The Hague **Security Delta**

Notitie Risico-Analyse in Onzekerheid

Artificial Intelligence (Kansen en Bedreigingen)

HSD
The Hague **Security Delta**

info@thehaguesecuritydelta.com | thehaguesecuritydelta.com | @HSD_NL

Inhoudsopgave

1. Risicoanalyses om de dynamiek voor te blijven

De wereld verandert snel. Dat er sprake is van een 'technologische tsunami' wordt breed erkend. Desondanks worden we niet zelden overvallen door rap opkomende technologieën die, naast nieuwe kansen, ook nieuwe bedreigingen inhouden die soms veel verder gaan dan wat we op grond van onze ervaringen uit het verleden konden verwachten. Dit moet en kan beter. Niemand kan de precieze aard en timing van de grote veranderingen en de daaruit voortvloeiende nieuwe risico's voorspellen, maar goede analyses stellen ons wel in staat om de range aan mogelijkheden te verkennen en ons zo actief voor te bereiden op wat komen gaat.

Deze notitie is bedoeld om voor een concrete casus – de snelle vooruitgang in *Artificial Intelligence* (AI) - te schetsen hoe een scenariogedreven risicoanalyse ons in staat stelt om (1) de kansen die nieuwe technologie biedt in kaart te brengen om (2) de nieuwe of verschuivende risico's en dreigingen die hierbij opdoemen te karakteriseren, teneinde (3) de oplossingsrichtingen te verkennen om deze risico's en dreigingen tegen te gaan of te beheersen. Het is daarbij van belang te beseffen dat deze nieuwe risico's en dreigingen opkomen in een dynamische systeemcontext. Nieuwe patronen van verknoping en fragmentatie, van samenwerking en conflict, van winnaars en verliezers ontstaan: op mondiale schaal (geopolitieke orde), diep in onze samenleving (het maatschappelijk weefsel) en op alle tussenniveaus (in de rol van Europa versus de lidstaten bijvoorbeeld). Er is sprake van complexe(re) verbindingen omdat interne en externe veiligheid in elkaar overvloeien. In het veiligheidsecosysteem worden andere en meer verschillende spelers van belang. Er ontstaan daarbij nieuwe risico's en dreigingen die onmogelijk goed vallen te snappen en aan te grijpen vanuit de bestaande structuren, omdat deze de ordening uit het verleden weerspiegelen.

Deze notitie kan worden gezien als een vertrekpunt voor een proces waarin de snelle technologische ontwikkelingen op het gebied van AI (en andere gebieden[1]) met een veiligheidsbril op worden bezien. Zo dient de notitie als input voor een HSD-Café in november 2017, waarin een hands-on scenario-oefening rond het thema wordt georganiseerd. Op deze wijze wil HSD Office de partners in HSD en alle overige belanghebbenden informeren over de mogelijke veiligheidsimplicaties van de snelle ontwikkelingen op, bijvoorbeeld, het gebied van AI, om vervolgens in scenariogedreven risicoanalyses gezamenlijk te bepalen welke innovatieve veiligheidsoplossingen nodig zijn. Deze notitie en het bijbehorend proces bieden tevens houvast voor de agendering en de toekomstige ontwikkeling, rol en positionering van HSD.

[1] Tegelijk met deze notitie over AI verschijnt een soortgelijke notitie over Distributed Ledger Technology (block chain).

2. Eerste Kennismaking

Als mogelijke *game changer*

> "The twenty-first century will be dominated by algorithms. 'Algorithm' is arguably the single most important concept in our world. If we want to understand our life and our future, we should make every effort to understand what an algorithm is, and how algorithms are connected with emotions. An algorithm is a methodical set of steps that can be used to make calculations, resolve problems and reach decisions. An algorithm isn't a particular calculation, but the method followed when making the calculation." [2]

Kunstmatige Intelligentie (artificial intelligence, AI) wordt al een aantal decennia gezien als een van de belangrijkste potentiële *game changers* van onze tijd. Ondanks succesjes op deelgebieden zijn de hooggespannen verwachtingen nog niet waargemaakt. Dat gaat veranderen. De afgelopen paar jaar is er grote vooruitgang geboekt. Slimme, zelflerende algoritmen leren beslissingen nemen of patronen herkennen door een grote hoeveelheid bestaande beslissingen of patronen te analyseren. We beginnen langzamerhand de impact van AI op ons dagelijkse en professionele leven echt te merken. Denk aan verbeterde automatische vertalingen, de opkomst van op spraakherkenning gebaseerde digitale assistenten, slimmere zoekalgoritmes en geautomatiseerd juridisch of medisch advies. In de woorden van Tesla-baas en visionair Elon Musk: "to some degree we [humans] are already a cyborg — you think of all the digital tools that you have — your phone, your computer".

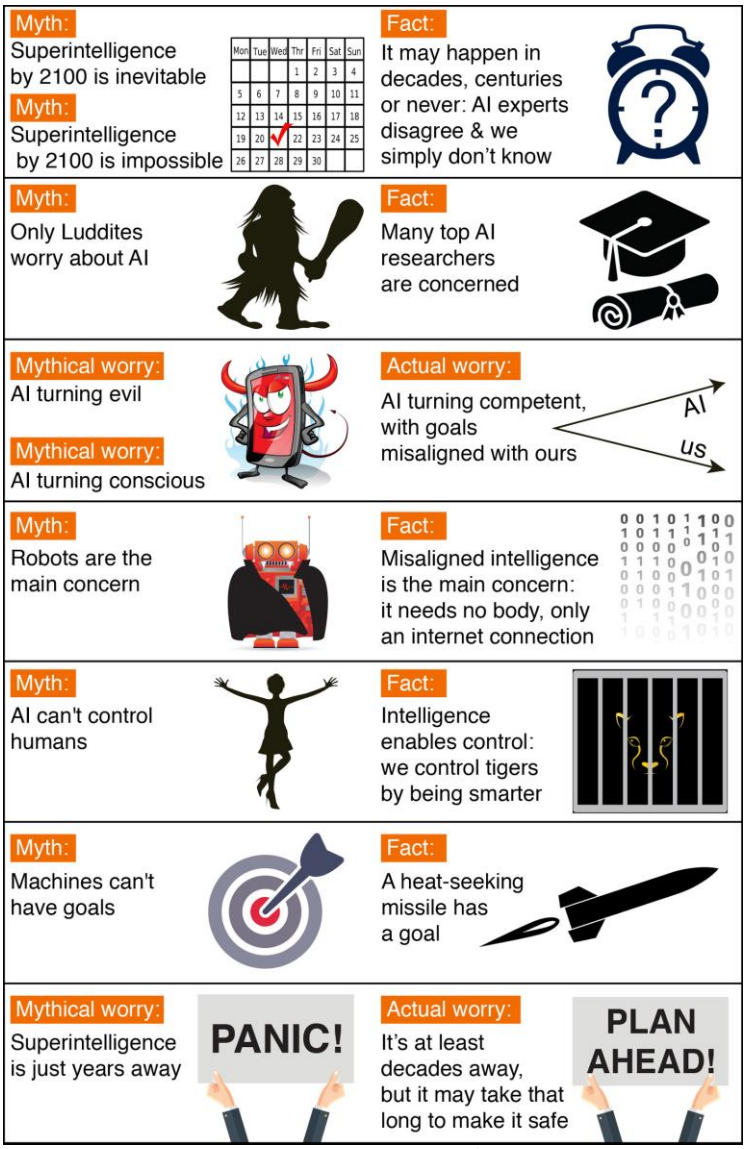
AI is op bepaalde deelgebieden inmiddels gelijkwaardig aan menselijke intelligentie of begint deze zelfs te overtreffen. Die deelgebieden worden groter en veelvuldiger en beginnen op elkaar aan te sluiten. Er wordt gewaarschuwd voor de gevolgen van steeds krachtiger AI-systemen, en niet door de minsten. In 2015 schreven onder meer Elon Musk en Stephen Hawking een open brief waarin stond dat AI onder menselijke controle moest worden gehouden, daarbij implicerend dat AI ook 'out of control' zou kunnen raken.[3] Andere zijn verontrust over het verdwijnen van banen onder invloed van AI, en de sociale en politieke gevolgen daarvan.[4] Tegelijk ziet iedereen de voordelen van AI-toepassingen. Dezelfde open brief beschrijft (vooral) deze voordelen, alsmede de snelheid waarmee deze zich kunnen gaan aandienen: "there is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable."

Vooral in een specifiek deelgebied van AI, *machine learning,* lijken de ontwikkelingen in een stroomversnelling geraakt. Experts verwachten de komende jaren een verdere versnelling. Veel van het onderzoek naar AI vindt plaats in Amerikaanse bedrijven als IBM, Google en Apple. De (opgetelde) investeringen van de Amerikaanse private sector is vele malen groter dan het budget dat de Amerikaanse federale overheid beschikbaar stelt.[5]

AI heeft ook bij de huidige stand van zaken al duidelijke implicaties voor veiligheid. AI kan nu al arbeidsintensieve activiteiten zoals analyse van satellietbeelden of tijdkritische toepassingen in het kader van cyber defence grotendeels automatiseren. Maar AI heeft de potentie om dergelijke

[2] Yuval Noah Harari, Homo deus: A brief history of tomorrow, 2016.
[3] Zie https://futureoflife.org/ai-open-letter
[4] Zie onder meer Edward Luce, The Retreat of Western Liberalism, 2017.
[5] McKinsey Global Institute, Artificial intelligence. The next digital frontier?, 2017.

deelgebieden te ontstijgen en het hele veiligheidsdomein te transformeren; een strategische impact vergelijkbaar met die van vliegtuigen, kernwapens, computers en communicatietechnologie. Onder invloed van AI zullen processen en structuren, actoren en samenwerkingsvormen, werkwijzen en prioriteiten in het veiligheidsdomein drastisch veranderen, zowel aan de dreigings- als aan de bestrijdingskant.



Figuur 1: De grootste mythes over AI[6]

6 "Benefits & Risks of Artificial Intelligence", http://futureoflife.org/background/benefits-risks-of-artificial-intelligence/

Hoe uniek is de mens?

