnnen zijn. Het is van belang te realiseren dat niet alle risico's *kenbaar* zijn. Ter illustratie: een vliegtuigbouwer zou het risico op neerstorten kunnen baseren op de faalkans van individuele componenten en de effecten wanneer dat falen optreedt. De vliegtuigbouwer moet echter ook rekening houden met het gelijktijdig falen van componenten. Deze risico's zijn kenbaar, maar kunnen natuurlijk 'gemist' worden. Daarnaast zijn er risico's die de vliegtuigbouwer niet *kan* inschatten: zo zou achteraf kunnen blijken dat een component gevoelig blijkt voor straling, terwijl de vliegtuigbouwer dit aspect niet had geïdentificeerd of logisch had kunnen afleiden. Een laatste categorie, de *bekende onkenbare risico's,* bestaat uit de risico's waarvan wel bekend is dat ze *kunnen* bestaan, maar waarvan de precieze omvang niet in te schatten is (Figuur 7). In de context van dit onderzoek is het belangrijk te realiseren dat de hier genoemde modellering alleen *kenbare* risico's kan betreffen.

	Kenbaar	**Onkenbaar**
Onbekend	Onbekende kenbare risico's	Onbekende onkenbare risico's
Bekend	Bekende kenbare risico's	Bekende onkenbare risico's

Figuur 7 Vier categorieën risico's

In deze studie hanteren we de definitie voor risico zoals voorgesteld in [31]: Een (kenbaar) risico bestaat uit een potentiële *risicogebeurtenis* (scenario), de *waarschijnlijkheid* dat deze plaatsvindt en het (negatieve) *effect* dat deze gebeurtenis heeft. Hoe hoger het product van waarschijnlijkheid en effect, hoe groter het risico. [32] Afhankelijk van de methode wordt een andere weging of vermenigvuldiging gebruikt (zie bijvoorbeeld figuur 5.7 in [31]). Het verlagen van de waarschijnlijkheid en/of het verlagen van de impact zijn dan logischerwijs manieren om het risico te *mitigeren*. Andersom kan het risico worden *aanvaard* wanneer de negatieve gebeurtenis zeer onwaarschijnlijk is, en/of wanneer de negatieve effecten klein of acceptabel zijn.

3.1.2 Maatschappelijke risico's

Telecominfrastructuren vervullen een belangrijke maatschappelijke functie en zijn aangemerkt als vitaal. [33] Steeds meer diensten worden digitaal geleverd, en zijn daarmee afhankelijk geworden van goed werkende, betrouwbare en altijd beschikbare telecominfrastructuur. Hierdoor hebben gebruikers bepaalde verwachtingen bij het gebruik van telecomnetwerken. Wanneer niet aan de verwachtingen wordt voldaan kan dit negatieve gevolgen hebben. Hieronder gaan wij in op welke maatschappelijke doelstellingen een telecomnetwerk vervult en hoe AI hier invloed op kan hebben.

Vanuit het perspectief van het telecomsysteem als geheel zou dan ook gekeken moeten worden naar *maatschappelijke effecten*. Wanneer toepassingen gebruik maken van telecominfrastructuur en deze niet goed kunnen werken als gevolg van (bijvoorbeeld) uitval levert dit de maatschappij wel degelijk kosten op. In missiekritische situaties kan er zelfs sprake zijn van letsel. Een voorbeeld is de 'noodknop' die aanwezig is op alle C2000-portofoons die door politie en brandweer worden gebruikt. Werkt die knop niet, dan kan een agent in nood zijn collega's niet op tijd oproepen, en slachtoffer worden in een gevaarlijke situatie. [34]

Wat zijn dan de relevante doelparameters waarop een eventueel risico betrekking kan hebben in telecominfrastructuren? In de dissertatie van Vriezekolk [31] wordt een modellering van telecomsystemen gehanteerd die is gebaseerd op *diensten* welke bestaan uit een netwerk van *nodes* en *links*, aan welke beide een set risicogebeurtenissen wordt gekoppeld. In het betreffende onderzoek wordt specifiek gekeken naar de effecten van *uitval* van de telecomdiensten als gevolg van deze gebeurtenissen. [31] Naast uitval zien we een aantal andere doelstellingen van telecominfrastructuur die (gezien de onderzoeksvraag) eveneens meegenomen dienen te worden in de analyse. We bespreken deze hieronder.

Beschikbaarheid van netwerken

De samenleving leunt steeds meer op de beschikbaarheid van telecomnetwerken. Telecom wordt een kritieke infrastructuur die een zo hoog mogelijke beschikbaarheid vereist. De toepassing van AI kan die beschikbaarheid vergroten, maar kan deze ook negatief beïnvloeden. Wanneer een AI-systeem faalt kan dit soms grote delen van het netwerk platleggen. Wanneer fouten zich van systeem A naar systeem B propageren kan dit een kettingreactie veroorzaken.

Voorbeeld: Kettingreactie

In een mobiel netwerk wordt op basis van een AI een "power management" procedure uitgevoerd, waarbij wordt bepaald welke frequentiebanden op een basisstation worden ingeschakeld. [35] Zijn er weinig gebruikers, dan wordt het aantal banden teruggebracht om energie te besparen. Het kan voorkomen dat voor een langere periode er geen gebruikers in de omgeving zijn waardoor het model het basisstation volledig uitschakelt. Dit heeft als gevolg dat een systeem om automatisch "neighbour relations" te bepalen (paren van basisstations waartussen een terminal kan bewegen in het netwerk) faalt. Voor andere basisstations lijkt het alsof er minder verkeer is. Hierdoor worden er meer basisstations uitgeschakeld, en propageert de fout zich door het netwerk.

Integriteit van informatie

Bij integriteit gaat het om de juistheid en betrouwbaarheid van informatie. AI kan in sommige toepassingsgebieden hier een negatieve impact op hebben. In een informerende rol kan AI de integriteit van informatie aantasten. Een AI-systeem kan mogelijk ruis uit een ander systeem verwerken tot onjuiste informatie.

Voorbeeld: AI-gegenereerde informatie zorgt voor obfuscatie

Een *predictive maintenance*-systeem maakt gebruik van vochtsensoren in de grond om te voorspellen wanneer corrosie plaatsvindt in kabels. Een kapotte sensor levert continu dezelfde data. Het AI-model, dat niet is voorbereid op de situatie "kapotte sensor", interpreteert deze data alsof de sensor werkt. Hierdoor wordt een corroderende kabel te laat opgemerkt. Een oplossing zou kunnen zijn om de sensoren redundant uit te voeren, of om meer (verschillende) informatie mee te nemen in het model.

Vertrouwelijkheid van gegevens

Telecomaanbieders hebben te maken met grote hoeveelheden gevoelige informatie; niet alleen de informatie die klanten over het netwerk uitwisselen, maar ook (meta)gegevens over klanten en dit verkeer. AI-toepassingen in telecominfrastructuren kunnen (deels) worden getraind op basis van deze (gevoelige) gegevens. Het is denkbaar dat een kwaadwillende uit deze AI-systemen gevoelige data kan terugleiden.

Voorbeeld: herleiden van persoonsgegevens uit AI

Een AI-systeem detecteert fraude op basis van kenmerken van abonnees. Het systeem wordt doorlopend getraind op basis van klantgegevens en informatie over fraudegevallen. Het is mogelijk dat één kenmerk (of een specifieke combinatie van kenmerken) slechts bij één abonnee voorkomt. Iedereen met toegang tot het fraudedetectiesysteem kan nu achterhalen of deze specifieke persoon gemarkeerd is (geweest) als fraudeur door deze gegevens in te voeren. Het systeem moet zo worden gebouwd dat het aantal personen waarop uitkomsten terug te herleiden zijn (de 'celgrootte') altijd boven een bepaalde ondergrens ligt.

Mogelijk onethische keuzes in telecominfrastructuren

Wanneer in een telecomnetwerk keuzes moeten worden gemaakt (bijvoorbeeld over welk verkeer voorrang krijgt, waar bepaalde capaciteit wordt ingezet, et cetera) kan het zijn dat daarbij bepaalde maatschappelijke *waarden* onvoldoende of niet in acht worden genomen. Als bijvoorbeeld wordt vereist dat hulpdiensten altijd een bepaalde minimumcapaciteit zouden moeten kunnen gebruiken op een mobiel netwerk, dan moet deze vereiste niet ondermijnd worden door de toepassing van een bepaald AI-algoritme. In tegenstelling tot de eerdergenoemde effecten betreft dit een effect op maatschappelijk niveau, waarbij de doelstelling niet een specifieke toepassing is (en het daarbij leveren van de gevraagde connectiviteit) maar de maatschappelijk gewenste (ethisch verantwoorde) uitkomst.

In sommige gevallen kan een beslissing daarnaast worden genomen op basis van gegevens die hier niet voor gebruikt zouden mogen worden, of op zijn minst discutabel is. Er zijn diverse voorbeelden bekend, zoals applicaties die het batterijniveau van een smartphone meenemen om de kredietwaardigheid van de gebruiker te bepalen. [36]

Voorbeeld: Beslissingen op basis van ongewenste informatie

Een AI die toegang heeft tot alle data die beschikbaar is bij een telecomoperator maakt integraal beslissingen op basis van die gegevens. Doordat een AI zelf de relaties tussen de in- en output modelleert is het op voorhand niet bekend hoe die informatie zal worden gebruikt bij het maken van een beslissing. Mag een AI bijvoorbeeld gesprekken naar de klantenservice lagere of hogere prioriteit geven op basis van het gegeven of een klant al eerder een klacht heeft gemeld bij de klantenservice? Hetzelfde risico speelt wanneer telecomdata *buiten* de telecomsector in AI-toepassingen wordt gebruikt.

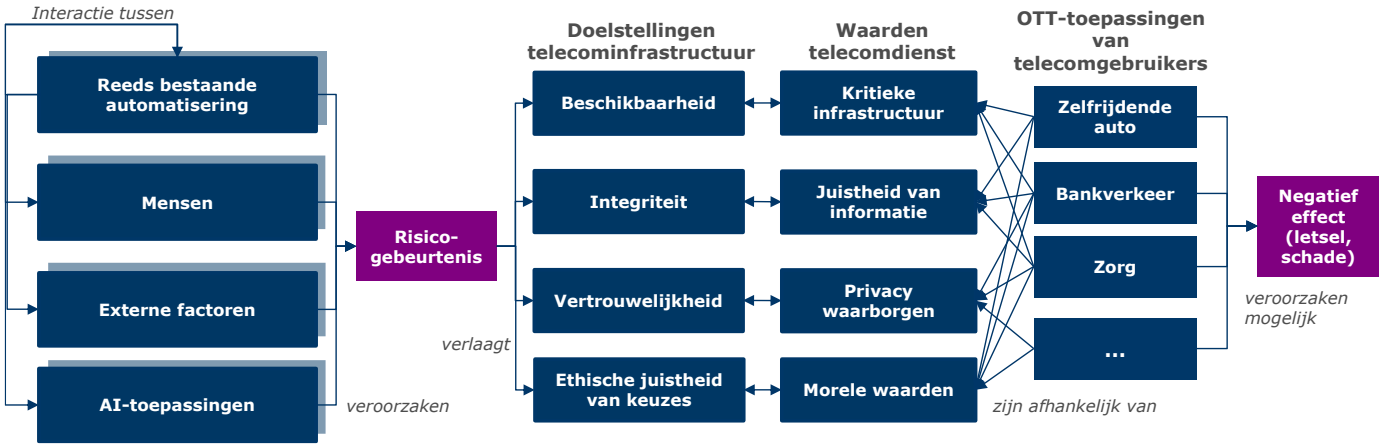
Voorbeeld: Pre-emptive of prioritering?

In een calamiteitenscenario is van belang dat hulpdiensten altijd kunnen communiceren. Op een netwerk kan dit op verschillende manieren worden ingericht. Zo kan er altijd capaciteit zijn gereserveerd voor hulpdiensten (een soort "vluchtstrook"), of kan capaciteit van gebruikers met lagere prioriteit worden afgenomen zodra de hulpdiensten deze nodig hebben. [37] In het verleden is gebleken dat de verschillende technieken in verschillende scenario's hun eigen voor- en nadelen kennen. Een mens of systeem dat een telecomnetwerk ontwerpt en beheert moet deze voor- en nadelen afwegen binnen een normatief kader. Een AI-systeem is mogelijk niet in staat zelf de passende normatieve afwegingen te maken.

3.1.3 Risicopropagatie in de telecomketen

Risico's moeten niet alleen in isolatie worden beschouwd, maar ook in relatie tot elkaar: wat gebeurt er als er twee gebeurtenissen tegelijkertijd optreden? Wat gebeurt er als de ene gebeurtenis het gevolg is van een andere gebeurtenis?

