ESSENTIELE V&G EISEN VOOR INDUSTRIËLE MACHINES UITGERUST MET MACHINE LEARNING

Datum
14 september 2018

TNO innovation for life

Rapportage voor
Ministerie van Sociale Zaken en Werkgelegenheid

ESSENTIELE V&G EISEN VOOR INDUSTRIËLE MACHINES UITGERUST MET MACHINE LEARNING

Rapport voor	Ministerie van Sociale Zaken en Werkgelegenheid
Datum	14 september 2018
Auteurs	Wouter Steijn, Liisa Janssens, Jan Harmen Kwantes, Dolf van der Beek & Anne Jansen (projectleider)
Projectnummer	060.31545
Rapportnummer	TNO 2018 R10499
Projectnaam	MAPA Robotica 18.204.2-11
Contact TNO	Anne Jansen
Telefoon	08886 60991
E-mail	Anne.jansen@tno.nl

Inhoudsopgave

Lijst met afkortingen 2

1 Inleiding 3
1.1 De machine richtlijn 5
1.2 Machine learning 6
1.3 Leeswijzer 8

2 Methode – Deel 1 9
2.1 Deskstudie 9
2.2 Interviews 9
2.3 Workshop Extern 10

3 Resultaten – Deel 1 12
3.1 Mens in controle 12
3.2 Transparantie borgen omtrent algoritmes en gedrag van de machine 13
3.3 Verantwoordelijkheden 17
3.4 Concept voorstel aanvullende essentiële V&G eisen voor machine learning 18

4 Methode - Deel 2 22

5 Resultaten – Deel 2 24
5.1 Definitief voorstel aanvullende essentiële V&G eisen voor machine learning 25
5.2 Overzicht definitief voorstel 28
5.3 Overwegingen 28

6 Referenties 31

A Appendix: Interviewprotocol 33

A.1 Protocol 1. Praktijkdeskundige 33

A.2 Protocol 2. Wetenschappelijke expert 34

Lijst met afkortingen

AI	– Artificiële intelligentie
CE	– Conformité Européenne
EFTA	– European Free Trade Association
EU	– Europese Unie
FIM	– Futuristische industriële machines
GDPR	– General Data Protection Regulation
GIM	– Geavanceerde industriële machines
HCI	– Human-computer interaction
HLMI	– High-level Machine Intelligence
ISO	– International Organization for Standardization
MiFID	– Markets in Financial Instruments Directive
MKB	– Midden- en kleinbedrijf
MR	- Machinerichtlijn
MRA	– Mutual Recognition agreement
OSHA	– Occupational Safety & Health Administration
R&D	– Research and Development
RL	– Reinforcement learning
SZW	– Ministerie Sociale Zaken en Werkgelegenheid
SML	– Supervised machine learning
TIM	– Traditionele industriële machines
UML	– Unsupervised machine learning
V&G	– Veiligheid en Gezondheid

1 Inleiding

De robotindustrie is bezig met een enorme snelheid een opmars te maken in diverse arbeidsdomeinen van zorg tot maakindustrie. In de afgelopen vijftig jaar zijn industriële machines niet alleen sneller en accurater geworden, maar zijn deze machines ook mobieler geworden. Daarnaast hebben deze machines meer bewegingsgraden gekregen, en kunnen deze machines dankzij verbeterde sensoren steeds beter met hun omgeving omgaan. Verder zien we een toenemende digitalisering (waarbinnen de ontwikkelingen parallel lopen aan die van automatisering; Eurofound, 2018), dit houdt in dat (industriële) machines steeds vaker via netwerken aan elkaar en aan het internet worden gekoppeld. Het is belangrijk om bij deze trends vooruit te blijven kijken en de machineveiligheid van morgen te definiëren, zodat een robot al in het ontwerp en de ontwikkeling ervan inherent veilig kan worden gemaakt.

In dit kader heeft TNO, in opdracht van en in samenwerking met het ministerie Sociale Zaken en Werkgelegenheid (hierna SZW), in de afgelopen jaren onderzoek gedaan naar de mate waarin de Europese richtlijn 2006/42/EG machines (hierna: de Machinerichtlijn) toekomstbestendig is. In eerdere rapporten heeft TNO onderzocht wat nieuwe risico's en mogelijke beheersmaatregelen zijn, van het koppelen van industriële machines met en via het internet (Steijn, van der Vorm et al., 2016), de introductie van robots op de werkvloer (Steijn, Luiijf et al., 2016), en de gedeelde werkvloer tussen mens en robot (Jansen et al., 2018). In dit rapport zal de focus liggen op de volgende onderzoeksvraag, geformuleerd door SZW:

Welke essentiële veiligheids- en gezondheidseisen dienen in de Machinerichtlijn te worden opgenomen voor machines met machine learning?

Ontwikkelingen met betrekking tot industriële machines. Industriële machines worden aangestuurd door besturingssystemen. Bij de traditionele industriële machine (TIM) zijn de handeling 'scripted'. Dat wil zeggen, dat de TIM binnen een gestructureerde omgeving zijn voorgeschreven (geprogrammeerde) taak uitvoert (zie Figuur 1).

Figuur 1[1] Voorbeelden van traditionele industriële machines (TIMs) die een vaste taak hebben. Vaak in een afgesloten ruimte en zonder machine learning

[1] Figuren van https://techcrunch.com/2017/03/19/y-combinator-has-a-new-ai-track-and-wants-startups-building-robot-factory-tech-to-apply/?guccounter=1 en http://dutch.foodmakingmachines.com/sale-4110180-industrial-automatic-dough-forming-machine-steamd-bun-making-machine.html

Ontwikkelingen met betrekking tot industriële machines (vervolg)

Recentelijk doen meer geavanceerde industriële machines (GIM) hun intrede op de werkvloer. Deze GIMs worden ook wel robots of cobots genoemd, maar werken nog steeds aan een afgebakende taak binnen een gestructureerde omgeving, en kunnen binnen deze context nieuw acties leren (zie Figuur 2). Door middel van ontwikkelingen op het gebied van machine learning kunnen GIMs hun acties verbeteren op basis van nieuwe data, waardoor expliciete programmering van het gehele spectrum aan wenselijke acties niet meer noodzakelijk is. De focus van de ontwikkelingen bij de GIMs is om deze minder afhankelijk van menselijke operators te laten functioneren. Technische innovaties richten zich daarbij op capaciteiten die beter in staat zijn om hun omgeving te herkennen en erop te reageren (Eurofound, 2017).

Figuur 2[2] Voorbeelden van geavanceerde industriële machines (GIMs) die meerdere taken kunnen uitvoeren, bij de mens op de werkvloer zijn, en over machine learning technologie beschikken.

Naar de toekomst toe zullen industriële machines (ofwel FIMs, futuristische industriële machines) nog verder verbeterd kunnen gaan worden op het gebied van real-time verwerking van informatie, probleemoplossing, mobiliteit, sensoriek, en leer- en aanpassingsvermogen. In een recent onderzoek (Grace et al., 2018) geven AI experts aan dat er 50% kans is dat binnen 45 jaar er een AI is die op alle fronten beter is dan de mens (ofwel High-level Machine Intelligence; HLMI). Deze cijfers zijn gebaseerd op schattingen van 352 wetenschappers die binnen het vakgebied van machine learning hebben gepubliceerd. Op deelaspecten worden er al eerder ontwikkelingen verwacht. GIMs of FIMs die in ongestructureerde omgevingen (bv. Bouwplaatsen) kunnen functioneren worden over 15 jaar verwacht (Robotics VO, 2013).

De Machinerichtlijn is in Nederland geïmplementeerd via het Warenwetbesluit machines. Het Warenwetbesluit machines verwijst naar bijlage 1 van de Machinerichtlijn. Hierin zijn essentiële veiligheids- en gezondheidseisen (V&G eisen) opgenomen. Deze eisen moeten door fabrikanten en ontwerpers in acht worden genomen bij het in de handel brengen van machines.