> "Consciousness is the biologically useless by-product of certain brain processes. Jet engines roar loudly, but the noise doesn't propel the aeroplane forward. Humans don't need carbon dioxide, but each and every breath fills the air with more of the stuff. Similarly, consciousness may be a kind of mental pollution produced by the firing of complex neural networks. It doesn't do anything. It is just there. If this is true, it implies that all the pain and pleasure experienced by billions of creatures for millions of years is just mental pollution. This is certainly a thought worth thinking, even if it isn't true. But it is quite amazing to realise that as of 2016, this is the best theory of consciousness that contemporary science has to offer us." [7]

Zelfbewustzijn wordt gezien als een van de belangrijkste kenmerken van de mens, een emergente eigenschap die uit de complexiteit van het menselijk brein ontstaat. Waar planten leven en dieren hun leven beleven, daar is de mens in staat ook zijn eigen beleven te beleven. Mensen zijn zich bewust van zichzelf, van anderen en van hun omgeving, alsmede van de interactie tussen deze elementen, en kunnen daarop reflecteren en van leren. Ondanks dat AI meer en sneller informatie kan verwerken dan een mens, betekent het niet dat daarmee ook bewustzijn ontstaat. Maar intelligentie en bewustzijn worden in AI ontkoppeld. Mensen dreigen zo hun waarde te verliezen. Alleen bewuste (menselijke) wezens konden tot voor kort taken uitvoeren die veel intelligentie nodig hebben, zoals het spelen van schaken, autorijden, ziekten diagnosticeren of terroristen identificeren. Maar we ontwikkelen nu nieuwe vormen van niet-bewuste AI die dergelijke taken, allemaal gebaseerd op patroonherkenning, veel beter dan mensen kunnen verrichten. Dit brengt een nieuwe vraag op: welke van de twee is echt belangrijk, intelligentie of bewustzijn? Ooit een leuk tijdverdrijf voor filosofen, in de eenentwintigste eeuw is dit een belangrijke praktische vraagstelling geworden. En het is ontnuchterend om te beseffen dat voor veel menselijke activiteiten het antwoord is: intelligentie is noodzakelijk en bewustzijn facultatief. Het huidige wetenschappelijke antwoord op deze kwestie kan samengevat worden in drie eenvoudige principes:

1. Organismen zijn algoritmen. Elk dier - inclusief Homo sapiens - is een verzameling van organische algoritmen die door natuurlijke selectie gevormd worden door miljoenen jaren van evolutie.
2. Algoritmische berekeningen worden niet beïnvloed door de materiaal waaruit de rekenmachine bestaat. Of een abacus van hout, ijzer of plastic is, twee kralen plus twee kralen is gelijk aan vier kralen.
3. Daarom is er geen reden om te denken dat organische algoritmen dingen kunnen die niet-organische algoritmen nooit kunnen repliceren of overtreffen.

(Zelf)bewustzijn is ook nodig voor empathie, het vermogen om zich te verplaatsen in een ander en zich de gevoelens van die ander voorstellen, zonder dat men die gevoelens per se deelt en meevoelt. Superintelligente robots zonder gevoel worden door Jos de Mul gekwalificeerd als psychopaten, met alle risico's van dien. "Psychopaten met een bovenmenselijke intelligentie, niet gehinderd door de karakterzwakheden van menselijke psychopaten, kunnen - en zullen - de nodige ellende aanrichten". [8]

Zelfbewust of niet, een wezenlijke vraag is hoe kunstmatige breinen hun superieure intellectuele vermogens zullen inzetten. Politiek wetenschapper Charles Rubin gelooft dat AI-breinen niet

[7] Yuval Noah Harari, Homo deus: A brief history of tomorrow, 2016.
[8] "Robots Maken Mensen Overbodig", https://www.trouw.nl/home/robots-maken-mensen-overbodig~af04c799/

ontworpen kunnen worden om uitsluitend goedgezind richting te mens te zijn. Hij stelt zelfs dat het idee van te onderscheiden kwaadwillige of vriendelijke AIs überhaupt geen betekenis heeft: “elke welwillendheid kan niet worden onderscheiden van kwaadwilligheid”.[9] AI-onderzoeker Rodney Brooks schrijft daarentegen: “ik vind het een fout om ons nu zorgen te maken over het feit dat we in de komende paar honderd jaar kwaadwillige AI ontwikkelen. Ik denk dat de zorg voortkomt uit een fundamentele fout, waarbij er geen verschil wordt gemaakt tussen de recente vooruitgang in een bepaald aspect van AI, en de enorme en complexiteit van het bouwen van verstandige intelligentie”.[10]

De opkomst van machines die op en boven het niveau van de menselijke intelligentie presteren betekent niet alleen dat er een samenleving kan ontstaan waarin AI-systemen een belangrijk deel van het werk van de mens overnemen; maar ook dat ons hele mentale zelfbeeld van ‘de mens als koning (en eindpunt) van de schepping’ op kantelen staat. Veel meer *game changing* potentieel is nauwelijks denkbaar.

[9] Charles Rubin, *Artificial Intelligence and Human Nature*, The New Atlantis (Spring 2003), pp88–100.
[10] Rodney Brooks, *Artificial intelligence is a tool, not a threat*, 2014.

3. Technologische Stand van Zaken

Generaties van AI

In de ontwikkeling van AI kunnen we drie niveaus of generaties van steeds toenemend redeneervermogen onderscheiden:[11]

1. **Artificial Narrow Intelligence (ANI)** betreft AI die de menselijke intelligentie voor specifieke taken evenaart of zelfs overtreft.
2. **Artificial General Intelligence (AGI)** betreft AI die het scala van menselijke intellectuele prestaties voor elke taak evenaart.
3. **Artificial Super Intelligence (ASI)** is de aanduiding voor AI die de menselijke intelligentie voor elke taak overtreft.

Van AGI, laat staat van ASI, zijn we zelfs volgens de optimisten nog decennia verwijderd. Maar ANI wordt op steeds meer terreinen gerealiseerd. Langzamerhand beginnen die terreinen elkaar te raken en te overlappen zodat we steeds complexere en uitgebreidere taken aan AI kunnen overlaten; denk bijvoorbeeld aan zelfrijdende auto's. De snelle voortgang van de afgelopen paar jaar is vooral verbonden aan een deelgebied van AI, namelijk *machine learning*: het brede onderzoeksveld binnen AI dat zich bezighoudt met de ontwikkeling van algoritmes en technieken waarmee computers kunnen leren zonder menselijke tussenkomst. En daarbinnen is het weer vooral *deep learning*, software die probeert de activiteit van de menselijke hersenen na te bootsen, die furore maakt. Deze software leert patronen te herkennen in digitale weergaven van geluiden, afbeeldingen en andere gegevens. Het basisidee van een kunstmatig 'neuraal netwerk' die de werking van de menselijke neocortex simuleert is reeds decennia oud. Door verbeteringen in algoritmes en steeds krachtiger computers kunnen computerwetenschappers nu veel meer lagen van virtuele neuronen modelleren. Met deze grotere gelaagdheid is er veel vooruitgang geboekt in spraak- en beeldherkenning.

Hype of hoop?

Er zijn momenteel vier belangrijke drivers achter de snelle vooruitgang in AI-technologie[12]: decennia van exponentiële groei in computergebruik; vooruitgang bij de implementatie van *machine learning* technieken; toenemende beschikbaarheid van grote datasets om *machine learning* op los te laten; en stevige en snel toenemende commerciële investeringen.

Mede omdat de AI-agenda zich, voorlopig althans, concentreert op *machine learning* en ANI-toepassingen, is enige voorzichtigheid op zijn plaats over hoe ver de vooruitgang op AI-gebied kan reiken. In Gartners '*hype cycle*' staat *machine learning* op het topje van wat de 'piek van opgezwollen verwachtingen' wordt genoemd, gedoemd om spoedig terug te vallen in de 'vallei van desillusie' (zie Figuur 2). Tegelijk merkt Gartner *smart machine technologies* aan als "the most disruptive class of technologies over the next 10 years due to radical computational power, near-endless amounts of data, and unprecedented advances in deep neural networks that will allow organizations with smart machine technologies to harness data in order to adapt to new situations and solve problems that no one has encountered previously."

[11] Zie o.a. HCSS, *Artificial intelligence and the future of defense. Strategic implications for small- to medium-sized force providers*, 2017. In deze studie voor het ministerie van Defensie zijn de voortgaande ontwikkelingen in AI in kaart gebracht en de implicaties daarvan voor defensie en veiligheid geduid.
[12] Greg Allen and Chan Tahniel, *AI and National Security*, 2017, http://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf

Figuur 2: Machine Learning in de Gartner *hype cycle* 2016[13]

Per saldo lijkt de belofte dat AI onze wereld gaat veranderen de komende periode toch echt werkelijkheid gaat worden. Over de snelheid waarop is verschil van inzicht, en zeker ook over waar dit toe kan leiden.[14] De opkomst van AI markeert wel een (hernieuwde) aandacht voor wat menselijke intelligentie is. Volgens sommige onderzoekers zijn onze hersenen in feite niets anders dan een hoop algoritmes die op *wetware* werken (als equivalent van de op silicium gebaseerde computerchips). Deze computermetafoor voor hoe de hersenen werken wordt door anderen weer bestreden. Zij stellen dat de 'intelligentie' die AI ten toon spreidt iets heel anders dan de menselijke intelligentie; met als implicatie dat 'de mens' nog lang niet verouderd of achterhaald is.[15] Dit is niet de plaats om dit debat uit te diepen, maar dat de verschillende posities op de lange termijn tot wezenlijk andere toekomstbeelden leiden – en op de middellange termijn tot andere onderzoeksprioriteiten en ethische en juridische overwegingen - mag duidelijk zijn.

Ontwikkelagenda

Generaties AI

Hierboven zijn al de opvolgende generaties van steeds geavanceerdere AI-systemen genoemd: Artificial Narrow Intelligence (ANI) voor specifieke taken; Artificial General Intelligence (AGI) voor alle taken op gelijkwaardig niveau als de mens; en Artificial Super Intelligence (ASI) als de overtreffende trap, iets waar we ons principieel geen voorstelling van kunnen maken. ANI is al realiteit, vaak onopgemerkt, zoals in onze zoekmachines of de assistenten en vertaalprogramma's op onze

13 Zie http://www.gartner.com/newsroom/id/3412017
14 Een radicale visie is de hypothese van 'technologische singulariteit': in een lerend AI-systeem zou op een gegeven moment een kettingreactie van zelfverbeteringscycli kunnen optreden, resulterend in een krachtige superintelligentie die kwalitatief de menselijke intelligentie vele malen zal overtreffen. Dit zal leiden tot (letterlijk) onvoorstelbare veranderingen in de menselijke beschaving. De bekende uitvinder en AI-expert Ray Kurzweil voorspelde dat dit punt (de singulariteit) rond 2045 zal worden bereikt.
15 Zie bijvoorbeeld het werk van de onlangs overleden filosoof Hubert Dreyfus.

mobiele telefoons. Denk ook aan de expertsystemen Deep Blue (schaken) en Watson (Jeopardy!) van IBM en Googles AlphaGo (go) die de wereldkampioenen de baas zijn. ANI is verder gerealiseerd in High-Frequency Trading algoritmen, automatische spam filters enz. Wanneer we de komst van AGI kunnen verwachten is controversieel, maar diverse experts verwachten de volledige ontwikkeling in het midden van deze eeuw. Het is niet ondenkbaar dat ASI relatief snel daarna verschijnt (de singulariteit-hypothese). In deze notitie gaan we vooral in op ANI, de toepassing van AI op specifieke taken binnen afgebakende domeinen. Daarbij houden we wel in het oog dat deze taken en domeinen zich steeds verder uitbreiden, om op termijn (al dan niet geleidelijk) over te gaan in iets dat als AGI kan worden aangemerkt.

Koppeling aan datasets

Het beschikbaar komen van grote hoeveelheden bruikbare data is een belangrijk onderdeel van de AI-revolutie. De veelzijdigheid van open source-tools zoals TensorFlow, Torch en Spark, gekoppeld aan de beschikbaarheid van enorme hoeveelheden rekenkracht via diverse cloud-providers, betekent dat er tegenwoordig eenvoudig op *machine learning*-gebaseerde systemen kunnen worden gebouwd. De flessenhals is dan de beschikbaarheid van goede data en niet zozeer de algoritmen die op de data kunnen worden losgelaten. De ‘democratisering’ van allerlei (big) databestanden, nu vaak nog verstopt of bewust afgeschermd in gesloten systemen, is daarom een noodzakelijke stap om AI-ontwikkelingen te versnellen.