Figuur 8 toont schematisch hoe AI-toepassingen, mensen, externe factoren en reeds bestaande automatisering (al dan niet in interactie met elkaar) risico-gebeurtenissen kunnen veroorzaken. Deze gebeurtenissen hebben negatieve impact op de *doelstellingen* van telecominfrastructuur, waardoor waarden van telecomdiensten in gevaar komen, en er (door niet goed functioneren van de afhankelijke toepassingen) negatieve effecten kunnen ontstaan. In dit onderzoek richten we ons op aantasting van de doelstellingen van telecominfrastructuren.



Figuur 8 De manier waarop (o.a.) AI-toepassingen risico-gebeurtenissen veroorzaken en uiteindelijk leiden tot negatieve maatschappelijke effecten

Inzet van AI-toepassingen in telecominfrastructuren kan een *verschil* veroorzaken in het totale risiconiveau op systeemniveau: elk risico dat al bestond en niet wordt vergroot of verkleind door toepassing van AI valt buiten beschouwing in dit onderzoek. Dit verschil kan (zie Figuur 8) op de volgende manieren ontstaan:

- **Inzet van een AI-toepassing op zichzelf.** We bespreken deze risico's op toepassingsniveau in paragraaf 3.2.

- **Interactie tussen AI-toepassingen en andere systemen.** We bespreken dit eveneens op toepassingsniveau in paragraaf 3.2 als correlatie van waarschijnlijkheid en/of effect.

- **Vervanging van een mens door een AI**. Het door een AI laten uitvoeren van een taak brengt risico's met zich mee, en die kunnen zowel hoger of lager zijn dan wanneer een mens die taak uitvoert. In dit onderzoek worden de risico's van menselijk handelen in telecominfrastructuren niet in kaart gebracht. Het in paragraaf 3.2 gepresenteerde model kan echter wel worden gebruikt om de risico's van de vervangende AI-toepassing te bepalen om zo de afweging bij het inzetten van een mens-vervangende AI-toepassing te kunnen maken.

- **Cyber(on)veiligheid van AI-toepassingen.** Uiteraard zijn ook AI-toepassingen onderhevig aan cyberdreigingen en bijbehorende veiligheidsrisico's. Dit bespreken we in paragraaf 3.1.4.

- **Inzet van AI-toepassingen voor risicomitigatie.** Anders dan bij de voorgaande gaat het hier om AI-toepassingen die specifiek zijn bedoeld om risico's te verkleinen. De positieve effecten moeten daarbij uiteraard opwegen tegen de eventuele nieuwe risico's. We bespreken dit in paragraaf 3.1.5.

3.1.4 Cyber(on)veiligheid van AI-toepassingen

AI-toepassingen zijn informatiesystemen, en derhalve zijn zij onderhevig aan alle cyberdreigingen die daarbij kunnen spelen (aantastingen van confidentialiteit, integriteit en beschikbaarheid van informatie). Berghoff et al. [38] geven een analyse van zwakheden uitgesplitst naar de verschillende fasen in de levenscyclus van een AI-toepassing. Hierin komt een groot aantal risico's naar voren dat in feite reeds bestond voor niet-AI-gebaseerde informatiesystemen in telecominfrastructuur: waar met data gewerkt wordt, is informatiebeveiliging nu eenmaal nodig. In zekere mate gelden de door Berghoff et al. [38] gevonden risico's ook in situaties met reguliere systemen en mensen die beslissingen maken (en daarbij te maken krijgen met gemanipuleerde informatie). Tabel 1 geeft een overzicht van de gevonden zwakheden die (in onze ogen) *nieuw* zijn bij inzet van AI.

Tabel 1 Mogelijke zwakheden van AI-toepassingen ten aanzien van informatiebeveiliging [38]

Fase	Vertrouwelijkheid	Integriteit	Beschikbaarheid
Planning	• Het (deels) gebruiken van bestaande modellen, welke kwaadaardige elementen bevat • Backdoors en bugs in gebruikte software-frameworks voor machine learning		
Dataverzameling	• Grote hoeveelheden data zijn nodig om een model te trainen; deze concentratie van data is mogelijk risicovol.	• "Poisoning attack" waarbij trainingsdata wordt gemanipuleerd om het uiteindelijke resultaat van de AI-toepassing te beïnvloeden (o.a. met	• Er ontstaat een bias in de trainingsdata voor een subset van de cases, waardoor de uiteindelijk AI-

Fase	Vertrouwelijkheid	Integriteit	Beschikbaarheid
		verborgen "triggerpatronen"[9])	toepassing niet goed werkt voor deze subset
Training	• Training wordt vaak op gedeelde (cloud)infrastructuur uitgevoerd, waarbij confidentialiteit moeilijker te garanderen is.	• Training wordt vaak op gedeelde (cloud)infrastructuur uitgevoerd, waarbij (meer) mogelijkheden tot sabotage kunnen bestaan.	
Testen en evalueren		• Door manipuleren van de testset kan (in de feedbacklus naar training) een bias worden geïntroduceerd.	
Operatie	• Het model kan 'verborgen' informatie over broncases bevatten (bijvoorbeeld wanneer er in een bepaalde cel slechts één persoon valt), welke het model kan openbaren. • Backdoors en bugs in onderliggende (cloud)infrastructuur kunnen vertrouwelijkheid in gevaar brengen.	• Een aanvaller kan gewichten in een model manipuleren (en daarmee 'trigger patterns' introduceren of de uitkomsten beïnvloeden) zonder dat dit direct zou opvallen. • Backdoors en bugs in onderliggende (cloud)infrastructuur kunnen ingangen zijn voor sabotage • Adversarial attacks	• Wanneer een probleem gevonden wordt in een AI-systeem is het lastig dit te corrigeren zonder opnieuw te trainen; de doorlooptijd en daarmee termijn van onbeschikbaarheid kan onacceptabel hoog zijn.

3.1.5 Risicomitigatie op basis van AI

AI-toepassingen in de telecomsector kunnen niet alleen risico's veroorzaken, maar ook worden ingezet om risico's te mitigeren. Een aantal manieren waarop dat kan zijn de volgende:

- **Anomaly detection.** Hierbij kan een AI op basis van kleine signalen of de combinatie van signalen een bepaalde afwijking (bijvoorbeeld een falend onderdeel) vroegtijdig(er) detecteren. Het betreffende onderdeel kan dan bijvoorbeeld sneller worden vervangen, waardoor het risico op (latere) uitval afneemt. Een ander voorbeeld is het gebruik van AI-gebaseerde firewalls welke nieuwe vormen van gevaarlijk verkeer kunnen herkennen zonder dat dit verkeer eerder is waargenomen. Het risiconiveau wordt verlaagd doordat de *waarschijnlijkheid* van risico's afneemt.

  Anders dan bij de inzet van AI voor directe aansturingstoepassingen in telecominfrastructuur zijn de risico's hier kleiner. Wanneer het systeem bepaalde zaken terecht opmerkt ("*true positive*") is de meerwaarde hoog, terwijl het niet opmerken van zaken ("*false negative*") het risiconiveau ten opzichte van de situatie zónder AI niet verlaagt. Het onterecht opmerken van zaken die niet schadelijk zijn ("*false positive*") kan overigens wel een probleem vormen, al zullen deze in veel gevallen minder zijn dan de meerwaarde van de "true positives".

[9] Een niet vaak voorkomende combinatie van invoerparameters die in testen/valideren ongetest blijft, en in het AI-model een bepaalde uitkomst forceert.

- **Root cause analysis.** In een storingssituatie of bij een melding van uitval kan AI worden gebruikt om de oorzaak van een storing sneller te achterhalen. Hiermee kan doeltreffender en sneller worden gehandeld. Het risiconiveau wordt verlaagd door dat het negatieve *effect* wordt geminimaliseerd.

- **Simulatie.** Wanneer grote delen van telecominfrastructuur worden bestuurd met AI kan makkelijker worden gesimuleerd hoe het systeem zich gedraagt bij een calamiteit. Zo kan op basis van een kopie van het aansturende model worden getest wat er gebeurt als grote delen van het netwerk uitvallen, er verkeerde informatie wordt ingevoerd, et cetera. Een dergelijke 'brandoefening' kan in principe zelfs continu plaatsvinden. Voor een organisatie waarin systemen samenwerken met mensen is een dergelijke simulatie veel complexer te realiseren.

Bij inzet van AI voor risicomitigatie zal een afweging moeten worden gemaakt of inzet van de AI het netto risiconiveau verhoogt of verlaagt, en of het maximale risiconiveau niet wordt overschreden.

Voorbeeld: monteurs naar de juiste locatie sturen

In een telecomnetwerk zijn veel netwerkfuncties van elkaar afhankelijk; een fout in een systeem kan als onderliggende oorzaak een fout in een heel ander systeem hebben. Het achterhalen van de onderliggende oorzaak is soms lastig, zeker wanneer het (bij uitval) snel moet gebeuren. Een AI kan dit proces versnellen, door op basis van informatie uit het netwerk een inschatting te geven van de locatie van de 'root cause'. De AI zou het uiteraard bij het verkeerde eind kunnen hebben en daarmee het herstelproces juist vertragen. Om in te schatten hoe bruikbaar de AI is kan vooraf echter eenvoudig een groot aantal simulaties worden uitgevoerd, en worden gekeken of de AI daarbij de juiste conclusie trekt. Ook tijdens een calamiteit zou de AI meerdere inschattingen kunnen maken (waarbij steeds bijvoorbeeld gegevens uit een ander deelsysteem worden weggelaten, en wordt gekeken of dan nog steeds dezelfde conclusie geldt).

3.2 Toepassingsniveau

3.2.1 Theoretisch kader

Er zijn verschillende manieren om risico's te kwalificeren of kwantificeren. Een veelgebruikte methode is die van Fine & Kinney [32]. Hierbij wordt risico gemodelleerd als het product van *waarschijnlijkheid*, *blootstelling* en *effect*, en worden deze componenten gescoord volgens de onderstaande Tabel 2.

Tabel 2 Methode voor het scoren van risicocomponenten volgens de methode van Fine & Kinney [32]

Waarschijnlijkheid	Blootstelling	Effect
10 Zeer waarschijnlijk	**10** Voortdurend	**100** Catastrofaal, veel doden, of >10^7 schade[10]
6 Mogelijk	**6** Dagelijks tijdens werkzaamheden	**40** Ramp, meerdere doden, of >10^6 schade
3 Ongewoon, maar mogelijk	**3** Wekelijks of incidenteel	**15** Zeer ernstig, dodelijk slachtoffer, of >10^5 schade
1 Alleen op lange termijn mogelijk	**2** Iedere maand	**7** Substantieel, zware verwonding, of >10^4 schade
0,5 Zeer onwaarschijnlijk	**1** Enkele keren per jaar	**3** Belangrijk, lichamelijk letsel, of >10^3 schade
0,2 Vrijwel onmogelijk	**0,5** Zelden	**1** Aanzienlijk, eerste hulp noodzakelijk of >$100 schade
0,1 Absoluut onmogelijk		

Het product van bovengenoemde componenten geeft een indicatie van de omvang van het risico, en kan volgend onderstaande tabel worden omgezet naar een kwalitatieve indicatie.[11]

Tabel 3 Vertaling van de risicoscore naar kwalitatieve inschatting en maatregelen volgens de methode van Fine & Kinney [32]

Score	Risico	Maatregel
>320	Zeer hoog	Overweeg stopzetten activiteiten
160-320	Hoog	Onmiddellijke actie gewenst
70-160	Aanzienlijk	Correctie noodzakelijk
20-70	Matig	Aandacht vereist
<20	Laag	Acceptabel

Analoog aan de methode van Fine & Kinney is het mogelijk het risiconiveau van AI-toepassingen in de telecomsector te bepalen door het inschatten van de componenten *waarschijnlijkheid* en *effect*.[12]

[10] Deze bedragen zijn afkomstig uit de originele studie uit 1979 en slechts ter illustratie. In werkelijkheid dient gecompenseerd te worden voor inflatie en context (in de originele studie: wapensystemen in de VS).

[11] Zie de interactieve versie op [diasli.de]

[12] In de methode van Fine & Kinney wordt overigens verondersteld dat de waarschijnlijkheid, blootstelling en het effect kenbaar en met zekerheid vast te stellen zijn. We plaatsen hierbij de kanttekening dat er sprake kan zijn van onzekerheid op deze assen, en dat hier de meest pessimistische waarde zou kunnen worden gehanteerd wanneer een inschatting van het maximale risico gewenst is.

3.2.2 Waarschijnlijkheid

Kijken we naar de aspecten van AI-toepassingen in de telecomsector die bepalend zijn voor de *waarschijnlijkheid* dat een negatieve gebeurtenis zich voordoet, dan zien we een aantal categorieën, die sterk gerelateerd zijn aan de manier waarop de AI-toepassing werkt. Hieronder gaan we nader in op deze aspecten en geven we aan hoe deze zich verhouden tot de eigenschappen van AI uit paragraaf 2.2.

Autonomie

AI-toepassingen nemen vaak taken over die normaal door een mens worden uitgevoerd, of ondersteunen de mens. Dat betekent dat deze systemen over een zekere mate van autonomie beschikken. Er kan onderscheid worden gemaakt tussen twee vormen van autonomie: autonoom *leren* en autonoom *handelen*.