De Machinerichtlijn is opgesteld in een tijd waarin er nog geen praktijktoepassingen van machines met machine learning bestonden. Samuel was de eerste die met de term machine learning kwam in 1959. Hij onderzocht de mogelijkheid om machines te laten leren van ervaringen zodat alles niet expliciet geprogrammeerd hoeft te worden (Samuel, 1959). Machine learning is een onderdeel van Artificiële Intelligentie (AI) en maakt onderdeel uit van het besturingssysteem van een machine. Alhoewel de Machinerichtlijn wel V&G eisen bevat

[2] Figuren van https://www.nbt.nhs.uk/about-us/building-brunel/automated-guided-vehicle-system en https://www.talentica.cz/robot-nebo-kobot/

waar besturingssystemen aan moeten voldoen, is het de vraag of deze eisen nog voldoende zijn voor het veilige ontwerp van de machine, zodra de machine een lerende entiteit wordt waardoor het zijn gedrag kan aanpassen.

De V&G eisen in de machinerichtlijn zijn van toepassing op fabrikanten en ontwikkelaars van machines voor zowel de professionele context als de consument. In dit rapport zal voornamelijk gekeken worden naar industriële machines die gebruik maken van machine learning algoritmes tijdens de gebruiksfase. Dit in tegenstelling tot machines die alleen tijdens de ontwerpfase worden getraind met behulp van machine learning algoritmes. De verwachting is dat juist de mogelijkheid tot aanpassen van acties door middel van machine learning op de werkvloer tot nieuwe risico's kan leiden. De voorgestelde V&G eisen aan het eind van het rapport zullen dus voornamelijk van toepassing zijn op de eerstgenoemde categorie machines.

Om bovengenoemde onderzoeksvraag te beantwoorden is een deskstudie gedaan naar het onderwerp en zijn verschillende experts op het gebied van robotica en machine learning benaderd voor interviews. Op basis van de verzamelde informatie zijn vervolgens enkele onderwerpen vastgesteld die van belang zijn in relatie tot het opstellen van aanvullende V&G eisen. Op basis van deze informatie is een eerste voorstel gedaan van mogelijke aanvullende V&G eisen. Deze eisen zijn vervolgens in een vervolgtraject van twee interne workshops bij TNO getoetst, alvorens ook een definitieve versie van deze eisen is vastgelegd. Dit rapport is daarom in twee delen opgebouwd. Het eerste deel gaat over de initiële verkenning. In deel 2 wordt de validatie verder beschreven. Hieronder zal eerst een korte toelichting worden gegeven op de Machinerichtlijn en machine learning, alvorens de aanpak van het project wordt beschreven.

1.1 De machine richtlijn

Binnen de Europese Unie (EU) zijn vanaf de jaren tachtig van de vorige eeuw veel productrichtlijnen[3] verschenen. Dit type richtlijnen wordt ook wél aangeduid als productveiligheids-, nieuwe-aanpak of CE-markerings-richtlijnen. In Nederland is de Machinerichtlijn opgenomen in de Warenwetgeving; meer in het bijzonder het Warenwetbesluit machines.

Deze richtlijnen hebben de volgende doelstellingen:

1 Realiseren van een Europese interne markt voor producten door het harmoniseren van de gestelde eisen aan die producten.
2 Het bieden van een hoog niveau aan veiligheids- en gezondheidsbescherming voor de mensen die met die producten werken/omgaan en in voorkomende gevallen ook voor dieren en de omgeving.
3 Een gelijk speelveld voor conformiteitsbeoordelingsinstanties, waaronder de zogenaamde notified bodies.

[3] De reikwijdte van deze productrichtlijnen is de gehele Europese interne markt. Deze omvat de lidstaten van de EU, de lidstaten van EFTA en via een mutual recognition agreement (MRA) ook Zwitserland.

De veiligheids- en gezondheidseisen die gesteld worden aan machines zijn van belang voor fabrikanten van machines binnen de Europese interne markt. Dit is opgenomen in artikel 2 en artikel 5 van de richtlijn 2006/42/EG machines (ofwel Machinerichtlijn). Een fabrikant of diens gemachtigde die een machine in de handel wil brengen moet voor deze in gebruik mag worden genomen voldoen aan de gestelde eisen in de Machinerichtlijn. Hoe hij aan die eisen gaat voldoen is aan de fabrikant of diens gemachtigde. De fabrikant dient aan de hand van de V&G eisen uit bijlage 1 van de Machinerichtlijn het volgende te doen:

1 Inventariseren van de gevaren die de machine veroorzaakt.
2 Deze gevaren wegnemen.
3 Als dit niet mogelijk is, maatregelen nemen om de gevaren te beheersen.
4 De restgevaren te vermelden in de documenten die hij bij de machine levert.

Hiervoor is het noodzakelijk om een risicobeoordeling van de machine te maken om te beoordelen welke veiligheids- en gezondheidseisen uit Bijlage I van toepassing zijn. Vervolgens past de fabrikant deze eisen toe in de bouw van de machine.
Om te voldoen aan de V&G eisen kan de fabrikant als hulpmiddel gebruik maken van zogenaamde Europese geharmoniseerde normen. Dit zijn normen die een concretere uitwerking bevatten van de algemeen geformuleerde veiligheids- en gezondheidseisen.
De fabrikant geeft met CE-markering en de verklaring van overeenstemming bij de machine aan dat hij de machine heeft gebouwd conform de essentiële veiligheids- en gezondheidseisen van de Machinerichtlijn.

In bijlage I, artikel 1.2.1 van de Machinerichtlijn staan V&G eisen met betrekking tot besturingssystemen. Deze V&G eisen zullen ook van toepassing zijn voor machines met machine learning.

Het besturingssysteem:

1. Is bestand tegen de normale bedrijfsbelasting en tegen invloeden van buitenaf.
2. Storing mag niet tot een gevaarlijke situatie leiden.
3. Fouten in de besturingslogica mogen niet tot een gevaarlijke situatie leiden.
4. Menselijke fouten gedurende de werking mogen niet tot een gevaarlijke situatie leiden.

1.2 Machine learning

Machine learning is een complex onderwerp dat nog volop in beweging is. Wij zullen hier niet te diep op machine learning ingaan, maar willen het onderwerp wel kort introduceren in verband met de doelstelling: het formuleren van mogelijke essentiële V&G eisen voor machines met machine learning.

Machine learning is een onderdeel van het onderzoeksgebied van AI. Binnen het onderzoeksgebied van AI wordt geprobeerd om via computers intelligent gedrag te repliceren. Machine learning zorgt ervoor dat machines patronen kunnen herkennen in complexe data, en kunnen leren van ervaring (ofwel data) om zo hun prestaties of taakuitvoering te optimaliseren. Denk daarbij bijvoorbeeld aan het optimaliseren van de tijd waarin een taak kan worden uitgevoerd of het minimaliseren van foute beslissingen. Dankzij betere informatieverwerking, geheugencapaciteit, en rekenkracht (in vergelijking tot de mens), kunnen machines met machine learning ingezet worden voor taken a) die te moeilijk

zijn voor de mens om uit te voeren, b) die te complex zijn om te programmeren[4] en/of c) waarbij flexibiliteit nodig is[5] (Shalev-Shwartz & Ben-David, 2014).

Er bestaan verschillende vormen en classificeringen van machine learning. We bespreken hier ter illustratie; 'Reinforcement learning' (RL), 'Supervised machine learning' (SML) en 'Unsupervised machine learning' (UML)[6].

Bij reinforcement learning leert de machine op basis van positieve en negatieve feedback de juiste associaties tussen input en output te leggen. Hierdoor leert de machine het gewenste gedrag te optimaliseren.

Bij SML wordt de machine voorbeelden aangeboden (input) met bijbehorende gewenste acties (output) (Russell, & Norvig, 2010). Hiermee leert de machine wat 'goed' en wat 'fout' is. Het doel is om een algemene regel te genereren welke output de machine moet genereren op basis van bepaalde input. Dit wordt 'supervised leren' genoemd omdat de mens direct betrokken is bij het leerproces en het geleerde proces kan toetsen.