Zwermintelligentie (*swarm intelligence*)

Zwermintelligentie is een vorm van AI gebaseerd op het collectieve gedrag van gedecentraliseerde, zelforganiserende systemen. De samenstellende systemen zijn simpel - dus goedkoop - met een heel beperkt (‘dom’) gedragsspectrum. Het intelligente gedrag ontstaat op het niveau van de ‘zwerm’ van veel van deze samenstellende delen. Er is geen centrale structuur die het intelligente gedrag stuurt, maar dit gedrag ontstaat op macroniveau door de gezamenlijke interacties van de delen We spreken over emergent intelligent gedrag. Dergelijke ontwikkelingen zullen voor zowel nieuwe capaciteiten zorgen als bestaande capaciteiten goedkoper maken voor een breder spectrum van (commerciële) partijen. Ook in het veiligheidsdomein zal 'zeer goedkoop en talrijk', verbonden door relatief simpele samenwerkingsmechanismen, 'zeer duur en schaars' in veel gevallen gaan verslaan. Zo lijkt de business case voor veel goedkope en kleine drones, die samen zwermintelligentie tonen en zo dure hightech platformen overbodig kunnen maken, steeds aantrekkelijker te worden.[16]

Van ondersteuning naar samenwerking

Uit de literatuur blijkt dat AI het komend decennium op vrijwel elk toepassingsgebied een rol gaat spelen. Naarmate computers slimmer worden, gaan we van geautomatiseerde ondersteuning van door mensen uitgevoerde taken naar een model waarin intelligente machines samenwerken met mensen om het volledige potentieel van mens-machine-teams te bereiken. Echte samenwerking van mens en machine vereist wederzijds vertrouwen, niet alleen van de mens in de machine, maar ook van de machine in de mens. In wezen moet de besluiten die mens en machine nemen transparant zijn voor de wederpartij, om vertrouwen te wekken dat de 'juiste' beslissing wordt genomen.

[16] Het US Strategic Capabilities Office van het Department of Defense en het Naval Air Systems Command hebben eerder dit jaar swarming technologie getest. “Dropped by Navy F/A-18 jet fighter-bombers, these tiny unmanned aircraft were able to show collective decision-making, adaptive formation flying, and self-healing. Swarm behavior could enable unmanned aircraft, boats, submarines, and land vehicles to work as teams, fill in for lost or damaged drones, learn from their mistakes, and communicate new courses of action to the collective. The ability for unmanned vehicles to swarm has virtually unlimited potential. It could overwhelm enemy defenses, act as a secure and self-healing communications network, and provide pattern recognition for targeting, reconnaissance, and even search and rescue.” http://navaltoday.com/2017/01/11/video-us-navy-fa-18s-release-drone-swarms-in-technology-demonstration/

Een logische volgende stap is dat we zaken steeds meer overlaten aan kunstmatige breinen. Als we systemen hebben die ons beter kennen dan wij onszelf (zie voorbeeld hieronder), is het niet logisch dat we de autoriteit over beslissingen overdragen aan die systemen?

Een recente studie in opdracht van Facebook heeft aangetoond dat het Facebook-algoritme menselijke persoonlijkheden en neigingen nu al een beter kan beoordelen dan vrienden, ouders en echtgenoten. De studie werd uitgevoerd op 86.220 vrijwilligers met een Facebook-account die een lijst van 100 persoonlijkheidsvragen hebben ingevuld. Het Facebook-algoritme voorspelde de antwoorden van de vrijwilligers op basis van het controleren van hun Facebook Likes - welke webpagina's, afbeeldingen en clips ze hebben gemarkeerd met de Like-knop. Hoe meer Likes, hoe nauwkeuriger de voorspellingen. De voorspellingen van het algoritme werden vergeleken met die van collega's, vrienden, familieleden en echtgenoten. Het algoritme had slechts 10 Likes nodig om de voorspellingen van collega's te verslaan; 70 Likes om beter te presteren dan vrienden, 150 Likes voor familieleden en 300 Likes voor echtgenoten. Met andere woorden, als u 300 likes op uw Facebook-account hebt geklikt, kan het Facebook-algoritme uw meningen en verlangens beter voorspellen dan uw man of vrouw!

Ethische en juridische overwegingen

AI brengt een verscheidenheid aan ethische en juridische vraagstukken met zich mee over hoe een intelligente machine zich moet gedragen tegenover zowel mensen als andere AI-toepassingen. Waarden en begrippen als medemenselijkheid, respect, waardigheid, geluk, mededogen, vrijheid en privacy mensenrechten zullen een rol moeten spelen in het ontwerp, het gebruik en de evaluatie van AI. Ontwerpers en ingenieurs zullen steeds vaker gevraagd (moeten) worden om hun ontwerp toe te lichten en te verdedigen in termen van de genoemde morele waarden.[17] Dit probleem is uiteengezet door Wendell Wallach in zijn boek Moral Machines, waarin hij het concept van kunstmatige ‘morele agenten’ (Artificial Moral Agents, AMA) introduceerde. Voor Wallach zijn AMA's onderdeel van het onderzoekslandschap van AI, geleid door twee centrale vragen: "does humanity want computers making moral decisions" en "can (ro)bots really be moral".[18]

Controle en toezicht

Bovenstaande noopt tot vormen van toezicht op AI-ontwikkelingen. We gaan hier in het hoofdstuk *So what voor veiligheid* nader op in.

De noodzaak van experimenteren

Het is onmogelijk om te voorspellen hoe de toekomst van AI eruit gaat zien, maar AI heeft het potentieel om elk bedrijfs- en verdienmodel dat we vandaag kennen op zijn kop te zetten. We weten alleen nog nauwelijks hoe. Om te overleven moeten bedrijven nu beginnen met experimenteren of het risico lopen om weggeconcurreerd te worden door ondernemingen die er in slagen zijn om AI efficiënt en effectief toe te passen.

AI begint een essentieel onderdeel te worden voor onze dagelijkse routines. Met digitale assistenten zoals Siri en Alexa zijn we op huiselijke schaal al aan het uitvinden wat handig werkt en wat niet, waar (nu) grenzen liggen en waar we graag die grenzen zouden willen oprekken. Ook in

17 http://agconnect.nl/artikel/de-ethiek-van-kunstmatige-intelligentie
18 Wendell Wallach, Moral Machines, 2010.

professionele omgevingen is Concept Development & Experimentation (CD&E) onontkoombaar. Men moet vertrouwd raken met de technologie: hoe werkt het, wat kunnen we ermee en welke valkuilen en beperkingen zijn er? Omdat AI een generieke technologie is, niet specifiek ontwikkeld voor een bepaald toepassingsgebied, is de vraagstelling iedere keer weer vergelijkbaar maar zijn de antwoorden mogelijk net verschillend.

CD&E begin vaak met een proof of concept - experimenten om te bewijzen dat de technologie werkt. Al snel moeten ook de gebruikers worden betrokken: dekt het een (eventueel latente) behoefte, werkt het in de praktijk? Sowieso is een gezamenlijke inspanning essentieel, van de programmeurs die nieuwe data mining-technieken ontwikkelen tot aan aan de ondernemers die bereid zijn om iets anders te proberen. Bedrijfsdirecties en investeerders zijn cruciaal voor succes als zij innovators kunnen koppelen, bereid zijn om middelen te investeren en openstaan voor mislukkingen. Een eerste stap is om te investeren in het ‘leren’, om het 'universum van de mogelijke' te verkennen, om te begrijpen wat er nu gedaan kan worden en wat er in de toekomst kan gebeuren.[19]

De ervaring leert dat laboratoriumsituaties of zelfs ‘live’ pilotprojecten wel in staat stelt de meer voor de hand liggende veiligheidsproblemen in nieuwe ICT-toepassingen te ontdekken en te corrigeren, maar dat de noodzakelijke inkadering juist de ‘unknown unknowns’ buitensluit. Veel veiligheidslekken treden pas op tijdens daadwerkelijk gebruik en niet in experimenten, hoe uitgebreid ook. Wel is het goed mogelijk dat experimenteren de (mentale) flexibiliteit verhoogt om sneller fouten te her- én erkennen, en naar een oplossingsmodus om te schakelen.

De overheid moet overwegen hoe regelgeving kan bijdragen aan een omgeving waarin nieuwe operationele AI-toepassingen goed kunnen worden onderzocht teneinde overheden en bedrijven in staat te stellen om te experimenteren, zowel in gecontroleerde omgevingen als in de feitelijke praktijk, waarbij niet kan worden gewacht op 'perfecte' oplossingen.

[19] Nigel Duffy, *AI in Your Organization: Begin Now by Experimenting, and Learning*, 31 mei 2017, https://www.linkedin.com/pulse/ai-your-organization-begin-now-experimenting-learning-nigel-duffy

4. Implicaties voor de Samenleving

Ondersteuning van digitale assistenten

AI kan ons leven een stuk makkelijker maken. Dit is op dit moment vooral zichtbaar in de opkomende digitale 'persoonlijke assistenten' die ons door een complexe maatschappij proberen te loodsen. Siri is de stemgeactiveerde, pseudo-intelligente PA van Apple waarmee velen van ons dagelijks communiceren. Zij helpt ons om informatie te vinden, afspraken in onze agenda's te zetten en berichten te verzenden. Siri maakt gebruik van *machine learning*-technologie om vragen en verzoeken van individuen beter te voorspellen en te begrijpen. Alexa van Amazon is een vergelijkbare tool. Alexa's opkomst als slimme hulp in huis is het afgelopen jaar snel gegaan. Het apparaat helpt ons om het web te af te zoeken voor informatie, te winkelen, afspraken te maken en het huisalarm in te stellen. Alexa vormt een behoorlijke eerste stap om, in combinatie met een oprukkend IoT, het toekomstig 'slimme huis' te verwezenlijken.

Meer specifiek gerichte digitale PAs zijn wearables zoals AppleWatch of FitBit die de gezondheid van een persoon monitoren. In toenemende mate zijn dergelijke apparaten zowel met elkaar als met het internet verbonden. Hierdoor kan de gezondheid van een patiënt desnoods real-time in de gaten worden gehouden; een verhoogde hartslag of verlaagde bloeddruk kan zorgen dat een alarm afgaat bij een arts. Ook kunnen bepaalde onderzoeken die nu nog in het ziekenhuis plaatsvinden elektronisch op afstand worden uitgevoerd. Bovendien kan een patiën, met behulp van data-analyse gepersonaliseerde gezondheidszorg ontvangen.

Tegelijk ziet we de opkomst van '*cuddly robots*' die ook belangrijke sociale functies vervullen en als metgezel van de mens optreden. Nu wellicht nog als een soort huisdier, maar steeds meer als volwaardige gesprekspartner, begeleider en helper in één. Een toekomst waarin digitale assistenten en digitale metgezellen geïntegreerd worden in digitale 'buddies' die ons van kinds af aan vergezellen, assisteren en mee-ontwikkelen gedurende onze levensloop is zeer wel voorstelbaar.