Autonoom leren

Een hedendaags AI-model wordt ontwikkeld op basis van grote hoeveelheden (historische) data. Uit deze data 'leert' een algoritme de gewenste uitkomsten bij bepaalde invoer. Er zijn verschillende manieren om dit leren of 'trainen' vorm te geven:

- **Offline learning**. Hierbij wordt eenmalig of om de zoveel tijd een model getraind op basis van een 'statische' dataset. Zowel het model als de gebruikte data kunnen worden getest en gevalideerd, alvorens het model in productie wordt genomen. Ook bij offline leren bestaan (niet-AI-specifieke) risico's rondom informatiebeveiliging, bijvoorbeeld als gevolg van manipulatie van de trainingsgegevens (zie [38] en §3.1.4).

- **Online learning**. Hierbij wordt een model getraind zoals bij offline learning en daarna periodiek opnieuw hertraind op basis van nieuwe data. Ook hierbij is doorlopend testen en validatie mogelijk. Een punt van aandacht is dat de uitkomsten van de AI mogelijk invloed kunnen hebben op de data die worden gebruikt om te trainen, waardoor een soort 'zelfversterkend effect' kan ontstaan.

- **Continuous learning:** Hierbij wordt een model *continu* bijgewerkt aan de hand van binnenkomende gegevens. Denk hierbij bijvoorbeeld aan de loggegevens die worden gegenereerd in telecomapparatuur. Anders dan bij online learning zijn er geen verschillende 'versies' van het model meer te onderscheiden: een prikkel heeft in potentie direct invloed op de volgende beslissing van de AI. We zien twee vormen van risico's:

    - **(Autonome) model drift.** Er bestaat, zonder voldoende toezicht, een risico dat het model na verloop van tijd onjuiste uitkomsten zal genereren of zal tenderen naar een specifieke uitkomst.

    - **Poisoning attack.** Een aanvaller zou de data die het systeem gebruikt om te leren zodanig kunnen manipuleren, dat de uiteindelijke uitkomsten ook wijzigen. Als voorbeeld: een algoritme dat gevaarlijk verkeer zou moeten onderscheppen zou langzaam kunnen 'wennen' aan dergelijk verkeer (doordat een aanvaller het systeem geleidelijk blootstelt aan meer en meer van dit verkeer), en het op enig moment volledig kunnen doorlaten. [38]

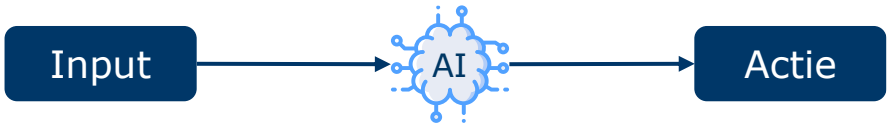
  Door Berghoff et al. [38] wordt overigens nog opgemerkt dat hoewel een aanvaller meer mogelijkheden heeft tot manipulatie indien een systeem continu lerend is, de effecten daarvan echter van tijdelijke aard zullen zijn.

De waarschijnlijkheid dat zich negatieve gebeurtenissen voordoen als gevolg van gebruik van AI-toepassingen is groter wanneer er in de leerfase van een AI-model onvoldoende wordt getest en gevalideerd of de gebruikte gegevens en methode inadequaat zijn. Dit risico is groter wanneer er sprake is van *online learning* en het grootst wanneer er sprake is van *continuous learning.*

Autonoom handelen

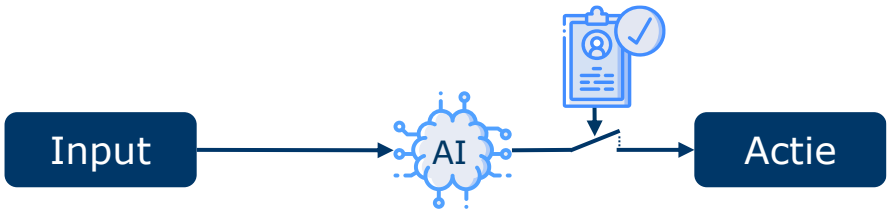
Een AI-toepassing kan op verschillende manieren worden ingezet:

- In een **closed loop**-scenario voert het AI-systeem direct handelingen uit. Een mens heeft hooguit de mogelijkheid het AI-systeem uit te schakelen. Een voorbeeld hiervan is spraakherkenningssoftware.

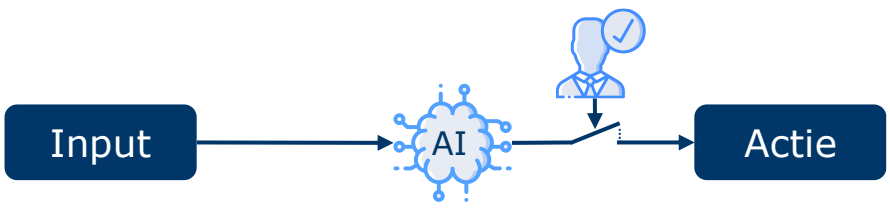


- In een **open loop**-scenario vervult het AI-systeem een ondersteunende rol. De AI presenteert aan een mens een uitkomst op basis waarvan deze persoon kan handelen. In dit scenario is er voor de mens een mogelijkheid om af te wijken van het advies en/of dit advies te controleren op basis van andere informatie. Een voorbeeld hiervan zijn expertsystemen die artsen helpen bij het stellen van een diagnose.

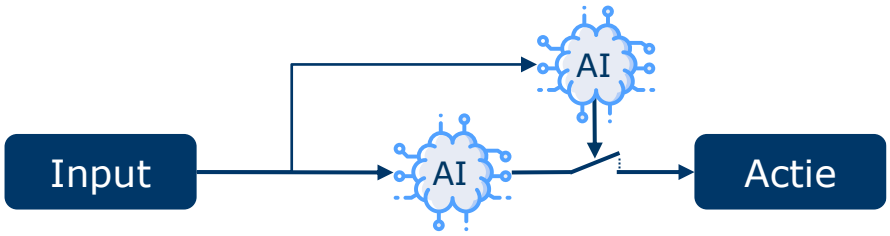
- In een **rule-constrained closed loop**-scenario kan een AI-systeem direct handelingen uitvoeren, maar is deze daarbij beperkt door bepaalde "harde" regels. Overschrijden van de regels leidt direct tot het uitschakelen van het systeem of het niet doorvoeren van de handeling. Een voorbeeld hiervan zijn autonome voertuigen, die vaak worden uitgerust met diverse 'fail safe'-regels die ervoor zorgen dat een auto een noodstop maakt in onveilige situaties.



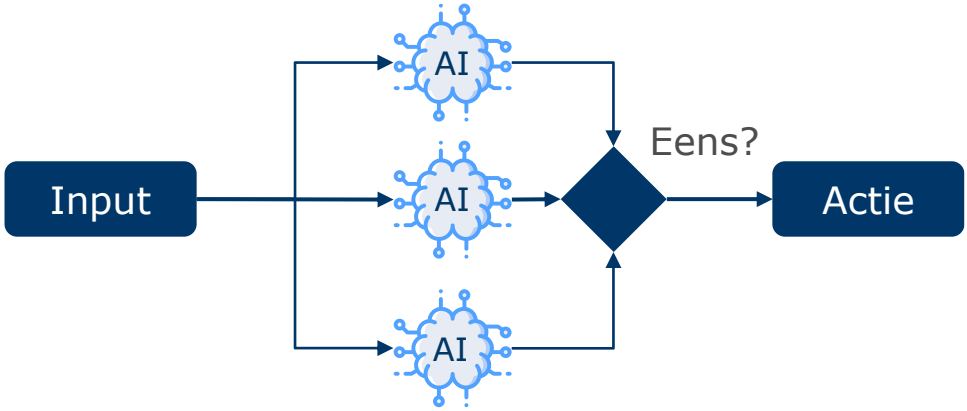
- In een **human-in-the-loop**-scenario kan een AI direct handelingen uitvoeren, maar kan een mens deze handelingen zo nodig tegenhouden of bijsturen. Een voorbeeld zijn autonome voertuigen waarbij een mens de handen aan het stuur moet houden.



- In een **AI-in-the-loop**-scenario wordt één of meerdere aanvullende AI-systemen gehanteerd om een AI-systeem dat handelingen uitvoert te controleren. Het controlerende AI-model krijgt de originele inputs en de beslissing van de AI te zien, en beoordeelt of de beslissing juist is.



Een andere implementatievorm is door meerdere AI-systemen dezelfde beslissing te laten maken, en deze alleen uit te voeren wanneer de beslissingen hetzelfde zijn. Hetzelfde principe wordt toegepast bij navigatiesystemen in vliegtuigen. Drie computers met verschillende implementaties van hetzelfde algoritme rekenen hier navigatieparameters uit, en alleen wanneer de drie uitkomsten exact gelijk zijn worden de uitkomsten gebruikt om het vliegtuig te besturen.



Merk op dat bij scenario's waarin mensen zijn betrokken ("human-in-the-loop" en "open loop") er een risico op gewenning bestaat. Na verloop van tijd kan het vertrouwen van de mens in de AI groeien en/of de aandacht afnemen (*"vigilance decrement"* [39]), waardoor deze afwijkingen minder snel zal opmerken, en er de facto sprake is van een *closed loop* (en daarmee een groter risico).

In volledig autonome scenario's neemt de impact van traditionele informatiebeveiligingsrisico's toe: een aanvaller die de gewichten van een autonoom AI-model kan wijzigen kan bijvoorbeeld langer ongemerkt blijven vanwege de complexiteit van de modellen.

De waarschijnlijkheid dat zich negatieve gebeurtenissen voordoen als gevolg van gebruik van AI-toepassingen is groter wanneer een AI-systeem rechtstreeks kan handelen. Hoewel de risico's te beperken zijn door controle door mensen is het zeer de vraag of een mens altijd de consequenties van een beslissing kan overzien en snel genoeg kan ingrijpen, en of na verloop van tijd niet een te groot vertrouwen ontstaat in de AI-systemen. In sommige situaties kan een AI beter presteren dan een mens, maar zelfs in deze gevallen wordt de waarschijnlijkheid op negatieve gebeurtenissen vergroot wanneer er geen toezicht is.

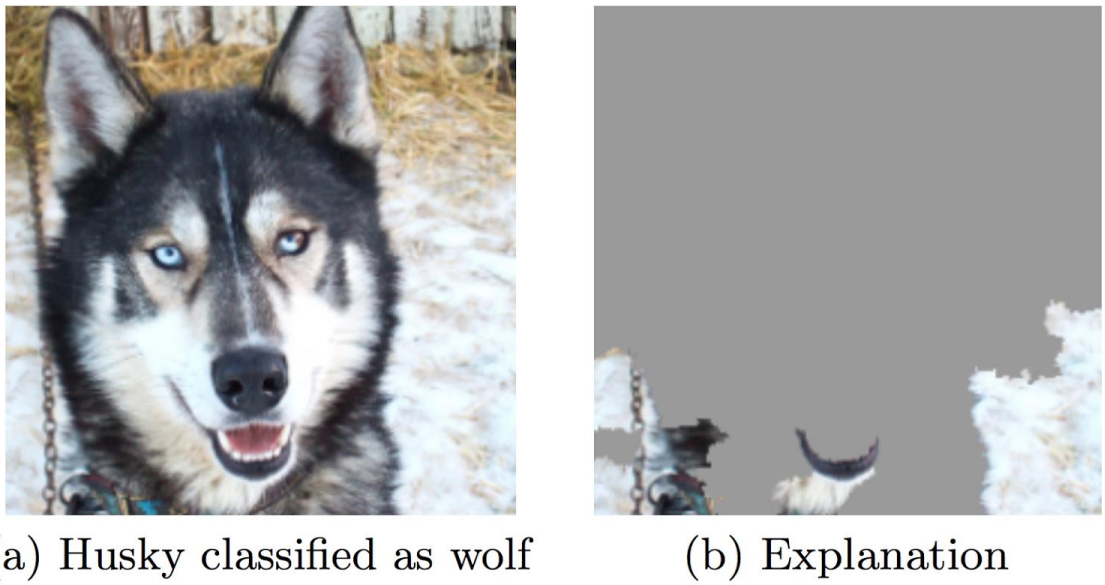
Onvoorspelbaarheid

De mate van voorspelbaarheid heeft grote invloed op de mate waarin waarschijnlijkheid van negatieve effecten kan worden ingeschat. Zoals we eerder aangaven zijn AI-toepassingen

op basis van *deep learning*, vanwege de hoge complexiteit van dergelijke AI-modellen, een stuk onvoorspelbaarder dan regelgebaseerde algoritmes. Een en ander is afhankelijk van de specifieke vorm van deep learning en de implementatie in de toepassing; de volgende elementen zijn daarbij van grote invloed op de voorspelbaarheid.