Bij UML kan een machine zich 'zelfstandig' aanpassen aan nieuwe situaties. Op basis van input vindt de machine zelf patronen of een structuur, die het vervolgens kan toepassen op nieuwe input. Iedere situatie levert nieuwe input op die de machine gebruikt om patronen te ontdekken die het kan gebruiken. Op deze manier koppelt de machine zelf bepaalde output aan bepaalde input zonder dat dit door een mens aangegeven wordt.

Het is belangrijk om op te merken dat in de huidige stand van zaken, machines met machine learning alleen nog kunnen leren afhankelijk van hun programmering en algoritmes. Er is dus nog geen sprake van een echt zelfstandig lerende entiteit, maar van een machine die dankzij een complexer besturingssysteem een breder scala aan input kan verwerken en zijn output op basis hiervan kan optimaliseren. Een voorbeeld hiervan is dat machines vooralsnog altijd doelfuncties meekrijgen (waarmee de machine zijn doel berekent).

Voorbeeld: AlphaGo
AlphaGo (Silver et al., 2017) wist in principe oorspronkelijk niks van het spel Go. De regels zijn tijdens een trainingsfase via SML aan AlphaGo geleerd door feedback te geven tijdens het spelen. Tijdens de daadwerkelijke toepassing verbeterde het systeem zich vervolgens elke keer na winst of verlies door middel van reinforcement learning. De laatste versie van AlphaGo heeft hierdoor bovenmenselijke prestaties neergezet door met 100-0 te winnen tegen de menselijke kampioen Go.

[4] Hier vallen taken onder die mensen 'natuurlijk' uitvoeren, zoals spraakherkenning, maar ook taken die te moeilijk voor de mens zijn zoals weervoorspellingen.
[5] Traditioneel kunnen machines niet van hun 'script' afwijken, met machine learning kunnen zij echter reageren op veranderende situaties in hun omgeving.
[6] Zie ook: https://www.e-sites.nl/blog/476-machine-learning-een-korte-toelichting-op-de-techniek-en-toepassing.html

1.3 Leeswijzer

In het volgende hoofdstuk wordt de methode beschreven waarmee bestaande literatuur, praktijkdeskundige, en wetenschappelijke experts zijn geraadpleegd om een inzicht te krijgen van de nieuwe risico's die meekomen met machines met machine learning. In hoofdstuk 3 worden vervolgens de resultaten beknopt weergegeven om uiteindelijk tot een concept voorstel voor aanvullende essentiële V&G eisen te komen met betrekking tot machines met machine learning. Dit concept voorstel is vervolgens intern bij verschillende TNO experts uit relevante vakgebieden getoetst. De gevolgde methode van deze toetsing staat beschreven in hoofdstuk 4. In hoofdstuk 5 zijn de belangrijkste discussiepunten weergegeven die tijdens deze toetsing naar voren zijn gekomen. Op basis van deze discussiepunten is een definitief voorstel voor essentiële V&G eisen voor machines met machine learning vastgelegd.

2 Methode – Deel 1

De resultaten in dit deel van het rapport zijn verkregen door middel van het bestuderen van beschikbare documentatie waaronder de Machinerichtlijn, interviews en een workshop met praktijkdeskundigen en wetenschappelijke experts op het gebied van robotica. Hieronder wordt elk van de gehanteerde methoden kort toegelicht. In Hoofdstuk 3 zijn de belangrijkste punten die tijdens de deskstudie naar voren zijn gekomen geïntegreerd met de resultaten uit de interviews en de workshop.

2.1 Deskstudie

Het onderwerp van dit rapport is dusdanig complex en vernieuwend dat het noodzakelijk is om in de (grijze) literatuur op zoek te gaan naar aanknopingspunten voor essentiële veiligheids- en gezondheidseisen. Verder voeren we een analyse uit van de Machinerichtlijn in zijn huidige vorm.

2.2 Interviews

Er zijn semigestructureerde interviews gehouden met twaalf experts op het gebied van robotveiligheid, mens-machine interactie, cognitieve engineering, artificiële intelligentie, ethiek en wetgeving, robotontwikkeling, -implementatie en -gebruik. Tijdens de interviews werd de discussie over machines met machine learning in de vorm van robots besproken. Daarom zal bij de bespreking van de resultaten ook het woord robots worden gehanteerd.

Tabel 1 Achtergrond geïnterviewden

Functie	Specialisatie
Praktijkdeskundige	Technologie, arbeid en privacy op de werkplek
Praktijkdeskundige	R&D Robotics
Praktijkdeskundige	Machinerichtlijn deskundige
Wetenschappelijke expert	Interfacing Law & Technology
Wetenschappelijke expert	Intelligent Control and Robotics
Wetenschappelijke expert	AI deskundige
Wetenschappelijke expert	Specifieke focus op robots en AI
Wetenschappelijke expert	AI deskundige
Wetenschappelijke expert	Integrated Systems Engineering
Wetenschappelijke expert	Intelligent man-machine systems
Wetenschappelijke expert	Robot en AI ethiek
Wetenschappelijke expert	AI en HCI deskundige

2.2.1 Deelnemers interviews

Op basis van de literatuur- en internetscan is een actoranalyse gemaakt waarbij met name actoren zijn geselecteerd, die kennis hebben van robotsystemen in het algemeen en specifieke kennis bezitten ten aanzien van AI. Deze experts zijn vervolgens gebeld of per mail uitgenodigd. Het doel was om tot minimaal tien deelnemers te komen.

In totaal zijn er 32 uitnodigingen in maart verstuurd. In Tabel 1 is in een overzicht een anonieme beschrijving van de uiteindelijk geïnterviewde deelnemers opgenomen.

2.2.2 Interviewprotocol

De interviews waren semigestructureerd, dit houdt in dat er vooraf een protocol is opgesteld met vragen die dienden als handvat voor het interview. Tijdens de interviews is voornamelijk en doelbewust doorgevraagd naar datgene waar de geïnterviewde persoon veel over kon vertellen. De interviews duurden ieder een uur tot 1,5 uur. In Appendix A staat het gehanteerde interviewprotocol. Vragen werden indien nodig aangepast aan de achtergrond van de geïnterviewde.

Naast het protocol hebben geïnterviewden vooraf informatie ontvangen om het gesprek te stroomlijnen. Deze informatie was gebaseerd op bestaande literatuur en eigen kennis binnen TNO. Samengevat betrof deze informatie een overzicht van drie categorieën waarin onderscheid werd gemaakt tussen industriële machines[7]:

Categorie 1: Traditionele industriële machine (TIM)
Deze machines zijn zodanig geprogrammeerd zodat zij een simpele taak geautomatiseerd kunnen uitvoeren. Hierbij zijn deze machines vaak locatie gebonden en is er een fysieke afstand tussen de machine en de mens (bijv. door een veiligheidskooi). Deze machines kunnen niet opereren zonder directe inmenging van de mens (bijv. door middel van directe bediening, ofwel door het aanleveren en verwijderen van te bewerken materialen.

Categorie 2: Geavanceerde Industriële machine (GIM)
Deze machines zijn in staat om meerdere of complexere handelingen uit te voeren. Hierbij kunnen sommige types dankzij machine learning ook hun originele programmering ontstijgen. Dat wil zeggen dat zij op basis van data hun taak anders en efficiënter kunnen gaan uit voeren. Deze machines zijn niet meer per se locatie gebonden en sommige kunnen hun omgeving ‘zien’ via sensoren en daarop reageren. Zij opereren naast de mens op de werkvloer, en er is meer sprake van samenwerking dan bediening, al volgt deze samenwerking wel een ‘script’.

Categorie 3: Futuristische industriële machine (FIM)
Deze machines hebben een AI die menselijke intelligentie benaderd. Zij zijn creatief, en kunnen problemen zelf oplossen. In plaats van machines die bediend worden, worden zij ‘agents’ op zich en is er sprake van een robot collega.