AI en het Internet of Things

Het Internet der Dingen (Internet of Things, IoT), de online verbondenheid van alledaagse apparaten, stelt ons in staat om enorme hoeveelheden gegevens te verzamelen over talloze activiteiten en processen. Maar welke inzichten zitten verborgen in de grote en groeiende hoeveelheden data? Voor mensen is het simpelweg onmogelijk om al deze gegevens te beoordelen en te begrijpen. De enige manier om deze IoT-gegenereerde gegevens te kunnen verwerken richting *actionable intelligence* is met AI - om zo stedelijke overheden te helpen bij het voorspellen van ongelukken en misdrijven, artsen real-time inzicht te geven in hoe pacemakers functioneren of zinvolle communicatie tussen zelfrijdende auto's te bewerkstelligen.[20]

De combinatie van grootschalige 'dataficatie' en AI kent ook nadelen. Zo kunnen AI-systemen beslissingen nemen op basis van big data zonder dat duidelijk is welke gedachtegang en afweging hieraan vooraf gingen. Wanneer deze beslissing niet de juiste blijkt is het vervolgens moeilijk te zorgen dat het een volgende keer beter gaat. Lag het aan de data - en welke dan? - of aan de redenatie? Zeker als de beslissing ingrijpende ongelukken tot gevolg heeft, waar ligt de verantwoordelijkheid voor het falen? Ligt deze bij de fabrikant van de machine die draait op AI-technologie? Of is de eigenaar of de operator van de machine verantwoordelijk?[21]

20 https://www.wired.com/insights/2014/11/iot-wont-work-without-artificial-intelligence/
21 https://qz.com/989137/when-a-robot-ai-doctor-misdiagnoses-you-whos-to-blame/

‘Bots’ in de media

AI kan ook worden ingezet voor mediamanipulatie. Online ‘bots’ verspreiden (nep)berichten via sociale media en passen lerende algoritmes toe om het publieke en politieke debat te beïnvloeden. Bots kunnen worden geprogrammeerd om eenzijdige politieke boodschappen breed te verspreiden en de illusie te wekken van publieke steun. Dit lijkt een steeds breder ingezette tactiek om de publieke discussie te vormen en te sturen. Zo zijn er aanwijzingen dat AI-technologie systematisch misbruikt is om de burgers te manipuleren tijdens de recente Amerikaanse[22] en Franse[23] presidentsverkiezingen[24].

Naast het beïnvloeden van het (online) debat kan AI ook worden gebruikt om individuele kiezers te te manipuleren. Tijdens de Amerikaanse presidentsverkiezingen heeft het data science bedrijf Cambridge Analytica een reclamecampagne uitgewerkt om de emoties van zwevende kiezers op basis van hun individuele psychologie te beïnvloeden. Deze geavanceerde operatie op microniveau was gebaseerd op grote hoeveelheden data en machine learning. Het probleem met het gebruik van AI in politieke campagnes is niet de technologie zelf, maar eerder de verborgen motivatie van het gebruik ervan en de doelgerichte berichten die de psychologische kwetsbaarheid van individuen uitbuiten. Dit is een verontrustende trend die politieke en maatschappelijke onrust of zelfs instabiliteit kan veroorzaken. Tegelijk kunnen dezelfde algoritmen die worden gebruikt om te misleiden en te misinformeren ook worden ingezet om maatschappelijke betrokkenheid te vergroten. Een ethische benadering van AI kan helpen om een kiezers te informeren en te betrekken. Nieuwe startups zoals Factmata en Avantgarde Analytics bieden deze technologische oplossingen al. Het gebruik van AI in verkiezingscampagnes zal niet verdwijnen, daarvoor is het al te waardevol gebleken. Het is daarom belangrijk dat AI in een ethisch kader geplaatst wordt.

AI en *jobless growth*

AI-toepassingen zijn momenteel vooral hulpmiddelen, ondersteunend bij het werk dat door mensen wordt uitgevoerd. Op die manier kunnen schaal- en efficiencyvoordelen worden behaald. Maar steeds vaker zullen AI-applicaties menselijk handelen en denken gaan vervangen. Twee parallelle ontwikkelingen tekenen zich af. Enerzijds een concentratie van grotere rijkdom in de handen van bedrijven die toegang hebben tot en goed gebruik (kunnen) maken van AI-technologie, en anderzijds een mogelijk grootschalige afname van het aantal beschikbare banen: *jobless growth*. De sociale en economische modellen die uit het industriële tijdperk stammen kunnen deze disruptieve veranderingen niet goed aan.[25]

Steeds meer bedrijven gaan over tot automatisering om efficiënter te kunnen opereren. Dit zet de rol van de werknemer en het begrip ‘arbeid’ onder druk. Maar ook het gevoel van waarde en voldoening dat mensen uit hun werk halen komt in het gedrang. In de literatuur wordt er verschillend gedacht over de percentages van banen die in de toekomst worden overgenomen door AI-gebaseerde automatisering, maar dat er banen zullen verdwijnen is duidelijk. Op basis van een uitgebreide studie onder 702 verschillende beroepen concludeerden Frey en Osborne dat de transportsector (o.a. taxichauffeurs en banen in de logistieke sector) en office support functies (denk aan receptionistes en beveiligers) een grote kans hebben te worden weggeautomatiseerd. Daarnaast

22 “On Twitter, a Battle Among Political Bots”, https://www.nytimes.com/2016/12/14/arts/on-twitter-a-battle-among-political-bots.html
23 “#MacronLeaks Changed Political Campaigning. Why Macron Succeeded and Clinton Failed”, https://www.weforum.org/agenda/2017/05/macronleaks-have-changed-political-campaigning-why-macron-succeeded-and-clinton-failed
24 Adam Segal, Artificial Intelligence Has the Power to Destroy or Save Democracy, Council on Foreign Relations, 7 augustus 2017, https://www.cfr.org/blog/artificial-intelligence-has-power-destroy-or-save-democracy.
25 Zie ook Elon Musk, Automation Will Force Universal Basic Income , Juni 2017, https://www.geek.com/tech-science-3/elon-musk-automation-will-force-universal-basic-income-1701217/

zijn ook veel werknemers in de verkoop- en dienstensector (kassamedewerkers, frontdesk medewerkers en accountants) vatbaar voor de effecten van AI. Ze schatten dat 47% van de banen de in VS in de komende twee decennia worden geautomatiseerd.[26] Tegelijk zal AI ook nieuwe banen creëren. Als auto's zelf kunnen rijden hebben mensen tijdens het autorijden meer tijd om andere dingen te doen en om producten of diensten te consumeren. Bovendien zullen ook zelfrijdende auto's onderhoud nodig blijven hebben. Tevens kan AI bijdragen aan vergroting van de productiviteit, wat op zich weer tot welvaartscreatie kan leiden en de basis kan vormen voor banengroei. Hoe de balans tussen vernietiging van bestaand werk en creatie van andersoortige banen zal doorslaan is onduidelijk.

Probability that computerisation will lead to job losses within the next two decades, 2013
(1=certain)

Job	Probability
Recreational therapists	0.003
Dentists	0.004
Athletic trainers	0.007
Clergy	0.008
Chemical engineers	0.02
Editors	0.06
Firefighters	0.17
Actors	0.37
Health technologists	0.40
Economists	0.43
Commercial pilots	0.55
Machinists	0.65
Word processors and typists	0.81
Real estate sales agents	0.86
Technical writers	0.89
Retail salespersons	0.92
Accountants and auditors	0.94
Telemarketers	0.99

Source: "The Future of Employment: How Susceptible are Jobs to Computerisation?" by C.Frey and M.Osborne (2013)

Figuur 3: Verdwijnende banen

In tegenstelling tot de industriële revolutie en de computerrevolutie, zal de AI-revolutie niet alleen bepaalde typen banen raken. Ook hoogopgeleide beroepsgroepen, zoals radiologen die tumoren opsporen in bodyscans, worden vervangen door machines die dit accurater kunnen. Tegelijkertijd is het zo dat door middel van AI bepaalde processen sneller kunnen worden uitgevoerd. Denk hierbij onder andere aan het uitvoeren van documentanalyses (bijvoorbeeld door juridische assistenten). Omdat dit sneller kan worden gedaan als het proces is geautomatiseerd, betekent dit ook dat de

[26] Carl Benedikt Frey en Michael A. Osborne, *The Future of Employment: How Susceptible Are Jobs to Computerisation?*, Technological Forecasting and Social Change 114 (2017), p.44

vraag naar dergelijke diensten kan stijgen, waardoor *an sich* meerdere assistenten nodig zijn.[27] AI lijkt vooralsnog minder geschikt voor banen die creativiteit, empathie en interdisciplinair denken vergen. Voorbeelden hiervan zijn mantelzorgers of maatschappelijk medewerkers; werk dat een genuanceerde menselijke interactie vereist.

Een oplossing voor het potentiële probleem van massawerkloosheid kan zijn "werk van liefde". Dit type banen, momenteel veelal vervuld door vrijwilligers, kunnen moeilijk door AI worden overgenomen (of anders bewust worden ontzien), en kunnen mensen een gevoel van voldoening geven.[28] Dit zou een nieuwe 'arbeidersklasse' kunnen creëren.[29] De uitdaging is dergelijke 'vrijwilligersbanen' in betaalde banen om te vormen.

Wat verder blijft is dat een relatief kleine groep juist profiteert van de AI-revolutie. Hoe kan de welvaart in een dergelijk asymmetrisch economisch systeem acceptabel verdeeld worden? Een deeloplossing zou een 'universeel basisinkomen' (UBI) kunnen zijn. Dit idee bestaat al sinds Thomas More's Utopia (1516) en stelt dat de staat een vast bedrag aan elk van zijn burgers verstrekt, ongeacht hun inkomen of werkstatus. Het kan een effectieve manier zijn om de vruchten van technologische vooruitgang redelijk te verdelen. Het concept van de 'verzorgingsstaat' krijgt hiermee een andere invulling krijgen en brengt de overheid in een andere positie. UBI heeft veel aanhangers in wetenschap en technologie.[30] Elon Musk pleitte al meerdere malen voor een universeel basisinkomen: "I think some kind of universal income will be necessary. The harder challenge is how do people then have meaning — because a lot of people derive their meaning from their employment." Er zijn ook criticasters van UBI. Zij betogen dat het huidige UBI-idee niet 'universeel' is. Alle tot nu voorgestelde UBI-initiatieven zijn immers strikt nationaal of lokaal.

Praktijken én normen verschuiven: wetgeving moet aanhaken

De huidige golf van automatisering creëert een kloof tussen de bestaande wetgeving en de veranderde realiteit op de werkvloer. Nieuwe arbeids- en arbeidswetgeving is dringend nodig om gelijke tred te houden. AI creëert niet alleen een kloof tussen wetgeving en de werkvloer, maar verstoort bijvoorbeeld ook wetgeving en regulering in het verkeer. Sinds 2016 hebben steden over de hele wereld verkeerswetgeving opgesteld om bestuurderloze auto's op de wegen toe te laten.

Wat filosofischer: de AI-revolutie heeft grote invloed op het denken over de positie en rol van de mens in 'de schepping'. De mogelijkheid van het ontstaan van kunstmatige breinen met intellectuele vermogens die uitstijgen boven die van de mens, roept vragen op over de verhoudingen tussen mens en machine, maar relativeert ook het onderscheid tussen mens en dier. Zo heeft de opkomst van AI gevolgen voor, bijvoorbeeld, ons denken over dierenwelzijn. Maar ook: waar liggen aansprakelijkheden als we steeds meer beslissingen aan kunstmatige breinen overlaten? Wet- en regelgeving, als codificering van onze normen en waarden, is op een breed front onderhevig aan veranderingen ten gevolge van de AI-revolutie. Een actieve wetgever is gewenst: "AI is een van de weinige gevallen waarin we proactief moeten reguleren in plaats van reactief. Want als we bij kunstmatige intelligentie alleen reactief handelen, is het snel te laat."