Transparantie

Sommige vormen van AI, met name die gebaseerd op *deep learning*, zijn opgebouwd uit een groot aantal lagen en coëfficiënten. Het is niet eenvoudig om hieruit af te leiden hoe het model zich gedraagt en op welke basis beslissingen worden gemaakt. Een sprekend voorbeeld is te vinden in Figuur 9. Een model werd getraind om verschillende diersoorten te kunnen onderscheiden. Bij het evalueren van het model bleken de uitkomsten weliswaar erg goed te kloppen, maar bleek dat het model de keuze om een dier te classificeren als wolf vooral te baseren op de aanwezigheid van sneeuw op de foto. Dat leidde ertoe dat nieuwe foto's van andere dieren met veel sneeuw ook de classificatie 'wolf' opleverden.



Figuur 9 Classificatie van dieren door een AI en de informatie die daarbij werd gebruikt [40]

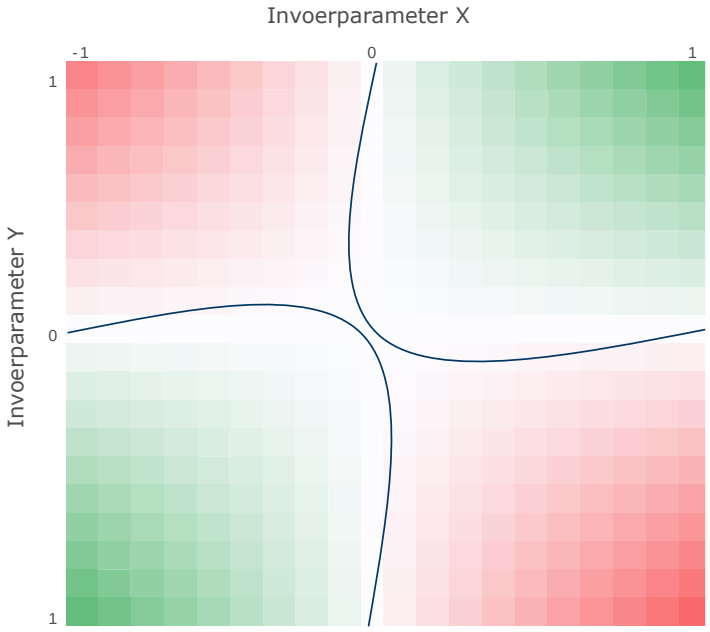
Het gebrek aan transparantie in AI-modellen maakt het voor een kwaadwillende persoon (binnen of buiten de organisatie) eenvoudiger om manipulaties te maken, en maakt dat deze langer onopgemerkt blijven. Zo kan een aanvaller met toegang tot een model gewichten wijzigen zonder dat dit direct zal worden gezien, maar waarmee wel een *trigger pattern* ontstaat in het systeem.

Een niet-transparant systeem verhoogt het risico op *adversarial attacks* [41]. Wanneer een AI niet transparant is kunnen niet robuuste eigenschappen (zoals sneeuw in het husky-wolf voorbeeld) mogelijk worden gebruikt ter classificatie. Wanneer het onbekend is welke eigenschappen exact worden gebruikt, kan een kwaadwillige persoon de niet robuuste eigenschappen manipuleren zonder dat de ontwikkelaar van het systeem hiervan op de hoogte is.

Er zijn ook scenario's denkbaar waarin transparantie van AI juist *niet* gewenst is. In een beveiligingssysteem waarvan de regels precies bekend zijn kan een aanvaller op zoek naar gaten (en wordt de *adversarial attack* eenvoudiger om uit te voeren). In een continu lerend niet-transparant systeem is het systematisch zoeken naar een lek lastiger (maar het systeem op zichzelf natuurlijk niet per definitie veiliger).

Non-lineariteit

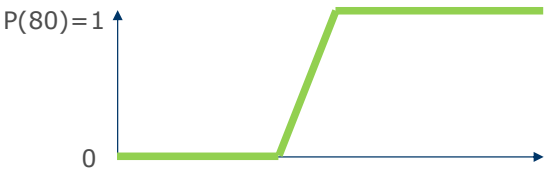
AI-modellen kunnen, afhankelijk van de wijze waarop zij zijn geïmplementeerd, sterk non-lineair gedrag vertonen. Enerzijds zorgt dit ervoor dat deze modellen zeer krachtige representaties kunnen vormen. Anderzijds maakt dergelijk gedrag het lastiger of onmogelijk om de waarschijnlijkheid van negatieve effecten in te schatten. Figuur 10 toont een voorbeeld van een model waarin op basis van twee parameters (hier 'x' en 'y') een uitkomst wordt geschat (hier aangegeven als rood en groen; dit kan bijvoorbeeld een classifier zijn die netwerkverkeer op basis van twee eigenschappen categoriseert als 'goed' of 'slecht'). Zoals uit de afbeelding volgt bestaan er op bepaalde punten scherpere overgangen dan andere. In het middelpunt is de uitkomst in dit voorbeeld het gevoeligst voor wijzigingen in de input: een kleine aanpassing kan ertoe leiden dat de uitkomst de ene of de andere kant uitvalt.



Figuur 10 Voorbeeld van een niet-lineair model met twee parameters, en 'decision boundaries' voor classificatie in twee categorieën (bron: Dialogic).

Een voorbeeld van non-lineair gedrag in de praktijk werd onlangs door onderzoekers ontdekt in een algoritme in zelfrijdende Tesla's dat snelheidsborden herkent. Door op een snelheidsbord met de aanduiding "35" de middelste poot van het getal "3" met enkele centimeters tape te 'verlengen' bleek de auto het getal op het bord ineens te gaan herkennen als "80". [42] Figuur 11 toont dit schematisch. De inschatting van een AI-model ("kans dat het een 80-bord betreft") neemt niet-lineair toe als gevolg van een zeer specifiek kenmerk.

35 → 35



Figuur 11 Niet-lineaire activatiefuncties in AI-modellen leiden tot niet-lineair gedrag van het model (bron: [42], visualisatie Dialogic)

De effecten van non-lineariteit hebben een grotere impact op de waarschijnlijkheid van negatieve effecten wanneer de gegevens die het model gebruikt kunnen worden gemanipuleerd door derden (zoals in het voorbeeld met het snelheidsbord) en/of wanneer de gegevens niet goed worden gevalideerd (en bepaalde invoerwaarden bijvoorbeeld buiten de grenzen kunnen vallen waarmee het model werd getraind).

De voorspelbaarheid van AI-modellen heeft directe invloed op de mate waarin negatieve effecten zich voordoen én de mate waarin dit met zekerheid is vast te stellen. Het risico wordt groter naarmate er onvoldoende transparantie is (de werking van het model is lastig te controleren) en er de mogelijkheid is tot non-lineair gedrag. Het risico is het grootst wanneer de gegevens kunnen worden gemanipuleerd of onvoldoende worden gecontroleerd.

Er zijn methoden om de voorspelbaarheid van AI-systemen te verhogen. Eén methode is het ontwikkelen van een simulatie waarin het AI-systeem kan worden getest alvorens te worden toegepast. Voor een classificatiemodel dat twee parameters (x,y) als input heeft die waarden tussen (-1, 1) kunnen aannemen is het eenvoudig om de complete input-outputruimte te verkennen. Door elke inputparameters op een as te zetten en de mogelijke waarden uit te proberen kan de *decision boundary*, het moment waarop het model besluit om keuze A of keuze B te maken, van een classificatiemodel worden bepaald.

Door de grote parameterruimte waarmee AI-systemen beslissingen maken, is het lastig om de gehele input-ruimte te verkennen. [38] Dit maakt het al snel erg ingewikkeld om in een hoge dimensionaliteit, soms meer dan 1000, inputvariabelen de *decision boundaries* te interpreteren.

Nieuwe situaties

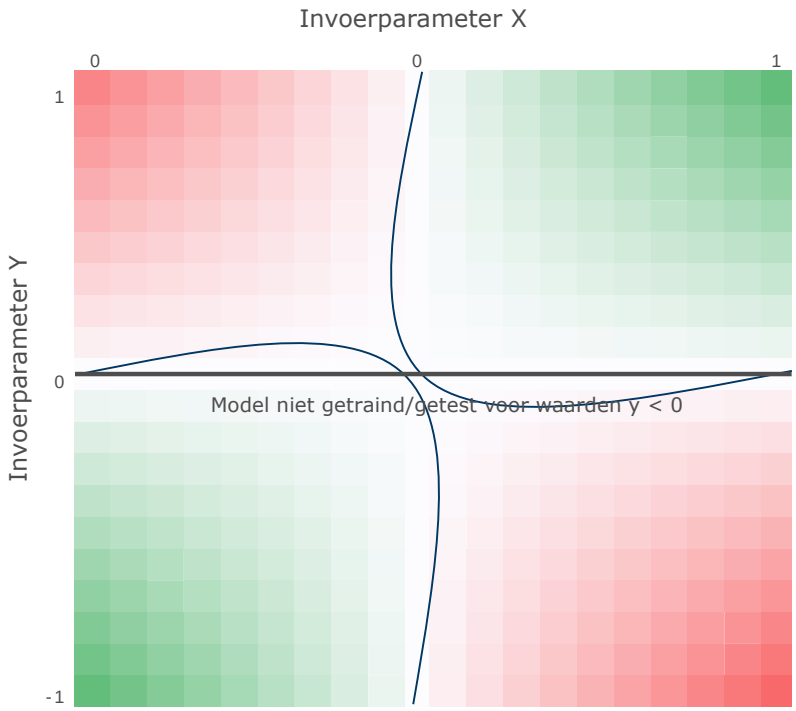
Een AI-model wordt veelal getraind op grote hoeveelheden data, die in het verleden zijn verzameld. Een AI-model leert hierbij welke outputs gepast zijn bij welke combinaties van inputs. Omdat dit leren gebeurt op basis van historische data wordt door het AI-model de aanname gemaakt dat *nieuwe*, nog niet eerder geziene combinaties van inputs te voorspellen zijn op basis van de eerdere combinaties. In sommige situaties kan deze aanname onjuist zijn. Zo is het aangetoond dat op basis van AI-modellen historische aandelenkoersen perfect te voorspellen zijn, maar zijn deze modellen allesbehalve in staat om toekomstige

aandelenkoersen correct te voorspellen. Het adagium *resultaten uit het verleden bieden geen garantie voor de toekomst* geldt dus ook voor AI-toepassingen.

Een AI kan slecht omgaan met nieuwe situaties, omdat 'begrip' van de achterliggende relaties ontbreekt. Een AI kijkt alleen naar input en output, en de achterliggende relaties zijn niet meer dan een 'black box'. De waarschijnlijkheid van risicogebeurtenissen neemt toe in situaties waar sprake kan zijn van nieuwe situaties.

Het is vanwege bovenstaande ook van belang dat de grenzen die worden gesteld aan invoer aan een AI-model, bekend zijn en nageleefd worden. Een model zou bijvoorbeeld getraind en getest kunnen worden binnen een bepaald bereik van een bepaalde invoervariabele. Technisch gezien zal zo'n model echter in staat zijn om ook uitkomsten te genereren búiten dit bereik (geïllustreerd in Figuur 12). Deze uitkomsten hebben echter mogelijk geen relatie met de werkelijkheid, omdat het model hier nooit mee is getraind: het model 'extrapoleert' de werkelijkheid, maar op een puur mathematische manier, en zonder dat vaststaat dat dat valide is.

AI-modellen kennen beperkingen als het gaat om de invoergegevens. Wanneer gegevens buiten het gevalideerde bereik worden ingevoerd zijn de uitkomsten mogelijk eveneens ongeldig. Het is van belang dat deze grenzen bekend zijn en worden afgedwongen in het gebruik van AI. Wordt dit niet gedaan, dan neemt de kans op risicovolle gebeurtenissen (als gevolg van onjuiste uitkomsten) toe.



Figuur 12 Voorbeeld van een model dat getraind is voor een bepaald invoerbereik, maar wel uitvoer (de kleur in de grafiek) zal geven voor waarden daarbuiten

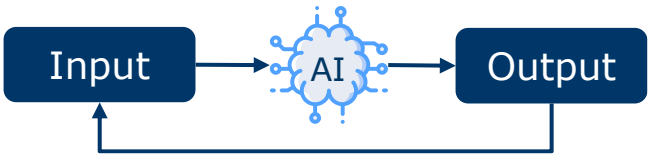
Correlatie

Het is denkbaar dat, voordat een negatief effect zich voordoet, er eerst meerdere gebeurtenissen dienen plaats te vinden. In vliegtuigen zijn bijvoorbeeld veel systemen dubbel

uitgevoerd. Negatieve effecten doen zich in principe pas voor bij het falen van beide systemen. De gebeurtenissen die (samen) leiden tot een negatief effect kunnen echter gecorreleerd zijn. In het vliegtuigvoorbeeld is het denkbaar dat beide systemen bijvoorbeeld hetzelfde mankement bevatten. Of dat ze door een gemeenschappelijke oorzaak beide uitvallen. Als gebeurtenissen samenhangen, spreken we over een correlatie tussen deze gebeurtenissen (nog even los wat de reden daarvan is). In de context van AI-toepassingen in de telecomsector zien we dat de waarschijnlijkheid van negatieve effecten op twee manieren kan worden beïnvloed door correlatie van gebeurtenissen:

- **De output van het ene AI-systeem wordt gebruikt als input door het andere.** Een fout in een eerder systeem kan zo leiden tot een fout in een later systeem, via alle in deze paragraaf beschreven mechanismen (bijvoorbeeld ongeldige invoerdata).