2.3 Workshop Extern

Op 4 mei 2018 is bij het platform voor Industrie & Veiligheid van de NEN een workshop “*Cobots als opkomend risico voor arbeidsveiligheid”*[8] gehouden van 1,5 uur. Aan de workshop namen 30 deelnemers en twee TNO-projectleden deel.

[7] De geïnterviewden kregen vooraf aan het interview onderstaande informatie in tabelvorm.
[8] https://www.nen.nl/Evenementen/Evenementdetailpagina/NENPlatform-Industrie-Veiligheid- 2.htm

Het doel van de workshop was om antwoord te geven op de volgende vragen:

1. Wat is er voor nodig om de samenwerking (in de toekomst) tussen werknemers en machine (robot) inherent veilig te maken op de werkvloer.
2. Op welke reeks menselijke waarden (bijv. de privacy van de mens) moet AI dan zijn afgestemd (in relatie tot het nemen van beslissingen in de uitvoering van het werk samen met een menselijke collega) en welke juridische en ethische status zou dit moeten hebben.
3. Hoe kunnen we het huidige wettelijk kader (o.a. Machinerichtlijn) aanpassen om de risico's te beheersen die aan AI zijn verbonden.
4. Wat is nodig om een constructieve en gezonde (kennis)uitwisseling mogelijk te maken tussen alle partijen in de keten (fabrikanten, systeem integratoren, eindgebruikers, AI-onderzoekers en beleidsmakers) en zo te streven naar een inherent veilige samenwerking tussen mens en robot?

Voor de workshop werden de deelnemers in vijf groepen van zes mensen verdeeld die vervolgens brainstormden over bovenstaande vragen. Het resultaat van de discussie is per groepje op een flipovervel gezet en in een plenaire terugkoppeling is per tafel door één iemand uit het groepje de uitkomsten van de discussie gepresenteerd. Vervolgens konden de andere deelnemers hierop weer reflecteren. De vragen 1 en 3 hadden direct betrekking op de onderzoeksvraag in dit rapport. De discussie omtrent deze vragen zijn verwerkt in de resultaten (deel 1).

3 Resultaten – Deel 1

Op basis van de informatie die is verzameld met de methoden zoals beschreven in de vorige sectie, kwamen drie onderwerpen naar voren die van belang zijn voor het opstellen van essentiële V&G eisen en de borging hiervan. Dit was ten eerste dat een machine niet bovenaan de beslissingshiërarchie mag staan. Met andere woorden, de mens moet altijd controle hebben over de machine en niet andersom. Ten tweede, werd er veel belang gehecht aan het transparant maken van de algoritmes van de machines en de gebruikte data. Dit geld zowel gedurende het gebruik als tijdens het ontwikkelproces. Deze transparantie is noodzakelijk om de kwaliteit te borgen, en om te voorkomen dat er 'bias' ontstaat in de machines, of dat de machines een 'black box' worden. Met de toenemende complexiteit van de besturingssystemen en het feit dat machines zelf onvoorspelbaar kunnen gaan bewegen op basis van wat ze leren, is transparantie van belang om te voorkomen dat machines onvoorspelbare 'black boxes' worden. Tot slot, zal voor de borging van de V&G eisen van belang zijn dat er duidelijkheid is over hoe verantwoordelijkheden verdeeld zijn over bijvoorbeeld ontwikkelaars, integrators, en de uiteindelijke gebruiker.

In dit hoofdstuk zullen bovenstaande onderwerpen verder toegelicht worden met behulp van de verzamelde informatie. Vervolgens wordt op basis van deze informatie een eerste versie van mogelijke aanvullende V&G eisen voor machines met machine learning voorgesteld. In de volgende hoofdstukken worden de stappen beschreven die zijn gezet om van een concept tot een definitief voorstel te komen van de V&G eisen.

Tijdens de interviews werd de discussie over machines met machine learning in de vorm van robots besproken. Daarom zal bij de bespreking van de resultaten ook het woord robots worden gehanteerd.

3.1 Mens in controle

Diverse van de geïnterviewde experts lijken het erover eens te zijn dat de robot niet bovenaan de beslissingshiërarchie mag staan. Een robot moet vooral gezien worden als een machine die op basis van zijn eigenschappen beslissingsondersteunend kan functioneren. Mensen moeten kunnen kiezen hoe en of ze beslissingen moeten of willen delegeren aan machines om door de mens gekozen doelen te bereiken.

Verder mag een robot ook niet ingezet worden op een manier waarop de robot zelf beslissingen gaat nemen die van invloed zijn op de mensen in zijn omgeving. Dit is bijvoorbeeld het geval bij zogenaamde 'trollyproblemen'. Trollyproblemen gaan over moreel moeilijke afwegingen in dilemma's over mensenlevens, waarin elke keuze een dodelijke afloop kent (bijvoorbeeld de keuze tussen een bekende redden, versus het redden van vijf vreemdelingen). Het is niet ondenkbaar dat autonome machines in de praktijk voor dit soort dilemma's komen te staan. Dit is nu al het geval voor bijvoorbeeld autonome auto's[9].

[9] Algemeen Dagblad (2018), Essay, de mens en zijn machine, we accepteren al lang dat robots mensen doden, 25 maart 2018.

Idealiter valt de verantwoordelijkheid van de keuze altijd tot de mensen te herleiden (bijv. op de wijze waarop een algoritme is geprogrammeerd).

Het Europees Economisch en Sociaal Comité heeft daarvoor in haar initiatiefwet artificiële intelligentie (EESC, 2017) eveneens gepleit voor een **human-in-command** benadering van AI met als randvoorwaarden de verantwoordelijke, veilige en nuttige ontwikkeling van AI, waarbij machines, machines blijven en mensen altijd betekenisvolle controle over deze machines moeten en zullen behouden. Een basisvoorwaarde voor betekenisvolle controle door de mens is dat de machine transparant is naar de gebruiker op de werkvloer die met de machine samenwerkt, transparant is zodat toezicht mogelijk is op het proces en transparant voor eventuele controles achteraf (bijv. bij een incidentanalyse). In de volgende sectie gaan we verder op deze vormen van transparantie in.

3.2 Transparantie borgen omtrent algoritmes en gedrag van de machine

Machines met machine learning introduceren nieuwe risico's op de werkvloer omdat de transparantie van wat en waarom ze iets doen zal afnemen. Het besturingssysteem van de robot dreigt zoals eerder beschreven een 'black box' te worden. Hierdoor worden de handelingen van de robot minder voorspelbaar voor de gebruiker die met de robot (samen)werkt. Tevens wordt het hierdoor achteraf moeilijker te bepalen waarom een robot een bepaalde handeling heeft uitgevoerd. Hieronder bespreken we de volgende drie typen transparantie die geborgd moeten worden in wet- en regelgeving om tot veilige robots te komen:

- Transparantie in gebruikte data sets en algoritmes (om bias te voorkomen);
- Transparantie in algoritmes van de machine (om de 'black box' te voorkomen);
- Transparantie in het ontwikkelproces (om kwaliteit te garanderen).

Valkuilen het nastreven van transparantie

Weller (2017) identificeert valkuilen in het nastreven van transparantie die voorkomen moeten worden. Ten eerste kan transparantie misleidend zijn als er ook nog informatie wordt achtergehouden. Ten tweede kan het vergroten van transparantie ertoe leiden dat privacy van individuen wordt aangetast. Ten derde moet voorkomen worden dat transparantie als doel op zich wordt gezien waardoor innovatie wordt afgeremd. Het vergroten van de transparantie moet altijd als achterliggend doel hebben om de veiligheid te verbeteren. Tot slot kan grotere transparantie leiden tot discriminerend gedrag (d.m.v. het ter beschikking stellen van betrouwbare informatie over bepaalde persoonskenmerken zoals etniciteit).