27 https://www.economist.com/news/special-report/21700758-will-smarter-machines-cause-mass-unemployment-automation-and-anxiety.
28 "Opinion | The Real Threat of Artificial Intelligence", https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html
29 "Universal Basic Income Is Neither Universal Nor Basic", https://www.bloomberg.com/view/articles/2017-06-04/universal-basic-income-is-neither-universal-nor-basic
30 "Robots Are Taking Our Jobs", http://www.theceomagazine.com/business/robots-are-taking-our-jobs/

5. *So What* voor de veiligheid

Patroonherkenning om onveiligheid te bestrijden

Hieronder gaan we vooral in op de risico's en bedreigingen die van AI kunnen uitgaan. Maar laten we niet vergeten dat AI ook veiligheid kan bevorderen. Het berichtje 'Brabantse speurders krijgen hulp van big data Crime Room'[31] is er een in een hele serie van vergelijkbare berichten. Slimme algoritmen leren patronen te herkennen die kunnen duiden op crimineel gedrag, in sommige gevallen zelfs al voordat er sprake is van een misdaad. AI is daadwerkelijk bezig om de praktijk van opsporing én preventie van misdaad te veranderen. Ook voor de internationale veiligheid wordt gekeken naar lerende systemen om ondermijnende acties van opponenten of opkomende crises beter en eerder te onderkennen. Dit stelt in staat om sneller te reageren en potentieel gevaarlijke situaties in de kiem te smoren.

Dergelijke *early warning*- en signaleringssystemen kennen diverse uitdagingen. Een belangrijke vraag is wanneer 'indicaties' dat er iets mis kan zijn overgaan in de 'relatieve zekerheid' dat er sprake is van ontoelaatbare of vijandelijke acties en die voldoende - wettelijke, politieke en/of morele - grond biedt om tot actie over te gaan. In sommige gevallen, met name in het cyber- en het informatiedomein, komt daar het attributieprobleem bij. Zelfs als het duidelijk is dat er geknoeid wordt met de systemen is het vaak lastig hard bewijs te leveren wie er precies achter zit. Deze uitdaging is des te groter bij een gelaagde dreiging: een statelijke actor die gebruik maakt van niet-statelijke groeperingen (zogenaamde 'proxies' zoals lokale warlords of hackercollectieven) om zijn doelen te bereiken. Een dergelijke tactiek is onderdeel van wat we tegenwoordig 'hybrid warfare' noemen: het op allerlei manieren trachten een tegenstander te beïnvloeden zonder hier openlijk voor uit te komen. Dit zet de besluitvormers voor het blok: wanneer is het genoeg en zijn (tegen)maatregelen gerechtvaardigd? Wat zijn rode lijnen en hoe worden die gecommuniceerd? Zeker in de internationale veiligheid, waarbij veelal optreden in coalitieverband aan de orde is, zijn dit uiterst lastige beslisprocessen.

Gebruik van AI door tegenstanders

Veiligheidsorganisaties gebruiken patroonherkenning om criminelen of tegenstanders te bestrijden, maar omgekeerd kunnen ook opponenten AI inzetten. We hebben al het voorbeeld genoemd van 'bots' die het publieke en politieke debat beïnvloeden, en op een verborgen manier verkiezingen kunnen manipuleren. Bots worden ook gebruikt om met het internet verbonden systemen - en welk systeem is dat tegenwoordig niet - af te tasten op zwakke plekken, om daar vervolgens misbruik van te maken voor criminele doeleinden.[32]

De ontwikkeling van gemilitariseerde AI is een van de grootste zorgen. Momenteel onderzoeken meer dan 50 landen toepassingen voor robots op het slachtveld, waaronder de Verenigde Staten, China, Rusland en het Verenigd Koninkrijk. Veel mensen die zich bezighouden met de risico's van AI willen het gebruik van autonoom opererende wapenplatformen beperken: "Lethal autonomous weapons threaten to become the third revolution in warfare. Once developed, they will permit armed conflict to be fought at a scale greater than ever, and at timescales faster than humans can comprehend. These can be weapons of terror, weapons that despots and terrorists use against

[31] "Brabantse speurders krijgen hulp van big data Crime Room", https://ibestuur.nl/nieuws/brabantse-speurders-krijgen-hulp-van-big-data-crime-room

[32] Uiteraard kan hetzelfde gebeuren juist met het oog op het *dichten* van lekken. Vergelijk het onderscheid tussen 'black' en 'white' (of 'ethische') hackers.

innocent populations, and weapons hacked to behave in undesirable ways. We do not have long to act. Once this Pandora's box is opened, it will be hard to close. We therefore implore the High Contracting Parties to find a way to protect us all from these dangers."[33]

De (on)veiligheid van AI-systemen

Bij een toenemende afhankelijkheid van intelligente systemen is een garantie voor de integriteit en veiligheid van deze systemen cruciaal. Met enige regelmaat verschijnen berichten van AI-systemen die de mist in gaan: denk aan de beruchte Microsoft Nazi-twitterbot,[34] het geval waar zelfrijdende auto's de weg kwijtraken door stickers op borden,[35] of de security robot die verdrinkt in een Amerikaans winkelcentrum.[36] Hoewel de gevolgen soms fors kunnen zijn, zijn dit veelal gewone softwarefouten die ook in niet-intelligente systemen kunnen optreden.

Ook voor de mogelijkheden om AI-systemen te hacken geldt in grote lijnen dat de risico's vergelijkbaar zijn met het hacken van andere programmatuur. Voor een lerend AI-systeem dreigt er echter een belangrijke extra gevaar, namelijk in het moedwillig veranderen van de data op basis waarvan het systeem leert. Zonder dat er iets mis is met de algoritmes, kunnen zo toch verkeerde patronen aangeleerd worden. Het is dan, bijvoorbeeld, mogelijk om de detectiealgoritmes van creditcardmaatschappijen om de tuin te leiden in het toch accepteren van frauduleuze betalingen. Merk verder op dat ook zonder dat de data gecorrumpeerd wordt er een valkuil dreigt voor lerende algoritmes op basis van big data. Individuele afwijkingen (ver) van een statisch gemiddeld patroon worden gesignaleerd, ook als ze volstrekt toelaatbaar zijn. Wanneer dit automatisch tot actie leidt zonder dat er een gelaagde (al dan niet menselijke) controle plaatsvindt, kan dit problemen geven. Ook hele goede AI-systemen zullen dergelijke 'blind spots' hebben die kwetsbaarheden creëert.

Nu we besluitvorming naar AI-systemen beginnen over te hevelen, dringen technici en onderzoekers aan op het verkrijgen van diepere kennis van de kwetsbaarheden in de mysterieuze zone tussen 'input' en 'output' van AI-systemen. Niet alleen lijken AI-systemen soms foute beslissingen te nemen, maar is het soms ook nog onmogelijk om, één, precies te controleren óf het wel een fout was; twee, waar de fout dan in zat; en drie, om correcties te maken die er voor zorgen dat de fout niet nogmaals optreedt.

Op 6 mei 2010 onderging de beurs van New York een grote schok. Binnen vijf minuten daalde de Dow Jones met 1.000 punten en verloor $ 1 biljoen aan waarde. De beurs veerde weer terug naar zijn pre-crash niveau in iets meer dan drie minuten. Dat is wat er gebeurt als supersnelle computerprogramma's verantwoordelijk zijn voor ons geld. Deskundigen hebben sindsdien geprobeerd te begrijpen wat er gebeurd is in deze zogenaamde 'Flash Crash'. We weten dat algoritmen de schuld zouden hebben, maar we weten nog steeds niet precies wat er fout is gegaan. Een aantal snelle geesten heeft veel verdiend aan dit incident.

[33] "An Open Letter to the United Nations Convention on Certain Conventional Weapons.", https://futureoflife.org/autonomous-weapons-open-letter-2017/
[34] "Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours", http://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/
[35] "Confusing Self-Driving Cars by Altering Road Signs - Schneier on Security", https://www.schneier.com/blog/archives/2017/08/confusing_self-.html
[36] "Security Robot 'in Critical Condition' after Nearly Drowning on the Job.", http://www.cnn.com/2017/07/18/us/security-robot-drown-trnd/index.html

De dreiging van AI-systemen zelf

"The development of full artificial intelligence could spell the end of the human race. Once humans develop artificial intelligence, it will take off on its own and redesign itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete and would be superseded".[37]

In zijn boek *Superintelligence: Paths, Risks, Strategies* (2014), stelt de filosoof Nick Bostrom dat 'echte' kunstmatige intelligentie (AGI), als het wordt gerealiseerd, een gevaar kan vormen dat elke vroegere bedreiging van technologie, inclusief kernwapens, overstijgt. Als de mensheid een dergelijk ontwikkeling niet zorgvuldig managet, orkestreert het zijn eigen uitsterven. Het centrale zorgpunt in zijn redenering is dat een lerende AGI zichzelf zo snel kan verbeteren dat in heel korte tijd een superintelligentie (ASI) ontstaat met een intellectueel potentieel dat het menselijke brein met vele ordegroottes verslaat. Een dergelijk systeem zou een nieuwe vorm van leven inhouden. Bostroms angst is Darwiniaans: de mensheid wordt door een slimmere concurrent verdrongen. Het analogon is mensen en gorilla's: nauw verwante primaten, maar met één soort die de planeet domineert en de andere aan de rand van vernietiging.

De meningen van AI-experts zijn verdeeld, maar het merendeel is inderdaad bezorgd over de risico's van een eventuele ontwikkeling van AI toepassingen die de mens zullen voorbij streven in intelligentie.[38] Deze zorg heeft onlangs aandacht gekregen na verschillende media optreden van vooraanstaande experts zoals Stephen Hawking, Bill Gates en Elon Musk.

Als het idee van een super-AI die de mensheid uitroeit te vergezocht is, hier een dichter bij huis-voorbeeld hoe AIs ons leven kunnen gaan bepalen.
Cortana is de persoonlijke AI assistent van Microsoft en vormt een integraal onderdeel van Windows. Gebruikers worden aangemoedigd om Cortana toegang te verlenen tot al hun bestanden, e-mails en applicaties, zodat ze hen kan leren kennen en advies kan geven. Ze kan je eraan herinneren om iets te kopen voor de verjaardag van je vrouw (en zelfs het cadeau uitkiezen), een tafel in het restaurant reserveren en je vragen je medicijn een uur voor het eten te nemen. Het kan je waarschuwen dat als je niet ophoudt met lezen, je te laat komt voor een belangrijke zakelijke bijeenkomst. Cortana kan je waarschuwen dat je bloeddruk te hoog is en je dopamine-niveau te laag en je, op basis van statistieken over de afgelopen jaren, in deze toestand ernstige fouten kan maken. Dus doe voorzichtig en onderteken niets!
Zodra Cortana evolueert van 'orakel' naar 'agent', kan ze namens haar meester rechtstreeks met de Cortana's van anderen gaan praten. Het kan onschuldig genoeg beginnen: mijn Cortana neemt contact op met jouw Cortana om een afspraak te maken over plaats en tijd van een vergadering. De volgende stap is dat een potentiële werkgever me vertelt dat ik geen cv hoef te sturen, maar simpelweg zijn Cortana met mijn Cortana laat praten over wie ik ben en wat ik kan. Of mijn Cortana wordt benaderd door de Cortana van een potentiële minnaar, om na te gaan of er een mooie relatie kan gaan bloeien - zonder dat hun menselijke eigenaren het weten.
Naarmate Cortana's meer autoriteit krijgen, kunnen ze elkaar beginnen te manipuleren om de belangen van hun meesters te vergroten, zodat succes op de arbeidsmarkt of de huwelijksmarkt in toenemende mate afhangt van de kwaliteit van je Cortana. Rijke mensen die de meest geavanceerde Cortana bezitten, hebben een voordeel ten opzichte van armere mensen met hun oudere versies.