- **De output van een AI-systeem wordt tevens gebruikt als input van ditzelfde systeem, of de systemen zijn anderszins gekoppeld.**[13] Om dezelfde redenen als hierboven kan dit leiden tot escalatie van foute uitkomsten.



In een telecomnetwerk zou een correlatiescenario er als volgt uit kunnen zien. Een basisstation in een mobiel netwerk bepaalt onterecht dat het radiosignaal in een bepaalde richting versterkt dient te worden. Een tweede basissignaal meet een naastgelegen signaal en past daar zijn eigen configuratie op aan, dat leidt tot eenzelfde fout, die wordt opgemerkt door het volgende basisstation. De fout propageert zich vervolgens als een olievlek door het netwerk.

Correlatie van gebeurtenissen kan de waarschijnlijkheid van negatieve effecten vergroten. In telecomnetwerken zijn systemen in hoge mate aan elkaar gekoppeld. Wanneer AI-toepassingen elkaars of de eigen output als input gebruiken, is de waarschijnlijkheid van correlatie het hoogst.

3.2.3 Effect

Kijken we naar de effecten van risicogebeurtenissen voortkomend uit AI-toepassingen, dan zien we twee bepalende factoren: *schade* (de 'zwaarte' van het effect) en *scope* (de reikwijdte van het effect). Daarnaast zien we dat effecten met elkaar gecorreleerd kunnen zijn; in andere woorden, elkaar kunnen versterken indien zij gelijktijdig optreden.

[13] Zie voor een voorbeeld van dit laatste [26], waarin een situatie wordt geschetst waarin meerdere AI "*agents*" autonoom opereren en leren, maar onderling observaties delen.

Schade

De *mogelijke* schade die een AI in een risicogebeurtenis kan veroorzaken hangt samen met het handelingskader van het algoritme (c.q. de mens, die op basis van het advies van AI een verkeerde beslissing zou nemen): hoe belangrijker de beslissingen, hoe groter het risico. Een groter aantal verschillende handelingsopties maakt dat het evalueren van de verschillende opties voor een toezichthouder op het algoritme daarnaast complex zou kunnen zijn.

De potentiële schade is groter wanneer er niet (tijdig) kan worden ingegrepen door een mens of wanneer de beslissingsruimte van het algoritme niet op een andere manier wordt beperkt. In dat kader gelden alle afwegingen die eerder werden gemaakt rondom autonoom handelen (p. 35).

Hoe invloedrijker de (indirecte) beslissingen van een AI-toepassing zijn, hoe groter de negatieve effecten bij een risicogebeurtenis. Wanneer een AI-toepassing autonoom handelt, zijn de negatieve effecten in sommige gevallen groter.

Scope

Naast het handelingskader van een AI-toepassing is de *scope* van handelingen van een AI van belang. In een telecomnetwerk kan de scope onder andere worden uitgedrukt in het aantal (potentieel) getroffen gebruikers of geografisch gebied. In een telecomnetwerk kunnen verschillende 'lagen' worden onderscheiden (zoals het toegangs-, transmissie- en coreniveau) waarbij de scope steeds groter wordt. Ten aanzien van de scope van een AI-toepassing in een telecomnetwerk zien we de volgende gradaties:

- **Volledig geïsoleerd.** Het algoritme maakt keuzes die effect hebben in een strak afgebakende omgeving. De uitkomsten van de AI heeft geen enkele invloed op andere systemen binnen de telecominfrastructuur. Een voorbeeld is een algoritme dat *beamforming* optimaliseert in een zendmast, of bijvoorbeeld ruisreductie verbetert in voor een bundel VDSL-lijnen. Een verkeerde uitkomst heeft alleen impact op de betreffende aansluitingen. De toepassingen zijn over het algemeen sterk *decentraal.* De scope is beperkt tot (1) een enkele of kleine groep gebruikers, (2) een geografisch sterk afgebakend gebied en/of (3) alleen het toegangsdeel van het netwerk.

- **Deels geïsoleerd.** Het handelingskader van een algoritme is goed afgebakend, maar er zijn manieren waarop een verkeerde beslissing impact kunnen hebben op een ander systeem. Hiervan is bijvoorbeeld sprake wanneer de output van een algoritme meetbaar invloed heeft op een ander systeem. Een andere denkbare route is dat het niet goed functioneren van een AI-systeem leidt tot het overschrijden van beveiligingsgrenzen (zoals bijvoorbeeld een aardlekschakelaar of zekering) die ervoor zorgt dat ook andere systemen uitvallen. De scope is beperkt, maar tot een grotere groep gebruikers, groter geografisch gebied, en/of meer dan alleen het toegangsdeel van het netwerk.

- **Niet geïsoleerd**. Dit betreft systemen die bedoeld zijn om andere systemen aan te sturen. Een fout in het aansturende systeem heeft directe gevolgen voor de werking van de aangestuurde systemen. De scope is in potentie het gehele netwerk, alle gebruikers, en het volledige geografische dekkingsgebied.

Bijzondere aandacht verdient *edge computing*, een techniek waarbij (al dan niet toepassingspecifieke, en mogelijk op AI gebaseerde) intelligentie wordt aangebracht in de randen van het netwerk. Hoewel de invloedssfeer van deze toepassingen lokaal is, bestaat er een risico op beïnvloeding van andere applicaties die dezelfde infrastructuur gebruiken.

Bij AI-toepassingen die zich decentraal op lokaal niveau bevinden (en risico's niet met elkaar gecorreleerd zijn) spelen over het algemeen kleinere negatieve effecten bij risicogebeurtenissen dan bij centrale AI-toepassingen die ontworpen zijn om andere systemen aan te sturen.

Correlaties

Eerder werd besproken dat correlatie tussen de waarschijnlijkheden van risicogebeurtenissen kan leiden tot nieuwe (of hoger dan verwachte waarschijnlijkheid voor bestaande) risico's. Ook aan de effectzijde kan correlatie risico's vergroten. Een analogie is het beveiligen van een gebouw tegen inbraak: wanneer het alarm niet wordt aangezet is er niet direct een vergroot risico op schade door inbraak; de deur is immers op slot gedraaid. Wanneer alleen de deur niet op slot wordt gedraaid, is er ook niet direct een vergroot risico (het alarm werkt immers nog). Echter, wanneer zowel het alarm niet is ingeschakeld en ook de deur niet op slot is, is het risico veel groter dan de som van beide risico's: een inbreker kan nu zonder problemen naar binnen, en er is sprake van schade.

Domino-effect

Wanneer systemen aan elkaar gekoppeld zijn, en beslissingen in het ene systeem het andere beïnvloeden, bestaat een risico op een "domino-effect": een fout in een systeem veroorzaakt een fout in het volgende systeem. Een dergelijk mechanisme lag aan de basis van de "flash crash" van Wall Street in 2010. Daarbij detecteerde één algoritme foutieve invoer als een afwijking, en besloot het om effecten te verkopen. Andere algoritmes zagen deze handeling als afwijkend en handelden op dezelfde manier, met als effect dat de koers uiteindelijk instortte en handel moest worden stilgelegd. [43] Het snelle handelen van de algoritmes en het feit dat niet bekend was dat de algoritmes "anomaly detection"-mechanismen bevatte, maakte dat de koersdaling zeer snel optrad. Uiteraard is niet gezegd dat menselijke beurshandelaren niet vatbaar zouden zijn voor hetzelfde: ze kunnen in paniek raken, met uiteindelijk hetzelfde effect.

Redundantie en variatie

Redundantie is een manier om risico's te mitigeren. Door elementen te dupliceren kan uitval van het ene element worden opgevangen door het andere. Daarnaast kan de output van twee elementen worden vergeleken en kan worden opgemerkt wanneer er een verschil ontstaat (dit werkt het beste wanneer de elementen volledig verschillende implementaties van dezelfde functie zijn). In telecomnetwerken is vaak redundantie aangebracht; netwerken worden bijvoorbeeld in ringen aangelegd, zodat het verbreken van een verbinding niet leidt tot volledig verlies van connectiviteit tussen de aangesloten locaties.

Een aan redundantie gelieerd concept is *variatie*. Het introduceren van redundantie om risico's te mitigeren is alleen effectief wanneer uitval van de redundante elementen onderling niet gecorreleerd is. Twee AI-systemen die redundant zijn ten opzichte van elkaar zullen - als zij verder exact gelijk zijn – exact dezelfde fout maken gegeven dezelfde invoer; in dat geval is er in de praktijk dus helemaal geen verlaagd risico door redundantie. Dit kan worden opgelost door variatie te introduceren. Hierbij zijn de twee redundante AI-toepassingen twee verschillende implementaties (compleet gescheiden implementaties, of bijvoorbeeld een oudere versie).[14] De kans dat beide systemen tegelijk tegen een fout aanlopen, is daarmee kleiner.

[14] [space.stackexchange.com] geeft een interessante kijk op hoe dit is geregeld in de Space Shuttle.

De negatieve effecten bij een risicogebeurtenis zijn kleiner bij een AI-toepassing die acteert in een redundant onderdeel van een telecominfrastructuur, zolang er geen correlatie is tussen het voorkomen van de risicogebeurtenissen op meerdere redundante opstellingen.

3.3 Risicobepaling

Zoals in het begin van dit hoofdstuk is toegelicht kunnen de daadwerkelijke risico's van toepassing van AI in telecominfrastructuren uitsluitend worden bepaald door dit op systeemniveau te analyseren. Niet alleen AI-toepassingen in isolatie, maar ook de interactie tussen AI-toepassingen en andere systemen speelt een rol bij het optreden van risicogebeurtenissen. De uiteindelijke negatieve effecten van zo'n gebeurtenis hangen sterk samen met het uiteindelijk gebruik van de telecominfrastructuur. Het valt buiten de scope van dit onderzoek, en is daarnaast in onze ogen niet mogelijk, om een volledig dekkend risicomodel te maken, zonder te kijken naar *specifieke* toepassingen en situaties.

Dat gezegd hebbende heeft het wel degelijk zin om te kijken naar de eigenschappen van AI-toepassingen en het directe effect dat deze hebben op het optreden van risicogebeurtenissen. We noemen dit het *toepassingsniveau*. Figuur 13 toont het model voor inschatting van de additionele risico's van *individuele toepassingen* van AI in telecominfrastructuren.

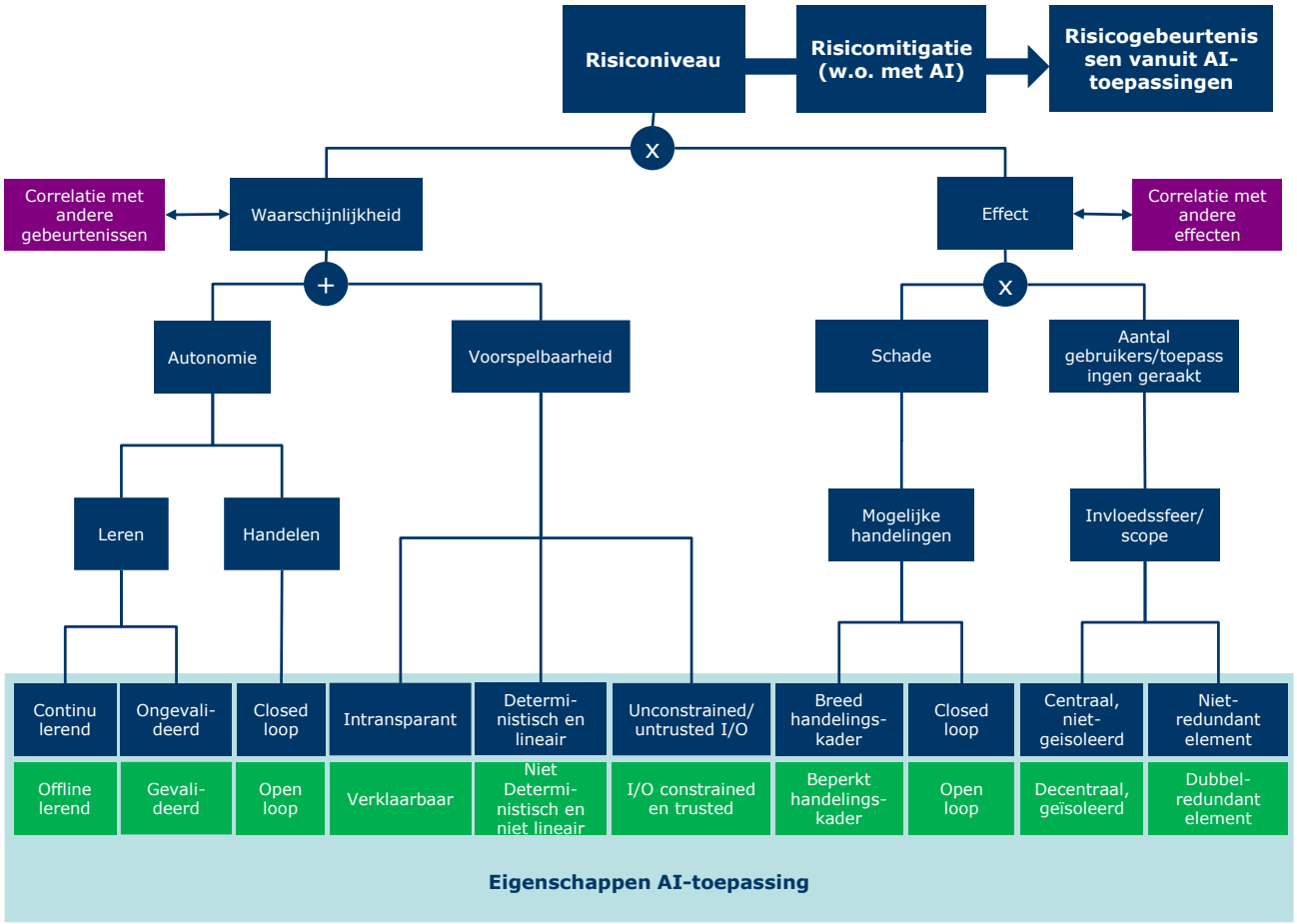
Zoals besproken is het risiconiveau sterk afhankelijk van de eigenschappen van de AI-toepassing, die (indirect) bepalend zijn voor zowel de waarschijnlijkheid als het effect van risicogebeurtenissen. Op de betreffende eigenschappen van AI-toepassingen zou beleid kunnen worden gevoerd.