3.2.1 Transparantie in gebruikte data sets en algoritmes om bias te voorkomen

Een robot moet altijd 'explainable' zijn. Met andere woorden, diens gedrag en beslissingen moeten kunnen worden herleid naar de geprogrammeerde algoritmes. Het is dus van belang dat de algoritmes en de data, op basis waarvan een robot functioneert transparant en begrijpelijk zijn. Op deze manier beheers je dat niet meer te herleiden valt waar(foute) handelingen van de robot door ontstaan.

Een van de experts noemt het voorbeeld van Microsoft onderzoek waarin een autonoom systeem is ontwikkeld op basis van machine learning algoritmes. Dit systeem kon worden ingezet om beslissingen van artsen -over het kwalificeren van een patiënt in een bepaalde

risicocategorie (hoog of laag)- te ondersteunen. Pijn op de borst en astma klachten werd als een laag risico bestempeld, deze classificering bleek echter door bias ontstaan te zijn. Deze groep patiënten werd namelijk vaak vroegtijdig opgenomen en had daardoor een laag overlijdensrisico. Om als mens in te kunnen grijpen (om 'potentiële bias' of 'trade offs' te voorkomen) dient er inzicht te zijn in het algoritme, welke datasets er zijn gebruikt, wat de hypothese was en hoe de verschillende performance metrics presteren waarop een diagnose is gebaseerd.

Vertaald naar de arbeidssituatie kan een verkeerde toetsing van het algoritme waarmee de robot functioneert er bijvoorbeeld toe leiden dat een lasrobot gaat lassen terwijl gebruikers nog geen oogbescherming hebben opgedaan. Mogelijk had de robot foutief geleerd dat lassen een laag risico handeling was omdat er nooit incidenten plaatsvonden (dit omdat gebruikers tot dan toe altijd oogbescherming op hadden terwijl de robot aan het lassen was).

Een andere expert geeft als voorbeeld een Chinees onderzoek, waarbij een autonoom systeem besliste wie een dader is of niet. Voor dit onderzoek werden foto's gebruikt van mensen van LinkedIn en werkelijke gedetineerden. Dit leverde een bias op doordat mensen op LinkedIn lachen voor hun carrière en mensen die de gevangenis in gaan nu eenmaal niet vrolijk (eerder boos) kijken. Als dit bijvoorbeeld een van de selectiefeatures worden om de beslissing op te maken is het afwegingskader onjuist en niet betrouwbaar voor daderselecties.

Een andere wetenschappelijke expert geeft aan dat een lerende machine voor het nemen van real-time beslissingen afhankelijk is van tijd, zijn perceptie en interpretatie van zijn omgeving. Wanneer een autonome drone iets moet vastleggen in zijn omgeving is het ene landschap bijvoorbeeld niet hetzelfde als het andere landschap. Kortom, de factoren in de omgeving kunnen bias veroorzaken en daarmee de beredenering die voorafgaat aan een beslissing beïnvloeden. De omgeving waarin een systeem is getest is vaak niet hetzelfde als een omgeving buiten de onderzoeksetting, wat risico's oplevert als systemen na labtesten zonder extern explorerend onderzoek in de industriële setting worden geplaatst. In je test omgeving kun je namelijk vaak beperkt rekening houden met de gehele context waarin de uiteindelijk praktijktoepassing gaat opereren.

Ook tijdens het programmeren van een robot kan bias optreden, bijvoorbeeld doordat de programmeur (onbewust) zijn eigen perceptie en eigen ervaringen toepast op het algoritme. Verder is het van belang om te voorkomen dat er bias in de robot ontstaat. Dit kan door goede databronnen als basis te gebruiken, maar ook door te voorkomen dat er externe invloed op de databronnen kan worden uitgeoefend. Daartoe heeft bijvoorbeeld het Europees Economisch en Sociaal Comité middels een initiatiefwet Artificiële Intelligentie (EESC, 2017) gepleit voor een Europese AI-infrastructuur, bestaande uit opensource en privacy respecterende leeromgevingen, levensechte testomgevingen en datasets van hoge kwaliteit voor ontwikkeling en training van AI-systemen.

Bias kan ook optreden door externe invloeden (bijv. cybercriminelen of hackers) of softwarefouten. Door geïnterviewden worden enkele voorbeelden gegeven hoe hiermee om kan worden gegaan. Bijvoorbeeld door gebruik te maken van ondersteunende software, waarmee getoetst wordt of implementatie van een algoritme correct gebeurt ('correct by design'). Verder kan gedacht worden aan het programmeren van een extra laag in de

software, die het eigen gedrag controleert en reguleert. Hierbij kan gedacht worden aan een beslissingskader, (waarin de wettelijke V&G eisen van de MD zijn voorzien). Een *'veiligheidslus'* in de software kan de robot uitschakelen als het geplande gedrag niet aan de veiligheidscriteria voldoet.

3.2.2 Transparantie in algoritmes tijdens gebruik om een 'black box' te voorkomen

Zoals eerder genoemd dreigen robots een 'black box' (in negatieve zin) te worden als gevolg van niet transparante datasets en algoritmes. Het zijn gesloten systemen die een input ontvangen, een output produceren zonder uitleg over hoe en waarom ze tot die output zijn gekomen. Hier zitten meerdere nadelen aan verbonden:

- Er kan ongemerkt bias optreden in de robot (zie ook voorgaande sectie 3.2.1).
- Het is niet te herleiden of de robot beslissingen neemt op basis van geldige criteria (onjuist of op een niet wetenschappelijke manier verkregen resultaten zijn niet geldig).
- Fouten die optreden in het systeem zijn moeilijker tot de bron te herleiden en daarmee lastiger te corrigeren.
- De robot kan van binnenuit of van buiten het systeem gehackt worden en daarmee de integriteit van de software bedreigen.

'Algoritmische transparantie' is het onderwerp dat dit jaar geadresseerd zal worden door de Europese Commissie die een drieledige aanpak in relatie tot AI heeft voorgesteld (2018). Het gaat hierbij niet over openbaarmaking van de broncode van een algoritme als zodanig. Het kan verschillende vormen aannemen, afhankelijk van de situatie, inclusief zinvolle uitleg (zoals vereist in de General Data Protection Regulation (GDPR)[10] ten aanzien van bijvoorbeeld geautomatiseerde beslissingen op basis van persoonlijke gegevens), of rapportage aan de bevoegde autoriteiten (zoals vereist in Markets in Financial Instruments Directive (MiFID II)[11]).

Verder heeft het Europees Economisch en Sociaal Comité met een initiatiefwet 'Artificiële intelligentie' (EESC, 2017) gepleit voor o.a. de ontwikkeling van een normeringssysteem ter verificatie, validatie en controle van AI-systemen, op basis van een breed spectrum aan standaarden op het gebied van veiligheid, transparantie, begrijpelijkheid, verklaarbaarheid en ethische waarden.

Ook het Future of Life Institute heeft op een conferentie zijn 21 Asilomar AI Principles gelanceerd (2017). Ten aanzien van transparantie benadrukken zij dat er transparantie moet zijn ten aanzien van falen: als een AI-systeem schade aanricht, moet het mogelijk zijn om vast te stellen waarom. Een van de geïnterviewde experts gaf aan dat het daarom zinvol zou zijn dat de robot dan beschikt over een 'black box' in de positieve zin van het woord zoals in de vliegtuigen aanwezig is, zodat ook achteraf de oorzaken van het falen beter onderzocht kunnen worden.

[10] Zie voor meer informatie omtrent de GDPR: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

[11] Zie voor meer informatie omtrent de MiFID: https://ec.europa.eu/info/business-economy-euro/banking-and-finance/financial-markets/securities-markets/investment-services-and-regulated-markets-markets-financial-instruments-directive-mifid_en.

Om risico's voor gebruikers te beheersen dient de robot ook zijn intenties duidelijk te maken via zijn interface. Denk daarbij bijvoorbeeld aan een robot die zijn intentie om te gaan bewegen duidelijk maakt door middel van een 'led lampje', wat de te bewegen richting aangeeft. Een ander voorbeeld om intenties te duiden, is door de voorspelde posities bijvoorbeeld te plotten op de operator control interface, zodat de operator meekrijgt waar de robot zich in de toekomst zal bevinden en daarop kan anticiperen.