37 "Stephen Hawking Warns Artificial Intelligence Could End Mankind", http://www.bbc.com/news/technology-30290540
38 Vincent C.Müller en Nick Bostrom, Future Progress in Artificial Intelligence: A Poll Among Experts, AI Matters (1), 2014.

AI principes en veiligheid

Een belangrijk deel van de leidende onderzoekers, ontwikkelaars en toepassers van AI maakt zich zorgen over de mogelijke gevolgen van de AI-revolutie. Een consortium waarin de verschillende belangengroeperingen zitting hebben heeft een aantal principes vastgesteld om met de veiligheidsrisico's en uitdagingen van AI om te gaan. De *Asilomar AI Principles* vormt een lijst van 23 richtlijnen die wetenschappelijke onderzoekers, wetenschappers en wetgevers moeten aanhouden voor het veilige, ethische en voordelige gebruik van AI. De principes zijn (nog) niet wettelijk vastgelegd, maar zijn bedoeld om de manier waarop onderzoek gedaan wordt te beïnvloeden.

Asilomar AI Principles[39]

Artificial intelligence has already provided beneficial tools that are used everyday by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.

Research Issues

1. Research Goal: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
2. Research Funding: Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies
3. Science-Policy Link: There should be constructive and healthy exchange between AI researchers and policy-makers.
4. Research Culture: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
5. Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Ethics and Values

6. Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
7. Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.
8. Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
9. Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
10. Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
11. Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
12. Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
13. Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
14. Shared Benefit: AI technologies should benefit and empower as many people as possible.
15. Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

39 https://futureoflife.org/ai-principles/

16. Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
17. Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
18. AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

Longer-term Issues

19. Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
20. Importance: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.
21. Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
22. Recursive Self-Improvement: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
23. Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

Bundeling én distributie van beveiligingskennis

Zoals voor veel ICT-systemen geldt ook voor AI-toepassingen dat individuele gebruikers en veel organisaties niet in staat zijn om alle veiligheidsaspecten te overzien; laat staan afdoende maatregelen te nemen en alle ontwikkelingen bij te houden. Er ontstaan centrale punten waar beveiligingskennis gebundeld wordt om op basis van deze kennis beveiligingsproducten en -diensten te leveren. Idealiter ontstaat een markt van dergelijke kenniscentra (in een mix van bedrijven en publieke instanties) die elkaar scherp houden en aanvullen. Het gevaar bestaat echter dat schaalvoordelen het nieuwe toetreders zeer lastig maken, waarmee de diversiteit van deze markt onder druk komt; een dergelijke tendens is inderdaad zichtbaar in (segmenten van) de cyber security markt. Een actief overheidsbeleid om diversiteit van de markt te waarborgen is dan gewenst, zoals dat onder meer in het Verenigd Koninkrijk gebeurt.[40]

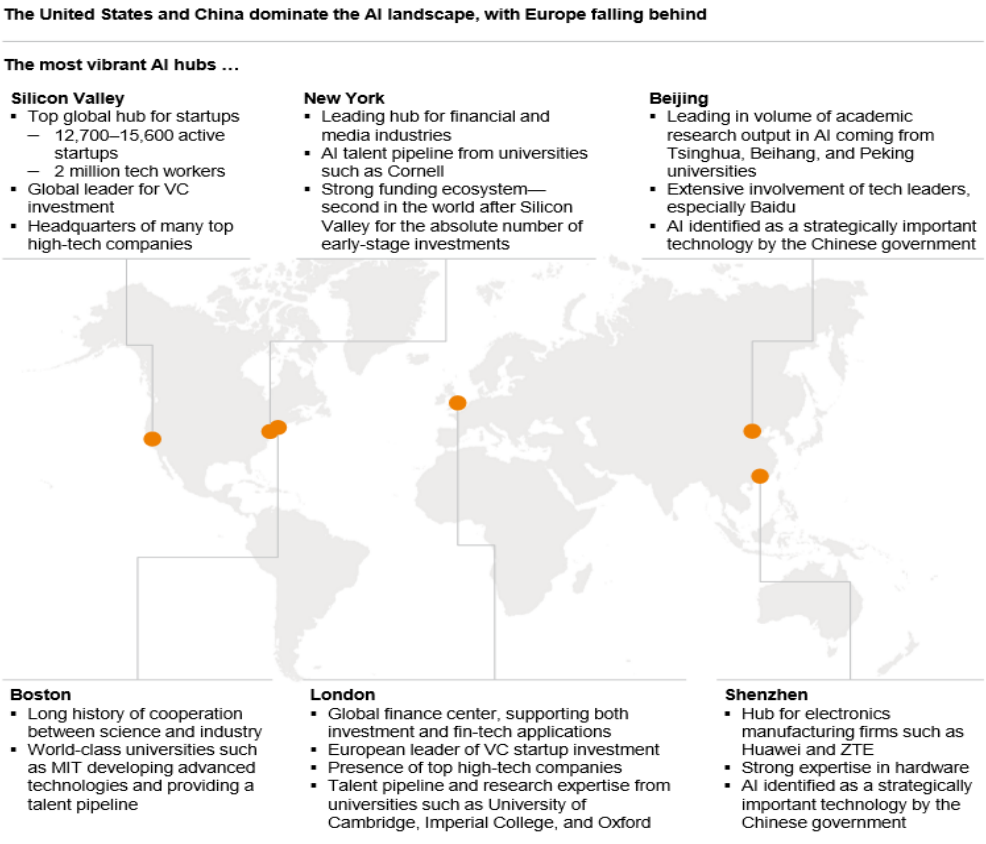
Bredere gevolgen voor maatschappelijke veiligheid

AI-systemen krijgen een steeds belangrijker rol in onze maatschappij. Wat kan dit in algemene zin betekenen voor de aard van de grote veiligheidsuitdagingen waarmee onze samenleving te maken heeft?

AI biedt de mogelijkheid om bestaande taken efficiënter en effectiever uit te voeren en door ondersteuning van besluitvormingsprocessen deze te verbeteren en te versnellen. In het veiligheidsdomein gaat het bijvoorbeeld om taken met zogenaamde ‘3D’-karakteristieken: *Dull* (bijvoorbeeld 24/7 bewakingstaken), *Dirty* (in een extreem klimaat) en *Dangerous* (in een vijandelijk omgeving). Intelligente machines kunnen in steeds complexere omgevingen steeds geavanceerdere ‘3D+’-taken aan en daarbij steeds autonomer werken. In het cyberdomein bijvoorbeeld is snelle herkenning van ongewenste toegang tot netwerken en manipulatie van informatie cruciaal.

[40] Zie onder meer “UK intelligence agencies turn to start-ups on cyber security”, https://www.ft.com/content/6cdfa82e-77bd-11e7-a3e8-60495fe6ca71

Aan de andere kant kan grootschalige toepassing van AI het fundamentele weefsel van onze maatschappij ingrijpend beïnvloeden en zo, als we niet uitkijken, bijdragen aan sociale ontwrichting en destabilisatie van onze samenleving. De schattingen over de timing en precieze percentages verschillen, maar dat de komende twee decennia een fors deel van de huidige banen door AI uitsterft lijkt een gegeven. Dit zijn ook banen waarbij niet-routinematige cognitieve vaardigheden nodig zijn, dus niet alleen lopende band werk. Een mogelijk volgende stap waarin AI niet alleen menselijke banen maar ook de mensheid als geheel overbodig maakt is veel controversiëler. Maar ook zonder deze stap is het game changing karakter van AI al immens. Werk betekent immers niet alleen inkomen maar ook de mogelijkheid tot zelfontplooiing en een nuttige bijdrage te leveren aan de maatschappij. Als een belangrijk deel van het banenbestand in de samenleving verloren gaat, zal er voor zowel de financiële als de sociale kant een oplossing gevonden moeten worden. Dit zal een forse aanpassing betekenen, en mogelijk zelfs een fundamentele omwenteling, in de processen en structuren waarop onze maatschappij is gebaseerd. Ongetwijfeld zal dit, zeker in een transitieperiode, tot sociale onrust en mogelijk tot sociale en politieke instabiliteit leiden.



Figuur 4: de mondiale dynamiek in technologische ontwikkelingen op het gebied van AI[41]

Ook in de geopolitieke ordening zijn er spanningen te verwachten als gevolg van de AI-revolutie. Zoals kolen- en staalproducerende landen in de begintijd van de industriële revolutie hun (economische en daarmee politieke) macht zagen toenemen, en later het bezit en de exploitatie van oliebronnen een machtsinstrument werd, zo heeft de AI-revolutie een soortgelijk potentieel om wereldwijde machtsverhoudingen te veranderen. Veel van de AI-technologieën worden momenteel bijna uitsluitend in de VS en China ontwikkeld (zie Figuur 4). Beide landen investeren met zowel

[41] Mc Kinsey Global Institute, Artificial intelligence. The next digital frontier?, 2017.

publieke als private middelen fors in het verder brengen van AI-kennis en -toepassingen. Het bezit van (superieure) AI-technologie en -kennis kan deze landen in een politieke en economische monopoliepositie brengen. Alle belangrijke besturingssystemen voor mobiele apparaten - Android, iOS en Windows - komen uit één land: de Verenigde Staten van Amerika. Wat als er met AI iets vergelijkbaars gebeurt en dat besloten wordt om de laatste ontwikkelingen in de AI-besturingssystemen niet (meteen) internationaal te delen? Bedrijven in die landen die wel toegang hebben tot de beste AI zullen concurrentievoordeel hebben. Dit kan ook op wereldschaal leiden tot welvaartverdelingsvraagstukken die met geopolitieke instabiliteit gepaard kan gaan.

6. Risicoanalyse in Onzekerheid

Hoe moeten we de (veiligheids)risico's inschatten van een opkomende technologie waarvan we de reikwijdte, de timing en het mogelijke disruptieve karakter nog niet goed kunnen doorgronden? En dan hebben we het nog niet eens over de *wisselwerking* met diverse andere grote trends waarvoor in meer of mindere mate hetzelfde geldt. Traditionele risicoanalyses, veelal gebaseerd op een technische 'maakheids'-gedachte en werkend binnen een afgebakend domein, kunnen slecht omgaan met onzekerheden en complexe verbanden. Een alternatieve manier van risicoanalyse stelt 'scenario's' en 'discussie' centraal. Vertrekpunt is de vraag wat een gewenst of vereist niveau van veiligheid is, of in meer kwantitatieve termen het acceptabel risico. Dit is geen uitkomst van een berekening maar onderwerp van debat en, uiteindelijk, een politiek-maatschappelijke keuze. In dit debat moeten de effecten die kunnen optreden en de (maatschappelijke) impact en kosten van verschillende maatregelen worden afgewogen. Nut en noodzaak van, in ons geval, AI-toepassingen moet worden afgezet tegen mogelijke effecten van de AI-revolutie op veiligheid en de impact van maatregelen om ongewenste effecten te mitigeren. Dergelijke afwegingen kunnen worden gemaakt door bedrijven voor de ontwikkeling van producten en diensten en door overheden om kaders en regels te stellen.