Risicomitigatie door de operator kan het risiconiveau van een toepassing terugbrengen tot een (voor de operator en maatschappij) acceptabel niveau. Hier ligt voor een toezichthouder een belangrijke beleidsvraag: wat is het voor de maatschappij acceptabele risiconiveau, en welke mitigatiemaatregelen dienen operators te nemen?

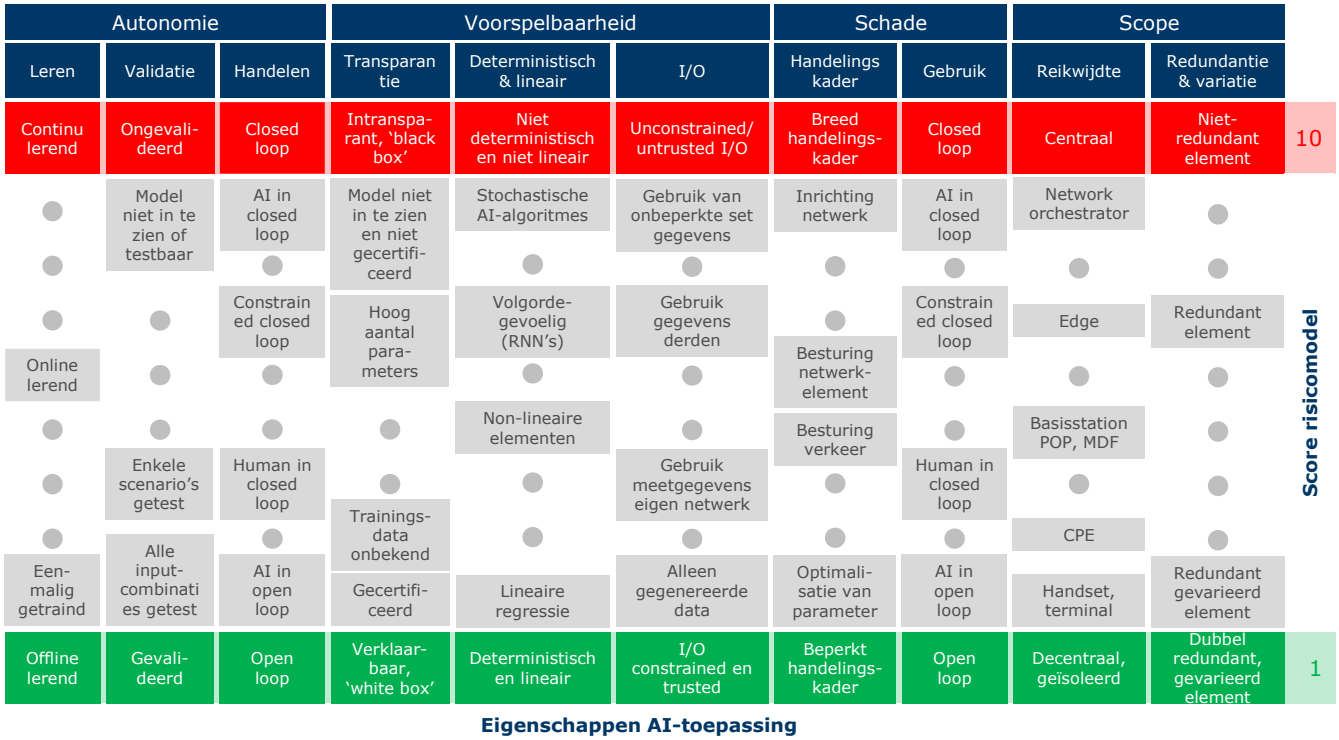
Scoren van risico-aspecten op toepassingsniveau

Hoewel de exacte weging van aspecten in het risicomodel onderwerp van discussie kan zijn, kunnen we op basis van ons onderzoek (met name literatuur en gesprekken met experts) een eerste aanzet geven voor een scoremodel. Figuur 14 toont dit model.

In Figuur 14 wordt aan iedere voor risico's relevante eigenschap van een AI-toepassing een score toegekend tussen 1 (laagste) en 10 (hoogste). De criteria die leiden tot een bepaald rapportcijfer zijn weergegeven in de grijze blokken, en zijn in de bovenstaande paragrafen uitvoerig toegelicht.



Figuur 13 Modellering van additionele risico's van toepassing van AI in telecominfrastructuren



Figuur 14 Beoordeling van AI-toepassingen in telecominfrastructuren aan de hand van het risicomodel

4 Rol van Agentschap Telecom

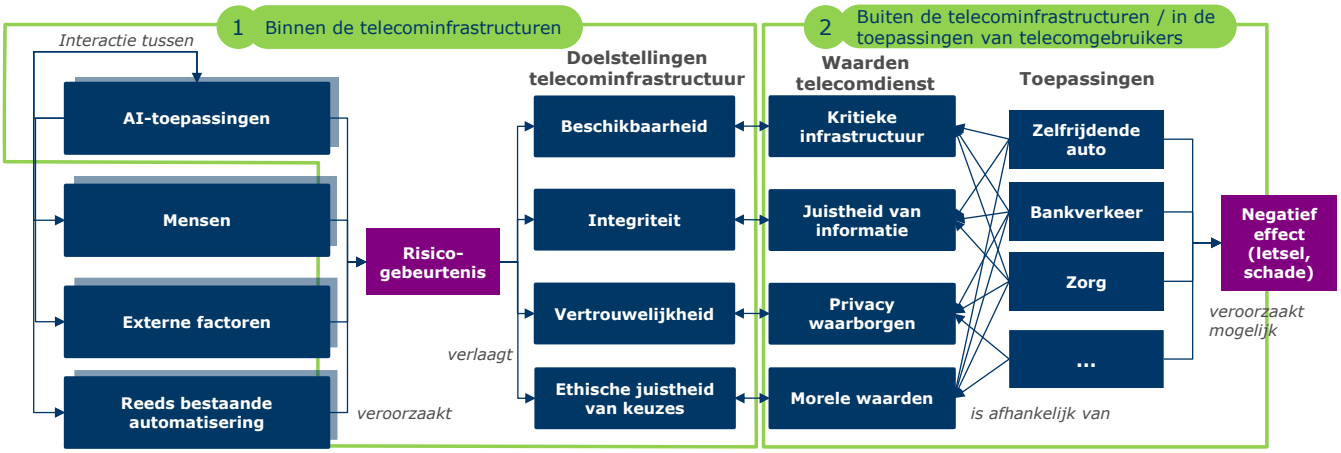
In dit hoofdstuk bespreken we de rol die Agentschap Telecom, als overheidsorganisatie/ toezichthouder, kan vervullen bij het omgaan met (additionele) risico's die ontstaan door het gebruik van AI in telecominfrastructuren.

4.1 Handelingskader

In paragraaf 3.1.3 bespraken we hoe de inzet van AI-toepassingen op systeemniveau kan leiden tot negatieve maatschappelijke effecten (waaronder letsel, schade en verlies van vertrouwen in het stelsel). De toezichthouder kan op twee verschillende punten in deze keten ingrijpen. (Figuur 15):

1. **Binnen de telecominfrastructuren.** De toezichthouder kan maatregelen treffen met als doel dat in telecominfrastructuren geen onverantwoorde AI-systemen worden ingezet die negatieve eigenschappen bezitten die (kunnen) leiden tot additionele risico's.

2. **Buiten de telecominfrastructuren/in de OTT-toepassingen van telecomgebruikers.** De toezichthouder kan de relatie tussen toepassingen en eisen aan telecominfrastructuren proberen uit te werken, om "onverantwoord" gebruik van telecominfrastructuur te beperken. Wanneer aan de toepassingszijde meer kennis en bewustzijn is over de risico's van onderliggende telecominfrastructuren kunnen risico's op toepassingsniveau beter worden ingeschat.



Figuur 15 Propagatie van AI-eigenschappen tot maatschappelijke negatieve effecten

Daarnaast zouden er eisen gesteld kunnen worden aan toepassingen en randapparatuur zelf. In dat laatste kader biedt de bestaande Europese regelgeving mogelijkheden. Zo moet randapparatuur voor telecommunicatie voldoen aan 'essentiële vereisten' zoals vastgelegd in de betreffende Europese richtlijnen. Richtlijn 2014/53/EU (bekend als de 'Radio Equipment Directive (RED)' bepaald bijvoorbeeld dat "'*[Tot bepaalde categorieën of klassen behorende radioapparatuur moet zo geconstrueerd zijn dat zij voldoet aan de volgende essentiële eisen:] radioapparatuur schaadt het netwerk of de werking ervan niet en maakt evenmin misbruik van de netwerkmiddelen waardoor een onaanvaardbare achteruitgang van de dienst wordt veroorzaakt*". [44] Een toezichthouder zou er zich hard voor kunnen maken dat deze aanvullende verplichting wordt 'geactiveerd' (waarbij de Europese Commissie bevoegd

is om gedelegeerde handelingen vast te stellen die een dergelijke 'activering' mogelijk maken). [45]

In dit onderzoek richten we ons met name op een telecomtoezichthouder, en daarmee primair op het eerste punt in de keten. Dat wil niet zeggen dat de toezichthouder niet óók in het toepassingsdomein zou kunnen acteren. In zekere zin begeeft het Agentschap Telecom zich al in dit domein in haar programma Telekwetsbaarheid, waarin bewustwording wordt gecreëerd rondom uitval onder gebruikers van telecominfrastructuren. [46]

4.2 Weging van risico's

De wijze waarop negatieve effecten worden gewogen hangt samen met het ambitieniveau ten aanzien van telecominfrastructuur, en is daarmee maatschappelijke afweging. Wanneer het uitgangspunt is dat de maatschappij (bepaalde) telecominfrastructuur moet kunnen gebruiken voor missiekritische toepassingen (met maatschappelijke meerwaarde), dan liggen de eisen ten aanzien van risico's hoger dan wanneer de ambitie lager is (bijvoorbeeld alleen bedrijfskritisch gebruik). Risico-effecten zijn dan ook te definiëren ten opzichte van de gewenste dienstniveaus van telecomnetwerken (dan wel het niveau van alle infrastructuren tezamen – voor kritische communicatie kan om risico's te spreiden een combinatie worden ingezet van meerdere verschillende infrastructuren)[15]

Een toezichthouder die risico's van AI in de telecomsector wil inschatten, zal allereerst een ambitieniveau moeten bepalen ten aanzien van de infrastructuren: wanneer maatschappelijk gewenst is dat missiekritische diensten kunnen werken op basis van de beschikbare telecominfrastructuren, dan wegen risico's zwaarder dan wanneer slechts 'best effort'-dienstverlening gewenst is.

Het eerder ontwikkelde risicomodel kan vervolgens de basis vormen voor de toezichthouder om op toepassingsniveau te bepalen welke AI-toepassingen risico's met zich meebrengen. Verder is beschreven hoe AI juist zou kunnen bijdragen aan het beperken van risico's. Een toezichthouder kan vervolgens op specifieke toepassingen reguleren. Een andere methode is om juist te kijken naar de eigenschappen van AI-toepassingen, te bepalen welke combinaties van eigenschappen leiden tot een (in de ogen van de toezichthouder) te hoog (additioneel) risico, en op basis van deze combinaties regulering samen te stellen.