Weller (2017) geeft een aantal voorbeelden van informatie die bijvoorbeeld robots kunnen terugkoppelen om de transparantie te verbeteren. Dit kan op verschillende niveaus. Ontwikkelaars willen kunnen zien of het systeem zijn taak goed of slecht uitvoert en de oorzaken daarvan. Dit geldt ook voor de gebruiker. Een gebruiker wil weten wat de robot gaat doen en waarom om zo beter vertrouwd te raken met de robot in geval van toekomstige (onvoorziene) situaties. Ook wil een gebruiker of onderhoudsmonteur weten hoe een bepaalde voorspelling of beslissing is bereikt en kunnen controleren of het onderliggende systeem nog werkt volgens wettelijke eisen van de Machinerichtlijn en de ontwerpspecificaties van de fabrikant en klant. Verder zouden gegevens bewaard kunnen worden, waarmee experts, zoals de ontwikkelaars van het algoritme (om verbeteringen te kunnen aanbrengen), of wetenschappelijke experts (in het kader van een incidentanalyse) kunnen herleiden waarom iets fout is gegaan.

3.2.3 Transparantie in het ontwikkelproces om kwaliteit te waarborgen

Naast het belang van transparantie in de data en algoritmes van een robot tijdens het gebruik, is het ook van belang dat er transparantie geschept wordt in het ontwikkelproces (d.w.z., van design tot prototype) van de algoritmes waar een robot mee werkt. Eén van de wetenschappelijke experts legt uit dat ontwikkelaars vaak meerdere tests uitvoeren omtrent de prestaties van de robot (denk hier aan zowel de uitvoering van een taak, alsmede het voldoen aan eisen uit de machinerichtlijn), maar dat zij de tests met slechte prestaties niet altijd vrijgeven. Hierdoor kan de ontwikkelaar een beter beeld van de prestaties van de robot schetsen.

Een bekende vorm van manipulatie van onderzoeksresultaten is probability-hacking (P-hacking). P-hacking is een term die voortkomt uit een strijd die woedt rond statistiek (Head, Holman, Lanfear, Kahn, & Jennions, 2015). De term is gerelateerd aan de maatstaf die wordt gebruikt voor significantie: de p-waarde. In de sociale wetenschap blijkt dat wanneer je een onderzoek herhaald er vaak iets heel anders uit komt dan de eerste keer. Hoewel P-hacking vooral binnen het sociale wetenschappelijke domein aandacht krijgt, is het voor andere domeinen van belang (bijv. bij het toetsen van technische specificaties voor een veilig ontwerp). Met oog op de ambitie om te innoveren en de concurrentie voor te zijn, zouden onderzoekers een onderzoek net zo lang met verschillende maatstaven kunnen blijven herhalen totdat deze een mooi resultaat opleveren om te publiceren. Hier wordt op gecontroleerd door de eis van repliceerbaarheid te stellen aan onderzoek. Met andere woorden, een onderzoek moet door andere onderzoekers exact te repliceren zijn om de resultaten te kunnen controleren.

Als robots in de praktijk komen die met P-hacking als goede en veilige robots worden gedefinieerd, dan levert dit een gevaar op voor de veiligheid van de mens. Het zal echter heel moeilijk te controleren zijn of bepaalde schade uit onveilige situaties verband hebben met P-hacking. Dit omdat de eerdere onderzoeken die een slechtere prestatie opleverden,

waarschijnlijk niet openbaar gemaakt zullen zijn door de ontwikkelaar. Alleen het onderzoek dat ‘aantoont’ dat de robot goed presteert zal openbaar zijn. Hierdoor ontstaat er dus een vertekend beeld over de kwaliteit van de robot. Bij de praktijk van P-hacking is het belang van succes groot waardoor de ontwikkelaar graag een goed beeld van de robot wil schetsen. Tegelijkertijd is de pakkans klein omdat P-hacking moeilijk te controleren is.
In de medische sector wordt deze situatie voorkomen door middel van een pre-registratie eis van het onderzoek naar de effectiviteit van een medicijn. Hierdoor wordt het inzichtelijk of er mogelijk meerdere onderzoeken zonder effect voorgingen aan een publicatie waar wel (toevallig) een effect gevonden is. Als resultaat is de kans kleiner dat niet werkende (of zelfs schadelijke) medicatie op de markt komen.

Er is op het moment geen juridische eis van een pre-registratie van onderzoeksopzet bij de ontwikkeling van robots. Hofman, Sharm en Watts (2017) stellen verschillende eisen aan de onderzoeksopzet voor het vaststellen van de prestaties van een robot. Deze eisen zijn gericht op het transparant houden van het onderzoek, waardoor onderzoeksresultaten die bij toeval zijn gevonden beter herkend kunnen worden. Verder verhogen deze eisen bijvoorbeeld de repliceerbaarheid van het onderzoek.

Een pre-registratie van de onderzoeksopzet zou openheid creëren omtrent het ontwikkelproces van de machine learning capaciteiten van de robot. Op deze wijze wordt het voor andere onderzoekers inzichtelijk wat de (ontwikkel)geschiedenis en prestaties van een algoritme zijn. Bij de pre-registratie van de onderzoeksopzet moet onder meer duidelijk worden:

- Welke type datasets er zijn gebruikt.
- Wat de verhouding tussen trainingsdata en validatiedata is.
- Hoe vaak en op welk soort ‘out of sample data’ is getest.
- Hoe de hypothese (over hoe de machine het beste kan leren) tot stand is gekomen;
- Alle pre-processing keuzes: het gaat in deze niet alleen om de keuze van de data, maar ook op welke wijze data wordt opgeschoond, hoe deze is gelabeld en welke mogelijke (alternatieve) labels er waren.
- Wat voor typen algoritmen er zijn gebruikt, of welke typen men van plan is om te gaan gebruiken.

Het uiteindelijke doel van deze eisen is dat een te gebruiken applicatie daadwerkelijk doet wat degene die het naar de markt brengt beoogt en beweert dat het doet.

3.3 Verantwoordelijkheden

Bij het traject van ontwikkeling van een robot naar uiteindelijke toepassing in de praktijk, zijn meerdere partijen betrokken: ontwikkelaars van verschillende robotdelen, integrator die de robotdelen installeren bij de klant en de klant die de robot in gebruik neemt.

In de praktijk blijkt dat het nog niet altijd duidelijk is waar bepaalde verantwoordelijkheden liggen om ervoor te zorgen dat het eindproduct aan alle noodzakelijke eisen voldoet en blijft voldoen. Zolang hier verwarring over blijft bestaan zullen aanvullende wettelijke eisen niet tot het gewenste effect leiden. De (potentiële) risico's van AI-systemen worden onderworpen

aan plannings- en mitigatie-inspanningen. Deze inspanningen dienen evenredig te zijn met hun verwachte impact (proportionaliteit).

Uit de interviews komt naar voren dat de verantwoordelijkheid op ketenniveau afgedekt moet worden. De fabrikant dient zelf zorg te dragen dat product of dienst voldoet aan de V&G eisen.

3.4 Concept voorstel aanvullende essentiële V&G eisen voor machine learning

Op basis van bovenstaande informatie zijn vervolgens de volgende mogelijke aanvullende V&G eisen geformuleerd (naast de al bestaande V&G eisen voor besturingssystemen). Deze eisen zijn per onderwerp onderverdeeld, zoals hierboven besproken met een verwijzing naar de relevante paragraaf uit Bijlage 1 van de Machinerichtlijn (MR).

3.4.1 Mens in controle

Bij de ontwikkeling van machines met machine learning zou het uitgangspunt moeten zijn om de mens altijd in controle te laten zijn van de situatie (en machine). Belangrijke uitgangspunten hierbij zijn:

- De mens geeft opdracht aan de machine, de machine geeft niet zelf opdracht aan de mens.
- De mens moet altijd kunnen ingrijpen en de machine veilig kunnen ‘overrulen’.
- Mens-machine interface moet zo ontworpen zijn dat de mens altijd weet wat de machine gaat doen.