De impact van AI op de samenleving kan niet op voorhand nauwkeurig worden voorspeld. Wel kunnen de mogelijke effecten in kaart worden gebracht als ondersteuning voor besluitvorming en als basis om te identificeren wat de kritieke processen en effectieve oplossingsrichtingen zijn. Hiervoor bestaan verschillende technieken. Twee mogelijke technieken zijn:

1. Adaptieve planning: uitwerken van verschillende hoekpunten die samen het speelveld omvatten.
2. Probabilistische risicoanalyse: uitwerken van kansverwachtingen rekening houdend met onzekerheden.

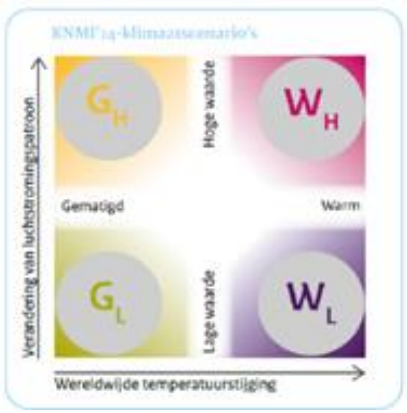
Adaptieve planning

Door het werken met verschillende scenario's (verhaallijnen) kan men het speelveld van mogelijke ontwikkelingen in kaart brengen. De scenario's schetsen bijvoorbeeld wat de effecten zijn van het mainstream worden van AI-toepassingen op een aantal momenten in de toekomst, ingekaderd door een aantal randvoorwaarden. Door verschillende aannames voor deze randvoorwaarden te hanteren en op een aantal hoekpunten in scenario's te beschrijven, worden de bandbreedtes van mogelijke effecten zichtbaar gemaakt. Er wordt geen uitspraak gedaan over de waarschijnlijkheid van ieder van de scenario's.

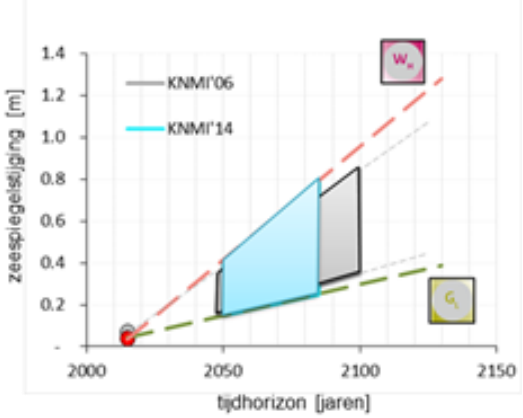
Op basis van deze scenario's kunnen de effecten in kaart worden gebracht). De bandbreedte en methodiek geven vervolgens handvatten om de potentie van diensten, producten en of beleid inzichtelijk te maken. Hierbij kan ook gebruik worden gemaakt van de 'reële opties'-aanpak om de kosten in te schatten van het zich voorbereiden op meerdere mogelijke toekomsten.[42] Per scenario kan in kaart worden gebracht wat het effect (de baten) zijn van deze maatregelen. Door de

[42] 'Reële opties' (*real options*) hebben een vergelijkbare functie als financiële opties, maar dan voor 'fysieke' investeringen. In onzekere tijden heeft het soms de voorkeur om risico's te spreiden en onomkeerbare keuzes zolang mogelijk te vermijden. Dit kan door een bredere portefeuille van verschillende opties te ontwikkelen, om vervolgens op basis van nieuw beschikbare informatie de meest geëigende optie(s) te lichten (en de overige te laten verlopen). Reële opties-theorie gebruikt de logica achter financiële opties, waarbij methoden bedacht om financiële opties (zowel 'put' als 'call') te waarderen als voorbeeld dienen om reële opties van een prijskaartje te voorzien.

scenario's onderling te vergelijken blijkt of deze maatregelen in alle gevallen werken of slechts voor een enkel scenario.



Deltascenario's gebruikt in het Deltaprogramma



Effecten van klimaateffecten uit deltascenario's van 2006 en 2014 op de zeespiegelrijzing als functie van de tijd

Figuur 5: Voorbeeld van deltascenario's voor waterveiligheid[43]

Figuur 5 visualiseert, als voorbeeld, hoe een dergelijke scenariogedreven analyse wordt gedaan voor de bescherming van Nederland tegen wateroverlast.[44] Uit de klimaatmodellen van het KNMI worden twee sleutel-variabelen gelicht: de wereldwijde temperatuurstijging en de luchtstomingspatronen. Dit zijn de twee assen in het figuur links. De klimaatmodellen geven een range aan waarbinnen deze variabelen kunnen veranderen; de vier scenario's in het figuur links – Gematigd/Laag, Gematigd/Hoog, Warm/Laag en Warm/Hoog – vormen de hoekpunten van deze range. Welk scenario werkelijkheid wordt bepaalt de zeespiegelrijzing voor de komende eeuw. Dit, op zijn beurt, is bepalend voor de mate waarin we bijvoorbeeld onze dijken moeten verzwaren. Vanuit de statistische modellen zijn alle scenario's denkbaar. Het is dan een politieke en maatschappelijke keuze waar we vanuit gaan in onze maatregelen tegen zeespiegelrijzing; waarbij duidelijk zal zijn dat minder risico lopen meer geld kost. De scenariogedreven analyse doen geen uitspraak over deze keuze, maar is wel van cruciaal belang om (1) de problematiek goed in kaart te brengen; en (2) de discussie over mogelijke maatregelen en hun kosten en baten te objectiveren.

Op vergelijkbare wijze kunnen we sleutelvariabelen definiëren voor de ontwikkelingen op het gebied van AI. Te denken valt bijvoorbeeld aan:

1. De economie: het aantal toepassingen in verschillende sectoren dat AI gebruikt en de mate waarin deze sectoren geïntegreerd zijn. Extremen: toepassing in een beperkt aantal gescheiden sectoren vs. brede en verstrengelde toepassing door de hele economie.
2. Evolutie of revolutie: de mate waarin (en evt. snelheid waarmee) AI leidt tot een vervanging van bestaande systemen of tot nieuwe toepassingen. Extremen: langzame vervanging omdat 'oude' systemen nog lang waarde blijven vertegenwoordigen vs. een kettingreactie omdat achterblijven geen optie is.

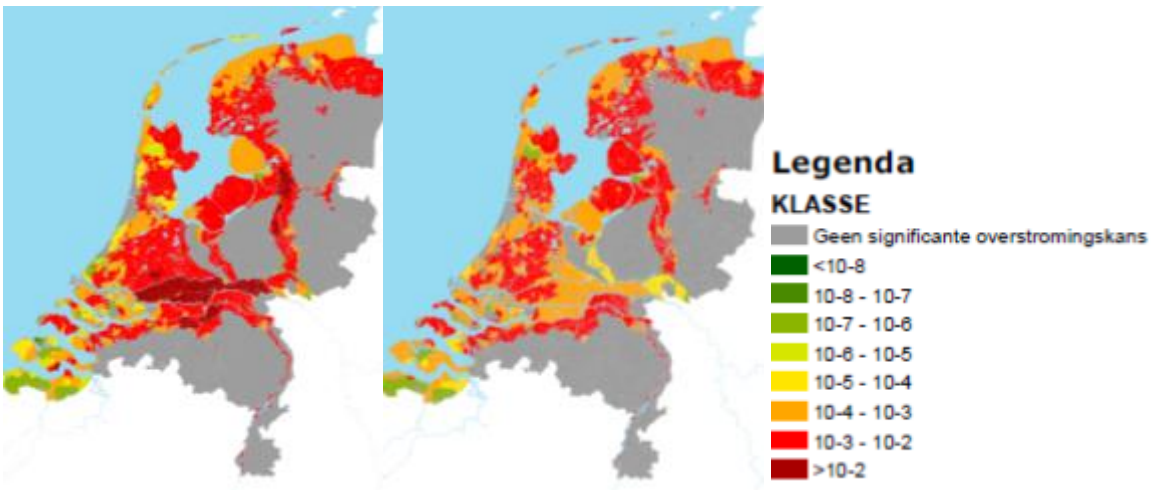
43 https://www.helpdeskwater.nl/onderwerpen/applicaties-modellen/applicaties-per/watermanagement/watermanagement/nationaal-water/kopie-werkt/uitvoer/
44 https://www.tudelft.nl/en/tpm/about-the-faculty/departments/multi-actor-systems/research/projects/adaptive-delta-management/

3. Welvaartsverdeling: de mate waarin valide politiek-maatschappelijke oplossingen gevonden worden voor een dreigende grote welvaartsongelijkheid tussen de ‘winners’ en ‘losers’ van de AI-revolutie. Extremen: grote onenigheid over aard van en oplossingen van het probleem dat per saldo nauwelijks wordt aangepakt vs. overeenstemming en implementatie van goede oplossingen (zoals een universeel basisinkomen).
4. Zingeving: De mate waarin valide politiek-maatschappelijke oplossingen gevonden worden om mensen die hun baan verliezen als gevolg van AI-revolutie zinvolle bezigheden te verschaffen. Extremen: grote onenigheid over aard van en oplossingen van het probleem dat per saldo nauwelijks wordt aangepakt vs. overeenstemming en implementatie van goede oplossingen.

Voor de combinatie van de extremen op deze assen kunnen duidelijk te onderscheiden scenario’s worden ontwikkeld die gezamenlijk het speelveld van mogelijkheden opspannen; om daarmee de discussie te voeden over robuuste maatregelen en oplossingen.

Probabilistische risicoanalyse

Een andere manier om risico-onzekerheid proberen te vangen is probabilistische risicoanalyse. Hierbij wordt getracht om de effecten van ontwikkelingen te kwantificeren rekening houdend met onzekerheden. Gebruik kan worden gemaakt van dezelfde scenario’s als bij adaptieve planning. Echter nu houden we rekening met ‘n’ (een al dan niet groot aantal) verschillende scenario’s met telkens net andere effecten. Door de risico’s in al deze scenario’s statistisch te benaderen kan ‘het’ risico worden vertaald in termen van verwachtingswaardes en bandbreedtes voor bijvoorbeeld euro’s per jaar of slachtoffers per jaar maar ook andere parameters als toegenomen werkloosheid per jaar etc. Op een vergelijkbare manier kunnen ook effecten van *maatregelen* in kaart worden gebracht.



Figuur 6: Plaatsgebonden overstromingskans voor (links) en na (rechts) versterken waterkeringen

Om de (bandbreedte van) effecten in kaart te brengen kunnen modellen gebruikt worden om de gevolgen te kwantificeren. Hierbij is het van belang om de parameters te benoemen waarin deze effecten worden beschreven. Voor de hand liggende parameters zijn kosten, werkloosheid,

vertrouwens- en welvaartsindices voor de bevolking en de impactcriteria zoals gebruikt in de Nationale Risicobeoordeling.[45]

De waarschijnlijkheid en onzekerheid kan in kaart gebracht worden met behulp van statistische en big data-modellen, en wellicht ook door gebruik te maken van gamingtechnieken. Gezien de aard van de technologie is de vraag of al deze informatie daadwerkelijk gemodelleerd kan worden, of dat hier te veel tijd (en geld) mee gemoeid is. Een alternatieve of aanvullende manier is door expertkennis te ontsluiten om zo de benodigde informatie rondom de scenario's in te schatten.

Hulpmiddelen op meerdere niveaus

Bovenstaande methoden kunnen bijdragen aan een debat of afweging over te maken keuzes over maatregelen of wat we onder een acceptabel veiligheids-/risiconiveau verstaan. De methoden schrijven deze afweging niet voor maar zijn een hulpmiddel. De methoden kunnen worden uitgewerkt op een hoog abstractieniveau, maar ook op groot detailniveau. Naarmate het detailniveau toeneemt is het, om de werklast binnen perken te houden, wel van belang om een goed beeld te hebben van de relevante én significante parameters die in meer detail worden uitgewerkt.