4.2.1 Toegepast op geïdentificeerde toepassingen

Het gepresenteerde risicomodel kan worden toegepast op de in het onderzoek geïdentificeerde AI-toepassingen, om een beeld te geven van de toepassingen die aanvullende risico's zouden kunnen veroorzaken. Tabel 4 toont hiervan een overzicht.

Op het eerste gezicht wekt de inschatting in Tabel 4 wellicht, onterecht, de suggestie dat bepaalde toepassingen beter kunnen worden geweerd uit telecominfrastructuren. We benadrukken echter dat in deze tabel uitsluitend wordt gekeken naar het *toepassingsniveau*. Zoals aangegeven kunnen er aanvullende risico's zijn als gevolg van inbedding van de toepassing binnen de telecominfrastructuur (door correlatie van risico's). Een AI-toepassing kan daarnaast ook worden ingezet om risico's juist te mitigeren. Tot slot is geen weging

[15] Niet voor niets wordt in de ISO31000:2009-standaard het begrip 'risico' breder gedefinieerd als (vrij vertaald) "de gevolgen van de onzekerheid op het halen van bepaalde doelen", waarmee in feite wordt onderkend dat het niet alleen gaat om risico's die direct letsel of schade opleveren, maar dat ook het niet halen van doelen tot dergelijke effecten kan leiden verderop in een keten.

toegekend aan de scores noch een koppeling gemaakt met de negatieve gevolgen: het model is niet normatief ten aanzien van het acceptabele risiconiveau noch de scores in Tabel 4.

Wel identificeert het model waar AI-toepassingen met welke eigenschappen kunnen leiden tot nieuwe risico's. Uit de tabel zou bijvoorbeeld kunnen worden afgeleid dat meer aandacht moet worden besteed aan validatie: zo kunnen risico's bij overigens beperkt risicovolle AI-toepassingen verder worden teruggedrongen.

Tabel 4 Gescoorde risico-aspecten van huidige en toekomstige toepassingen van AI in telecominfrastructuren[16]

Toepassing	Kansen: AI-mitigatie [17]	Autonomie			Voorspelbaarheid			Schade		Scope	
		Leren	Validatie	Handelen	Transparantie	Deterministisch & lineair	I/O	Handelingskader	Gebruik	Reikwijdte	Redundantie en variatie
Power management		1-6	1-10	7	5	5	4	2-5	7	5	5-7
Radio-optimalisatie		1-6	1-10	7	5	5	4	2	7	5	5-7
Optical network signal amplification		1	1	7	5	5	4	2	7	5	5-7
Path computation		6-10	1-10	7	5-10	5-9	4	10	7	10	5-7
Self-organizing networks		6-10	1-10	10	5-10	5-9	4-9	10	10	10	5-7
Performance monitoring	✓	1-10	1-10	1-4	5-10	2-5	4-9	2	1-4	10	5-7
Predictive maintenance	✓	6-10	10	1-4	5	2-7	4-9	2	1-4	10	5-7
Smart handovers		1-6	1-10	7	5	2-5	4	2-5	7	7	5-7
SDN/NFV		6-10	1-10	10	5-10	2-9	4-9	10	10	10	5-7
Optical network nonlinearity mitigation	✓	1-6	1	7	5	2-5	4	5	7	4	5-7
Bandwidth slicing / Resource allocation		6-10	1-10	7	5-10	2-9	4-7	5	7	5	5-7
Virtual topologies		6-10	1-10	10	5-10	2-9	4-9	10	10	7	5-7
Anomaly detection / malicious traffic detection	✓	10	10	10	10	10	4-10	2-5	10	5-10	5-7
Dynamische spectrumtoewijzing		6-10	1-10	7	5-10	2-9	4-9	7-9	7	7	5-7

[16] De cellen zijn als volgt gekleurd: groen: 1 t/m 4, oranje: 5 t/m 7, roze: 8 t/m 9, rood: 10. Bij een bereik is steeds de laatste van toepassing zijnde kleur uit deze reeks gehanteerd.

[17] De kolom "Kansen: AI-mitigatie" geeft aan of AI in deze toepassing specifiek wordt ingezet om andere, niet-AI-risico's, te mitigeren.

4.3 Instrumenten

Agentschap Telecom zou verschillende instrumenten kunnen inzetten om de additionele risico's van toepassing van AI in telecominfrastructuren te verlagen:

- **Voorlichting en bewustwording.** Door het voeren van campagnes binnen en buiten de telecomsector worden operators en gebruikers bewust van de additionele risico's die AI-gebaseerde systemen met zich meebrengen. Naast het feit dat beide groepen zich kunnen beraden op het mitigeren van de risico's kan ook een dialoog op gang worden gebracht om de dienstenniveaus en doelstellingen in lijn te brengen met de behoeften van eindgebruikers.

- **Het vereisen van transparantie**. Van operators kan worden verlangd dat zij inzicht geven in het gebruik van AI in het netwerk en de wijze waarop risico's zijn gemitigeerd. De toezichthouder zou hiervoor een model of format kunnen voorstellen. Dit "risicolabel" maakt het voor eindgebruikers inzichtelijk of het netwerk geschikt is voor de betreffende toepassing.

> ***Certificering als transparantie-garantie?***
>
> In veel vakgebieden buiten AI worden eisen gesteld aan producten. In deze gevallen onderschrijft de fabrikant via een verklaring (een *"supplier declaration of conformity"*) dat het product aan de eisen voldoet. Onderzoekers van IBM beargumenteren dat een soortgelijke verklaring of certificering zou kunnen worden ingezet om de betrouwbaarheid van AI te verhogen. [15] In de voorgestelde methode worden de aspecten *fairness* (eerlijkheid en evenwichtigheid van het algoritme), *explainability* (uitlegbaarheid van de resultaten), *robustness* (robuustheid, waaronder niet-lineaire effecten) en *lineage* (afkomst van trainingsdata) beoordeeld. De producent vult hiervoor een uitgebreide vragenlijst in over het product.

- **Het faciliteren van risicoanalyse en -mitigatie.** Een toezichthouder zou het delen van kennis rondom risico's van AI-toepassingen in telecominfrastructuren kunnen faciliteren. Een van de manieren waarop dit kan worden vormgegeven is het organiseren van dialoog tussen de verschillende operators. Ook zou de toezichthouder met individuele partijen kunnen spreken om te kijken welke ontwikkelingen er spelen en hoe risico's worden aangepakt.

- **Het ontwikkelen van criteria.** Een toezichthouder zou criteria kunnen ontwikkelen of aanwijzen ten aanzien van het gebruik van AI in telecominfrastructuren. Het zou kunnen gaan om generieke criteria (hoe wordt omgegaan met trainingsdata, controle op autonome systemen, et cetera) of specifieke criteria behorend bij specifieke toepassingen. In internationaal verband zijn diverse initiatieven opgestart om tot criteria te komen (zie o.a. [47]).

- **Het stellen van procesvereisten aan operators**. Een toezichthouder zou van operators kunnen verlangen dat zij processen inrichten om additionele risico's van AI te mitigeren, bijvoorbeeld door bepaalde controles of een mate van validatie/transparantie te verplichten.

- **Het voorstellen en het steunen van voorstellen voor het 'activeren' van additionele verplichtingen voor telecom-randapparatuur door de Europese**

Commissie. Zoals hierboven reeds aangegeven, kan de Europese Commissie een gedelegeerde handeling inzetten om additionele vereisten te stellen die helpen om schadelijke effecten van het gebruik van AI in randapparatuur te voorkomen.

Bovengenoemde instrumenten kunnen worden ingezet op het systeemniveau. Het gaat dan (bijvoorbeeld) over voorlichting of het stellen van transparantie-eisen op functioneel niveau, en gerelateerd aan de waarden van telecominfrastructuur.

De genoemde instrumenten kunnen ook op toepassingsniveau specifieke risico-eigenschappen van AI-toepassingen adresseren. Een overzicht van in onze ogen logische inzet is te vinden in Tabel 5. De 'lichte' instrumenten zijn waarschijnlijk het meest effectief wanneer zij gericht zijn op specifieke risico-eigenschappen. Andere 'zwaardere' instrumenten, zoals het stellen van standaarden of procesvereisten hebben een bredere scope.

Tabel 5 Instrumenten en risico-verhogende eigenschappen van AI-toepassingen in telecominfrastructuur

	Autonomie			**Voorspelbaarheid**			**Schade**		**Scope**	
	Leren	**Validatie**	**Handelen**	**Transparantie**	**Deterministisch & lineair**	**I/O**	**Handelingskader**	**Gebruik**	**Reikwijdte**	**Redundantie en variatie**
Voorlichting en bewustwording	✓	✓	✓	✓	✓	✓				
Het vereisen van transparantie		✓		✓	✓	✓				
Het faciliteren van risicoanalyse en -mitigatie	✓	✓	✓				✓	✓	✓	✓
Het stellen van standaarden		✓		✓	✓	✓		✓	✓	✓
Het stellen van procesvereisten	✓	✓	✓				✓	✓	✓	✓

5 Conclusie

In dit hoofdstuk beantwoorden we de onderzoeksvragen die in dit onderzoek centraal stonden (zie §1.2).

5.1 Beantwoording hoofdvraag

Wat zijn risico's van huidige en toekomstige inzet van AI in de telecomsector, en hoe kan Agentschap Telecom deze risico's beperken?

AI-toepassingen hebben specifieke eigenschappen welke risico's kunnen opleveren bij gebruik in telecominfrastructuren. De mate van autonoom leren en handelen, de mate van onvoorspelbaarheid, het handelingskader en de invloedssfeer van de AI-toepassing zijn bepalend voor de waarschijnlijkheid en de impact van de additionele risico's.

Boven op de *additionele* risico's van AI-toepassingen in telecominfrastructuren bestaan (nog steeds) gangbare risico's ten aanzien van informatiebeveiliging in de volledige levenscyclus van een AI-toepassing (planning, dataverzameling, training, testen en validatie en operatie). Verder zien we dat AI nadrukkelijk waarde kan toevoegen bij het mitigeren van risico's.

AI-toepassingen interacteren met andere AI-toepassingen, met mensen, met 'gewone' automatisering, en mogelijk met de buitenwereld. Het is daarom van belang om de toepassing van AI in de telecomsector te beoordelen op systeemniveau. Daarbij moet worden gekeken naar de uiteindelijke toepassingen die worden gerealiseerd op basis van de telecominfrastructuren, en het dienstenniveau dat zij van de infrastructuur vragen.

Aan Agentschap Telecom staan verschillende instrumenten ter beschikking om risico's van toepassing van AI in telecominfrastructuren te beperken: voorlichting en bewustwording, het vereisen van transparantie, het faciliteren van risicoanalyse en -mitigatie, het ontwikkelen van criteria en het stellen van procesvereisten. Daarnaast zouden op Europees niveau additionele verplichtingen ten aanzien van randapparatuur kunnen worden geactiveerd.

We bevelen aan om te starten met instrumenten op systeemniveau. Op specifieke risicoeigenschappen van AI-toepassingen kunnen eventueel specifieke instrumenten worden ingezet. In bredere zin zal een maatschappelijke discussie moeten plaatsvinden over het gewenste dienstenniveau van telecominfrastructuren.

In dit onderzoek is gekeken naar AI in de telecomsector, waarbij een specifieke definitie van AI is gehanteerd. Het is zeer denkbaar dat de conclusies ook (deels) toepasbaar zijn op autonome, zelflerende en datagedreven toepassingen in bredere zin. Ook in andere toepassingsdomeinen van het Agentschap Telecom spelen mogelijk vergelijkbare ontwikkelingen en risico's op het gebied van AI.

5.2 Beantwoording deelvragen

Hoe ziet het huidige gebruik van AI eruit in de telecomsector en sectoren die gebruik maken van digitale connectiviteit?

In de context van telecominfrastructuur betekent AI het gebruiken van algoritmes op basis van deep learning, getraind met behulp van grote hoeveelheden data, om taken te automatiseren die voorheen alleen (goed) door een mens zouden kunnen worden uitgevoerd.

Op dit moment zien we dat de meeste toepassingen van AI in telecominfrastructuren betrekking hebben op optimalisatie van specifieke parameters. Het betreft sterk afgebakende toepassingen. Het is daarbij niet altijd duidelijk of wat de fabrikant "AI" noemt ook daadwerkelijk betekent dat er algoritmes op basis van deep learning en getraind op grote hoeveelheden data worden gebruikt. Algoritmische optimalisatie wordt immers al jarenlang toegepast in telecominfrastructuren.

Welke ontwikkelingen worden de komende 5 jaar voorzien voor het gebruik van AI bij het verzorgen en gebruiken van digitale connectiviteit?