Verder is het van belang dat voor een machine met machine learning de fysieke omgeving gedefinieerd wordt, waarin de robot veilig kan worden ingezet (zie kader *Kaderen leerpotentieel machine).* Ook dient met betrekking tot de hardware voorop te staan dat de omgeving zo wordt ingericht dat de functionaliteit toeneemt en veiligheid en gezondheid van de gebruiker worden gewaarborgd.

Aanvullende V&G eisen:

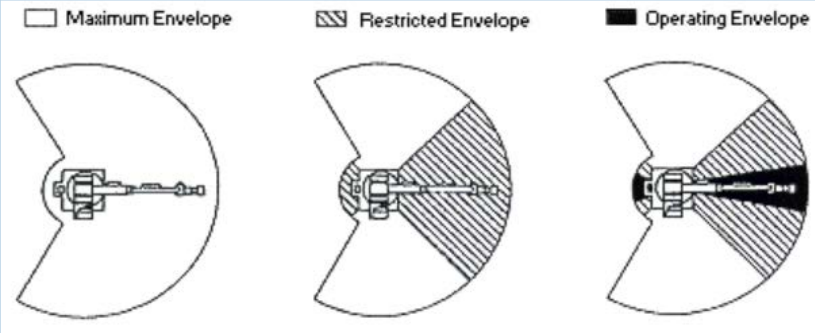
- Machines met machine learning technologie mogen niet in een situatie terecht komen ofwel geplaatst worden, waarin zij zelf afwegingen moeten maken in relatie tot schade aan mensen en of omgeving. (aanvulling op: Veiligheid en betrouwbaarheid van de besturingssystemen; paragraaf 1.2.1, Bijlage 1 MR).
- De machine met machine learning technologie moet worden voorzien van een noodstopfunctie die ten alle tijden kan worden uitgeschakeld/ kan worden overruled. De situatie is na het uitschakelen veilig. (aanvulling op: Stopzetting; paragraaf 1.2.4, Bijlage 1 MR).
- Machine learning mag niet leiden tot nieuw gedrag van de machine buiten zijn gedefinieerde taak en werkomgeving (aanvulling op: Veiligheid en betrouwbaarheid van de besturingssystemen; paragraaf 1.2.1, Bijlage 1 MR).

Kaderen leerpotentieel machine

Een manier om de risico's van de onvoorspelbaarheid van een machine met machine learning weg te nemen, is door het leerpotentieel te kaderen. Bij traditionele industriële machines is het werkbereik van de machine gedefinieerd (zie Figuur 3, uit Section IV: Hoofdstuk 4, OSHA Technical manual (OSHA, z.d.)). Het werkbereik is relatief klein (zwart), omdat de machine dezelfde beweging herhaald. Het maximale potentiële bereik (wit) is echter veel groter. Voor machines met machine learning is hun gehele maximale werkbereik potentieel hun werkbereik. Dit betekent dat er minder ruimte rond de machine veilig is voor de gebruiker om zich te bevinden (in figuur 3 is het veilige gebied alles wat buiten het gestreepte deel valt), omdat de specifieke acties van de machine niet meer vast staan.

Om dit risico te beheersen kan men het werkbereik (zwart) waarbinnen de robot nieuwe acties mag toepassen duidelijk definiëren en eventueel inperken.

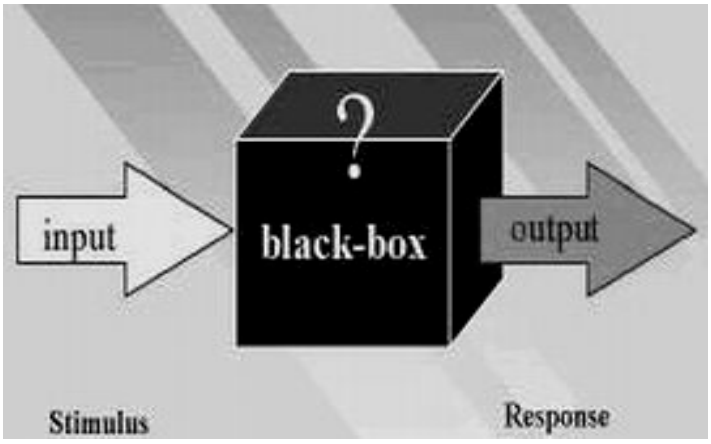
Het voor de gebruiker verboden gebied (restricted) zal hier vervolgens ook weer op aangepast moeten worden. In deze situatie blijft de robot in staat zijn gedrag aan te passen, maar is de onvoorspelbaarheid van zijn gedrag beperkt doordat deze gekaderd en gedefinieerd is.



Figuur 3 Visualisatie van de gedefinieerde bewegingsruimte van een industriële machine. Van links naar rechts de maximale bewegingsruimte (wit), de niet toegankelijke bewegingsruimte (gestreept) en de gebruikte bewegingsruimte (zwart)

3.4.2 Transparantie

Door toenemende complexiteit van het besturingssysteem en de ontwikkelingen op het gebied van machine learning (en AI), dreigt het besturingssysteem van de industriële machine een black box te worden (zie Figuur 4). Bij de traditionele industriële machines, was dit mogelijk ook al het geval, echter bij deze machines kon men erop vertrouwen dat bij een gelijke input, er een gelijke output zou komen. Tenzij er iets fout was met de machine. Het is echter wel problematisch als de machine van zijn oorspronkelijke programmering kan gaan afwijken en dus een andere output kan vertonen op dezelfde input. Dit gebeurt als de machine de input anders gaat beoordelen. Er moeten dus stappen genomen worden om het proces tussen input en output expliciet transparant te maken bij machines met leervermogen.



Figuur 4[12] Hoe de output op basis van de input wordt bepaald is onduidelijk.

Enerzijds moet de transparantie van de machine naar de gebruiker op de werkvloer toenemen. Dit kan door middel van communicatie van het huidige bewegingspatroon en mogelijke afwijkingen daarvan. Men kan bijvoorbeeld ook denken aan structurele (eind van de dag) feedback over wat de machine heeft geleerd en dat dit voldoet aan de op dat moment geldende V&G eisen. Anderzijds moet te allen tijde de software uit te lezen zijn. Hierdoor kan herleid worden waar het huidige bewegingspatroon op gebaseerd is en hoe een incident heeft kunnen plaatsvinden. Hiervoor is het belangrijk dat voor het ontwikkelproces al transparant is:

- Wat de originele code voor het algoritme is.
- Welke ruwe data is gebruikt voor het algoritme.
- Wat zijn de acties die de machine op basis van de data en het algoritme uitvoert?

Een typekeur zou inzichtelijk moeten maken welke vereisten er zijn toegepast en waaraan is voldaan om te spreken van transparantie van datasets en algoritmes. In de praktijk betekent dit dat ontwerpers, bouwers, en fabrikanten die een machine met machine learning maken, de machine of het productieproces moeten voorleggen aan een conformiteitsbeoordelingsinstantie (zie Verordening 765/2008 Artikel 2, lid 13) voordat de machine in de handel wordt gebracht. Deze instanties ijken, testen, certificeren en inspecteren de machineonderdelen.

3.4.2.1 Transparantie in algoritmes tijdens gebruik van de machine

Aanvullende V&G eisen:

- Een machine met machine learning technologie moet in staat zijn om gegeven zijn rol adequaat en passend te kunnen reageren op de mens (verbaal via woorden en/ of non-verbaal via gebaren, gezichtsuitdrukking of lichaamsbeweging). (aanvulling op: Ergonomie ;paragraaf 1.1.6, Bijlage 1 MR).
- Een machine met machine learning technologie moet in staat zijn om zijn intenties (wat hij gaat doen en waarom) op een begrijpelijke manier te communiceren naar de werknemer. (aanvulling op: Ergonomie ;paragraaf 1.1.6, Bijlage 1 MR).