Al ons ervaring met risicoanalyses leert ons één ding: in de complexe werkelijkheid ontstaan altijd onverwachte risicovolle situaties die niet uit de modellen vooraf kwamen. Juist het denken in en werken met scenario's geeft de mentale flexibiliteit om hierdoor niet in een kramp te schieten. Bovendien legt deze constatering nadruk op het *proces* van risicoanalyse: niet een eenmalige exercitie, maar een continue herijking op basis van steeds nieuwe gegevens en inzichten uit de theorie en uit de praktijk.

[45] Binnen de rijksbrede strategie Nationale Veiligheid wordt jaarlijks een Nationale Risicobeoordeling (NRB) opgesteld, waarin een aantal veiligheidsthema's wordt geanalyseerd in de vorm van scenario's die langs een vaste meetlat worden gelegd: de NRB-methodiek. De resultaten hebben tot doel beleidsmakers inzicht te geven in de relatieve waarschijnlijkheid en impact van de verschillende scenario's. Dit inzicht is van belang om capaciteiten te benoemen, beleid te formuleren en prioriteiten te stellen met als doel Nederland zo goed mogelijk voor te bereiden op verschillende soorten rampen en dreigingen. Zie https://www.nctv.nl/binaries/nat.risicobeoordeling-6-definitief_tcm31-32706.pdf

7. Ter Afsluiting

In deze paper hebben we de volgende lijn van redeneren gevolgd.

De **toepassingsmogelijkheden** van AI groeien snel. We beginnen langzamerhand de impact van AI-technologieën op ons dagelijkse en professionele leven echt te merken. We zijn op het punt aangeland waar AI op bepaalde deelgebieden inmiddels gelijkwaardig is aan menselijke intelligentie of deze zelfs begint te overtreffen. Die deelgebieden worden groter en veelvuldiger, en beginnen op elkaar aan te sluiten.

De ontwikkelingen op AI-gebied, meer specifiek op het gebied van lerende systemen, lijken in een stroomversnelling te zijn geraakt. De **uitdagingen** zijn deels technisch, maar hebben minstens zoveel van doen met de filosofische, ethische, sociaal-maatschappelijke en beheer(s)-aspecten verbonden aan een steeds belangrijke rol van AI in onze samenleving. Pilotprojecten zijn nodig, niet alleen bij de (mogelijke) invoering van steeds meer en steeds omvattender AI-toepassingen, maar zeker ook voor projecten waarin nieuwe benaderingswijzen en mogelijke oplossingen voor al deze aspecten worden onderzocht.

AI raakt het **wezen van ons menszijn**, en dwingt ons te (her)overwegen wat ons uniek - of misschien toch niet - maakt. Als we, zoals Harari dat in zijn boek *Homo Deus: a brief history of tomorrow* doet, op hoog abstractieniveau naar de mensheid kijken in het licht van de AI-revolutie, dan worden alle vraagstukken en ontwikkelingen overschaduwd door drie onderling verbonden processen en de daarbij horende sleutelvragen:

1. De wetenschap convergeert naar een allesomvattend dogma, dat zegt dat organismen algoritmen zijn, en leven gegevensverwerking is. De grote vraag is: is dit echt zo?
2. Intelligentie wordt ontkoppeld van bewustzijn. Dit roept de vraag op: wat is waardevoller, intelligentie of bewustzijn?
3. Niet-bewuste maar zeer intelligente algoritmen kennen ons binnenkort beter dan we onszelf kennen. Hierin verbonden is de omvattende, maar hele praktische vraag: wat gebeurt er met de samenleving, de politiek en ons dagelijks leven als gevolg van deze constatering?

AI als *technologie* begint zich buiten duidelijk afgebakende domeinen en relatief simpele taken te bewegen. In het veiligheidsdomein geeft snelle geautomatiseerde patroonherkenning op basis van meer en meer gegevens die in onze ‘gedataficeerde’ samenleving real-time beschikbaar komen de mogelijkheid om steeds beter **strategische signalering en *early warning*** te plegen; en wellicht ook steeds meer automatisch op te reageren of zelfs pro-actie op te nemen. Ook lijkt het erop dat slimme autonoom opererende systemen in het **militaire domein** een vlucht kunnen gaan nemen. In het cyberdomein zien we al een veelheid aan ‘bots’ actief, zowel aan de criminele of orde en stabiliteit-ondermijnende kant, als aan de kant van de misdaadbestrijding en ordehandhaving. Zoals voor alle technologie geldt ook voor AI dat heilzame en schadelijke toepassingen twee zijden van dezelfde medaille vormen.

AI als *game changer* lijkt de wijze waarop onze samenleving functioneert drastisch te gaan veranderen. Ook hoog-cognitieve taken, vooral als het gaat om patroonherkenning, kunnen met AI worden uitgevoerd, net zo effectief of zelfs effectiever dan de mens dat kan. Hoe ver dit kan gaan is controversieel. De vraag of lerende systemen op basis van analyse van goede voorbeelden ook ‘kunst’ kunnen produceren - en dus creatief zijn - leidt bijvoorbeeld tot grote discussies.[46] Wel is

[46] ”Slimme computers: kunnen ze straks ook kunst maken?”, https://www.volkskrant.nl/tech/slimme-computers-kunnen-ze-straks-ook-kunst-maken~a4506035/

duidelijk dat een steeds groter aantal type banen op de rol staat om te verdwijnen ten faveure van AI-oplossingen. Een **tweedeling** dreigt tussen de relatief kleine groep 'winnaars' van de AI-revolutie en een grotere groep 'verliezers' die steeds minder kans hebben op volwaardige banen; en daarmee hun financiële vooruitzichten maar ook een belangrijke pijler onder de zin van hun bestaan zien eroderen. Een analoge tweedeling kan zich ook afspelen op mondiale schaal, waarbij landen die de AI-revolutie goed weten te benutten een grote voorsprong kunnen bewerkstelligen op landen die achterblijven. Zoals altijd in de historie nemen bij dergelijke (dreigende) **forse verschuivingen van macht** - van lokale gemeenschappen tot op mondiale schaal - de risico's van **politieke en maatschappelijke instabiliteit** toe, met alle consequenties voor veiligheid van dien.

Al met al zijn we ons nog nauwelijks goed bewust van de gevaren die grootschalige AI-toepassingen met zich mee brengen – al klinken er meer en meer stemmen die op deze omissie wijzen. Speciale aandacht is er nodig voor het opzetten van een zeker **beheersregime** om te zorgen dat AI-ontwikkelingen niet 'out of control' raken. Het eindpunt van AI-ontwikkelingen kán letterlijk levensbedreigend zijn; niet zozeer voor individuen, maar voor de mensheid als geheel. De analogie met kernwapentechnologie dringt zich op - ook daar zijn formele en informele internationale afspraken gemaakt om het 'doomsday'-karakter van de technologie zo goed mogelijk te beheersen. Ook voor AI zullen er mondiaal principes moeten worden afgesproken waaraan alle partijen zich gebonden voelen, en internationale structuren om de afspraken te controleren en zo nodig af te dwingen.

Onze analyse ondersteunt en illustreert de behoefte aan **brede(re), veelal scenariogedreven risicoanalyses** in het licht van (technologische) ontwikkelingen met een potentieel - maar vooraf niet goed in te schatten - game changing karakter. De onzekerheid over de daadwerkelijke impact van de betreffende ontwikkeling / technologie, de timing ervan en (vooral) de wisselwerking met allerlei andere ontwikkelingen die elkaar kunnen versterken of juist dempen maken dergelijke brede analyses noodzakelijk. In het verlengde is het essentieel om in de meer macrogerichte risicoanalyses van de AI-revolutie de filosofische, ethische, sociaal-maatschappelijke en beheer(s)-aspecten verbonden aan een steeds belangrijke rol van AI in onze samenleving mee te nemen.

Colofon

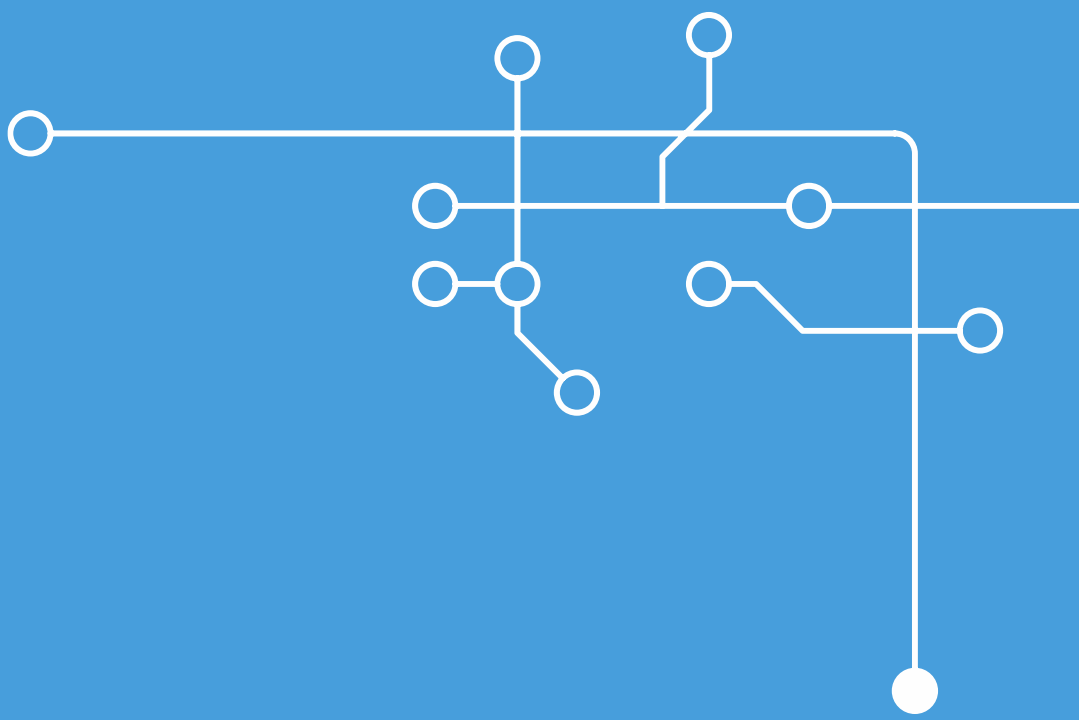
Notitie Risico-Analyse in Onzekerheid – Arificial Intelligence (Kansen en Bedreigingen)

Een publicatie in opdracht van
The Hague Security Delta
Wilhelmina van Pruisenweg 104
2595AN Den Haag
info@thehaguesecuritydelta.com
www.thehaguesecuritydelta.com
@HSD_NL

Auteur
Frank Bekkers en Karlijn Jans (HCSS).

Met dank voor de uitgebreide bijdrage van Bas Kolen en collega's van de TU Delft die de basis heeft gevormd voor het hoofdstuk *Risicoanalyse in Onzekerheid*.

Dank ook aan de inbreng van de volgende gesprekspartners: Bas Kolen van de TU Delft, Ronald Prins van Fox-IT, René Pluis van Cisco Systems en Stephan de Spiegeleire van HCSS.



The Hague Security Delta
Wilhelmina van Pruisenweg 104
2595 AN The Hague, The Netherlands
+31(0)70 204 51 80

info@thehaguesecuritydelta.com
www.thehaguesecuritydelta.com
@HSD_NL