Kijken we naar de komende vijf jaar, dan zien we deze toepassingen steeds geavanceerder worden. Een eindvisie, die wordt gedeeld door een aantal leveranciers van telecomapparatuur, is dat telecomnetwerken in hun geheel kunnen worden bestuurd door een AI. Hoewel het de vraag of dit al (volledig) binnen 5 jaar gebeurt, is deze eindvisie er wel degelijk een waar rekening mee gehouden moet worden.

Welke risico's ten aanzien van beschikbaarheid, authenticiteit, integriteit, vertrouwelijkheid, transparantie en voorspelbaarheid ontstaan er in de diverse sectoren als gevolg van het huidige en toekomstige gebruik van AI? Hoe kunnen de risico's voor de verschillende aspecten relatief ten opzichte van elkaar worden gewogen in een risicomodel voor digitale connectiviteit?

AI-toepassingen kunnen bepaalde eigenschappen bezitten die leiden tot aanvullende risico's voor telecominfrastructuren. Op basis van een risicomodel kunnen deze risico's worden ingeschat. Een schematische weergave van dit model is te vinden in Figuur 13. Deze eigenschappen hebben te maken met de volgende aspecten van AI:

- **De mate van autonoom leren en handelen van de AI-toepassing.** Wanneer hier in hoge mate sprake van is, neemt de waarschijnlijkheid van risicogebeurtenissen toe. Een belangrijke parameter is of de toepassing wordt gecontroleerd door mensen of regels.

- **De mate van voorspelbaarheid van de AI-toepassing.** Wanneer de modellen niet-deterministisch of sterk niet-lineair zijn, is het lastiger om te valideren of een AI-toepassing onder alle omstandigheden goed werkt. Een factor die daarbij meespeelt is welke data wordt gebruikt en of die manipuleerbaar is.

- **Het handelingskader van de AI-toepassing.** Wanneer de AI-toepassing een sterk beperkte invloed heeft op telecominfrastructuren is het effect van een risicogebeurtenis beperkter. Een toepassing met een breed handelingskader leidt in potentie tot grotere effecten.

- **De invloedssfeer van de AI-toepassing**. Een toepassing die op centraal niveau werkt en een telecominfrastructuur bestuurt is risicovoller dan een toepassing die op laag niveau een specifieke parameter optimaliseert.

Uitsluitend kijken naar de risico's van AI-toepassingen in isolatie geeft echter een onjuist, te beperkt beeld van de maatschappelijke risico's (en overigens voordelen) van de inzet van AI-toepassingen in telecominfrastructuren. Op systeemniveau beïnvloeden de volgende factoren de risico's:

- **Interactie tussen AI-toepassingen en andere systemen.** We bespreken dit eveneens op toepassingsniveau in paragraaf 3.2 als correlatie van waarschijnlijkheid en/of effect.

- **Vervanging van een mens door een AI**. Het door een mens laten uitvoeren van een taak brengt risico's met zich mee, en die kunnen hoger of lager zijn dan bij een AI-toepassing. In dit onderzoek worden de risico's van menselijk handelen in telecominfrastructuren niet in kaart gebracht. Het in paragraaf 3.2 gepresenteerde model kan echter wel worden gebruikt om de risico's van de vervangende AI-toepassing te bepalen om zo de afweging bij het inzetten van een mens-vervangende AI-toepassing te kunnen maken.

- **Inzet van AI-toepassingen voor risicomitigatie.** Op systeemniveau kunnen AI-toepassingen bijdragen aan het verlagen van het risiconiveau, bijvoorbeeld door het sneller detecteren van problemen of aanvallen, en het ondersteunen bij het vinden van oorzaken en oplossingen.

- **Cyber(on)veiligheid van AI-toepassingen.** Uiteraard zijn ook AI-toepassingen onderhevig aan cyberdreigingen en bijbehorende veiligheidsrisico's. Deze risico's worden mogelijk vergroot, omdat voor het trainen van AI-toepassingen grote hoeveelheden (soms gevoelige) data bijeen worden gebracht.

Tot slot is door ons geen weging toegekend aan de scores noch een koppeling gemaakt met de negatieve gevolgen: het model is niet normatief ten aanzien van het acceptabele risiconiveau. Wel identificeert het model wáár de toepassing van welk type AI-toepassing, en met welke eigenschappen daarvan, kan leiden tot nieuwe risico's.

Hoe kan Agentschap Telecom als toezichthouder en uitvoeringsorganisatie deze risico's beperken?

Aan Agentschap Telecom staan verschillende instrumenten ter beschikking om risico's van toepassing van AI in telecominfrastructuren te beperken: voorlichting en bewustwording, het vereisen van transparantie, het faciliteren van risicoanalyse en -mitigatie, het stellen van standaarden en het stellen van procesvereisten. We bevelen aan om te starten met instrumenten op systeemniveau. Op specifieke risico-eigenschappen van AI-toepassingen kunnen eventueel specifieke instrumenten worden ingezet. In bredere zin zal een maatschappelijke discussie moeten plaatsvinden over het gewenste dienstenniveau van telecominfrastructuren.

6 Referenties

[1] Horáková, J. (2006). *From Golem to cyborg: a note on the cultural evolution of the concept of robots* [www.ceeol.com] Slovenská Akadémia Vied - Kabinet výskumu sociálnej a biologickej komunikácie. pp. 83-98.

[2] MIT (2007). *Contributions and Impact* [jmc.stanford.edu]

[3] Buchanan, B.G. (2005). *A (Very) Brief History of Artificial Intelligence*

[4] Wright, J., en Vesonder, G. (1990). *Expert systems in telecommunications*

[5] Ng, A., Lee, H., Grosse, R., en Ranganath, R. (2011). *Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks*

[6] Krizhevsky, A., Sutskever, I., en Hinton, G. (2012). *ImageNet Classification with Deep Convolutional Neural Networks* vol. 25,

[7] LeCun, Y., Kavukcuoglu, K., en Farabet, C. (2010). *Convolutional networks and applications in vision*

[8] Tromp, J., en Farnebäck, G. (2015). *Combinatorics of Go* [tromp.github.io]

[9] Devlin, J., Chang, M., Lee, K., en Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [1810.04805]*

[10] Berner, C., Brockman, G., Chan, B., en Cheung, V. (2019). *Dota 2 with Large Scale Deep Reinforcement Learning* [arxiv.org] OpenAI.

[11] Barrat, J. (2015). *Our Final Invention: Artificial Intelligence and the End of the Human Era*St. Martin'S Griffin.

[12] Sutton, R. (2019). *The Bitter Lesson*

[13] Chollet, F. (2019). *On the Measure of Intelligence [1911.01547]*

[14] Ardila, D., Kiraly, A., Bharadwaj, S., en al., e. (2019). *End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography.*Nature.

[15] Arnold, M., Bellamy, R., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Natesan Ramamurthy, K., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., en Varshney, K. (2019). *FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity* [arxiv.org] vol. 63, pp. 6:1-6:13.

[16] Seth, Y. (2019). *BERT Explained – A list of Frequently Asked Questions* [yashuseth.blog]

[17] Zając, Z. (2016). *Bayesian machine learning* [fastml.com]

[18] Hochreiter, S., en Schmidhuber, J. (1997). *Long Short-Term Memory* [doi.org] vol. 9, MIT. pp. 1735-1780.

[19] Haeri, S., en Trajkovic, L. (2017). *Virtual Network Embedding via Monte Carlo Tree Search* pp. 1-12.

[20] de Vries, M., en Albers, J. (2019). *AI in het onderwijs: wat mag er?*Utrecht: Dialogic. pp. 31-46.

[21] Goodfellow, I., Shlens, J., en Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples [1412.6572]*

[22] Qiu, S., Zhou, S., Liu, Q., en Wu, C. (2019). *Review of Artificial Intelligence Adversarial Attack and Defense Technologies* [www.researchgate.net] vol. 9, p. 909.

[23] Chen, L., Ye, Y., en Bourlai, T. (2017). *Adversarial Machine Learning in Malware Detection: Arms Race between Evasion Attack and Defense*Athene, pp. 99-106.

[24] Finlayson, S.G., Chung, H.W., Kohane, I.S., en Beam, A.L. (2019). *Adversarial Attacks Against Medical Deep Learning Systems* [arxiv.org]

[25] Light Reading (2019). *Guavus Takes Jio's Big Data Challenge* [www.lightreading.com]

[26] Csáji, B.C. (2001). *Approximation with Artificial Neural Networks*Hungary: Eötvös Loránd University.

[27]Naderializadeh, N., Sydir, J., Simsek, M., Nikopour, H., en Talwar, S. (2019). *When Multiple Agents Learn to Schedule: A Distributed Radio Resource Management Framework* [arxiv.org]

[28]DARPA (2020). *Spectrum Collaboration Challenge* [www.spectrumcollaborationchallenge.com]

[29]Dialogic, Radicand Economics & iMinds (2016). *The impact of network virtualisation on the Dutch telecommunications ecosystem: An exploratory study* [www.dialogic.nl] Utrecht: Dialogic.

[30]ITU-T. Focus group on Machine Learning for Future Networks including 5G(FG-ML5G) (2019). *FG-ML5G-ARC5G. Unified architecture for machine learning in 5G and future networks* [www.itu.int] Geneve: ITU-T.

[31]Vriezekolk, E. (2016). *Assessing Telecommunications Service Availability Risks for Crisis Organisations* [research.utwente.nl] Universiteit Twente.

[32]Kinney, G., en Wiruth, A. (1976). *Practical Risk Analysis For Safety Management (No. NWC-TP-5865)*China Lake, CA: Naval Weapons Center.

[33]NCTV. *Overzicht vitale processen* [www.nctv.nl]

[34]RTL Nieuws (2020). *Agent aangevallen in Tilburg, noodknop werkt niet: 'Dit moet echt opgelost worden'* [www.rtlnieuws.nl]

[35]Sue, J., Brand, P., Brendel, J., Hasholzner, R., Falk, J., en Teich, J. (2018). *A predictive dynamic power management for LTE-Advanced mobile devices*Barcelona, pp. 1-6.

[36]King, L. H. (2019). *This startup uses battery life to determine credit scores* [money.cnn.com]

[37]Kassa, B. (2016). *Quality of Service. Priority and Preemption* [www.npstc.org]

[38]Berghoff, C., Neu, M., en von Twickel, A. (2020). *Vulnerabilities of Connectionist AI Applications: Evaluation and Defence* [arxiv.org]

[39]Parasuraman, R. (1986). *Vigilance, Monitoring and Search*New York: Wiley. pp. 41-1 - 41-49.

[40]Ribeiro, M.T., Singh, S., en Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* [arxiv.org]

[41]G. Fidel, R.B. A. S. (2019). *When Explainability Meets Adversarial Learning: Detecting Adversarial Examples using SHAP Signatures*

[42]The Next Web (2020). *Some Teslas have been tricked into speeding by tape stuck on road signs* [thenextweb.com]

[43]Business Insider (2016). *Here's what actually caused the 2010 "Flash Crash"* [www.businessinsider.com]

[44]Europese Unie (2014). *Richtlijn 2014/53/EU van het Europees parlement en de Raad van 16 april 2014 betreffende de harmonisatie van de wetgevingen van de lidstaten inzake het op de markt aanbieden van radioapparatuur en tot intrekking van Richtlijn 1999/5/EG* [eur-lex.europa.eu]

[45]DARE!! Measurements. *Radio Equipment Directive (RED)* [www.dare.nl]

[46]Agentschap Telecom (2020). *Telekwetsbaarheid* [www.agentschaptelecom.nl]

[47]ETSI. *Industry specification group (ISG) securing artificial intelligence (SAI)* [www.etsi.org]

[48]Korf, R.E. (1997). *Does Deep-Blue use AI?*

[49]Fidel, G., Bitton, R., en Shabtai, A. (2019). *When Explainability Meets Adversarial Learning: Detecting Adversarial Examples using SHAP Signatures*

Bijlage 1. Overzicht interviewrespondenten

Naam	Rol	Organisatie
Anne van Otterlo	Account CTO	Nokia
Gerwin Franken	Senior Regulatory Officer	Nokia
Patrick Blankers	Regulatory Officer	Ericsson
Jeroen Buijs	CTO	Ericsson
Jurjen Veldhuizen	Solutions Director	Huawei
Corine van Pinksteren	Regulatory officer	KPN
Sacha van der Wijer	Head of Advanced Analytics	KPN
Chris Molanu	Lead AI	KPN
Winifred Andriessen	Director Advanced Analytics	KPN
Simone Van Ginhoven	Regulatory officer	VodafoneZiggo
Aziz Mohammadi	Director advanced analytics	VodafoneZiggo
Michiel van Rijthoven	Lead data scientist	VodafoneZiggo
Frank van Berkel	Senior regulatory affairs counsel	T-Mobile
Miruna Anastasoaie	Lead AI	T-Mobile
Steven Latré	Professor Computational Science & Artifical Intelligence	Universiteit Antwerpen

Contact:
Dialogic innovatie & interactie
Hooghiemstraplein 33-36
3514 AX Utrecht
Tel. +31 (0)30 215 05 80
www.dialogic.nl