[12] Plaatje overgenomen van https://tvtropes.org/pmwiki/pmwiki.php/Main/BlackBox

- Machines met machine learning technologie dienen voor de gebruiker te onderscheiden welk deel van hun analyse gestoeld is op ‘supervised learning’ en welk deel gestoeld is op ‘unsupervised learning’, teneinde de totstandkoming van de analyse transparant te maken. (aanvulling op: Veiligheid en betrouwbaarheid van de besturingssystemen ;paragraaf 1.2.1, Bijlage 1 MR).

3.4.2.2 Transparantie in gebruikte data sets en algoritmes in het besturingssysteem

Aanvullende V&G eisen:

- Het gedrag van een machine met machine learning technologie moet (vooraf en achteraf) te herleiden zijn op basis van transparantie in toegepaste datasets, testomgevingen (incl. gebruikte scenario’s in trainings- en validatiemodellen van het algoritme) en afwegingskaders of toetsingscriteria voor beslissingen van algoritmen. (aanvulling op: Veiligheid en betrouwbaarheid van de besturingssystemen ;paragraaf 1.2.1, Bijlage 1 MR).
- De beslissingen genomen door een machine met machine learning technologie moet gelogd worden en bewaard blijven. (aanvulling op: Veiligheid en betrouwbaarheid van de besturingssystemen; paragraaf 1.2.1, Bijlage 1 MR).

3.4.2.3 Transparantie tijdens het ontwikkelproces

Aanvullende V&G eis:

- Alleen machine learning technologie met openbare pre-registratie van het research design van onderzoek, inclusief bijbehorend performance rapport van de prestaties van de machine, mogen in gebruik worden genomen. (aanvulling op: Veiligheid en betrouwbaarheid van de besturingssystemen; paragraaf 1.2.1, Bijlage 1 MR).

3.4.3 Verantwoordelijkheden

Er wordt geen aanvullende V&G eis voorgesteld op basis van dit onderwerp.

4 Methode - Deel 2

Op 31 Juli en 2 Augustus 2018 zijn (intern bij TNO) twee werksessies gehouden om de hierboven beschreven voorgestelde V&G eisen te valideren. In beide werksessies werden TNO experts uitgevraagd over de compleetheid, toepasbaarheid, en noodzakelijkheid van de geïdentificeerde V&G eisen.

Tabel 2 Deelnemers werksessies en hun specialisatie

Functie	Specialisatie
TNO expert 1	Neurale netwerken, 'deep learning' en intelligente beeldvorming.
TNO expert 2	Versterking van Visuele kwaliteiten in Robotsystemen.
TNO expert 3	Cybersecurity & ICT
TNO expert 4	Psychometrie en statistiek
TNO expert 5	Algemene Kunstmatige Intelligentie & Ethiek
TNO expert 6	Gedragsmodellering en Algoritmetesten

Deelnemers kregen ter voorbereiding onder andere de eerste versie van de voorgestelde V&G eisen toegestuurd gesorteerd op relevante paragraaf van de Machinerichtlijn (zie kader volgende pagina). Tijdens de werksessie zijn de voorgestelde V&G eisen beoordeeld op volledigheid en mate van belangrijkheid. Aan het eind van de werksessie zijn deelnemers gevraagd een score te geven aan de V&G eisen op basis van hun relevantie:

0 = niet belangrijk;
1 = enigszins belangrijk;
2 = belangrijk;
3 = zeer belangrijk.

Essentiele V&G eisen die zijn toegestuurd naar werksessiedeelnemers

Huidige essentiële V&G eisen van het besturingssysteem (doelvoorschriften):

- V&G eis 1: Het besturingssysteem van de machine Is bestand tegen de normale bedrijfsbelasting en tegen invloeden van buitenaf.
- V&G eis 2: Een storing in het besturingssysteem van de machine mag niet tot een gevaarlijke situatie leiden.
- V&G eis 3: Fouten in de besturingslogica mogen niet tot een gevaarlijke situatie leiden.
- V&G eis 4: Menselijke fouten gedurende de werking mogen niet tot een gevaarlijke situatie leiden.

Aanvullende essentiële V&G eisen (doelvoorschriften):

T.a.v. de ergonomie van het besturingssysteem (paragraaf 1.1.6, Bijlage 1 MR)

- V&G eis 1: Een machine met machine learning technologie moet in staat zijn om gegeven zijn rol adequaat en passend te kunnen reageren op de mens (verbaal via woorden en/ of non-verbaal via gebaren, gezichtsuitdrukking of lichaamsbeweging).
- V&G eis 2: Een machine met machine learning technologie moet in staat zijn om op een begrijpelijke manier te communiceren naar de werknemer wat hij gaat doen en waarom hij dat gaat doen.

T.a.v. veiligheid en betrouwbaarheid van de besturingssystemen (paragraaf 1.2.1, Bijlage 1 MR)

- V&G eis 3: Machines met machine learning technologie mogen niet in een situatie terecht komen ofwel geplaatst worden, waarin zij zelf afwegingen moeten maken in relatie tot schade aan mensen en of omgeving.
- V&G eis 4: Machine learning mag niet leiden tot nieuw gedrag van de machine buiten zijn gedefinieerde taak en werkomgeving
- V&G eis 5: Machines met machine learning technologie dienen voor de gebruiker te onderscheiden welk deel van hun analyse gestoeld is op ‘supervised learning’ en welk deel gestoeld is op ‘unsupervised learning’, teneinde de totstandkoming van de analyse transparant te maken.
- V&G eis 6: Het gedrag van een machine met machine learning technologie moet (vooraf en achteraf) te herleiden zijn op basis van transparantie in toegepaste datasets, testomgevingen (incl. gebruikte scenario’s in trainings- en validatiemodellen van het algoritme) en afwegingskaders of toetsingscriteria voor beslissingen van algoritmen.
- V&G eis 7: De beslissingen genomen door een machine met machine learning technologie moet gelogd worden en bewaard blijven.
- V&G eis 8: Alleen machine learning technologie met openbare pre-registratie van het research design van onderzoek, inclusief bijbehorend performance rapport van de prestaties van de machine, mogen in gebruik worden genomen.

T.a.v. stopzetting (paragraaf 1.2.4, Bijlage 1 MR)

- V&G eis 9: De machine met machine learning technologie moet worden voorzien van een noodstopfunctie waarbij deze ten alle tijden kan worden uitgeschakeld/ kan worden overruled. De situatie is na het uitschakelen veilig.

5 Resultaten – Deel 2

Na deelname aan de werksessies, is de deelnemers gevraagd de voorgelegde V&G eisen een score te geven op basis van hoe belangrijk deze zijn voor de veiligheid en gezondheid van gebruikers, die met deze machines werken. In onderstaande tabel zijn alle scores weergegeven. In de laatste kolom is ter illustratie van de mate van belangrijkheid een somscore weergegeven.

Duidelijk is te zien dat er twee V&G eisen die minder belangrijk worden geacht ten opzichte van de overige. Dit betreft V&G eis *5 Onderscheid SML en* UML en V&G eis *8 Pre-registratie onderzoek*. In de volgende sectie gaan wij verder in op de onderbouwing van het oordeel van de gedefinieerde V&G eisen en wat dit betekent voor het definitieve voorstel van V&G eisen.

Tabel 3 beoordeling voorgestelde essentiele V&G eisen

Essentiele V&G eisen voor besturingssystemen (Bijlage I, artikel 1.2.1 MR)	**Expert 1**	**Expert 2**	**Expert 3**	**Expert 4**	**Expert 5**	**Expert 6**	**Totaal**
V&G eis 1: Belasting en invloeden	2	2	3	3	3	3	16
V&G eis 2: Storing	3	3	3	3	3	3	18
V&G eis 3: Logica fout	3	3	2	3	3	3	17
V&G eis 4: Menselijke fout	3	3	3	3	3	3	18
Voorgestelde aanvullende essentiële V&G eisen voor machine learning							
V&G eis 1: Reageren op de mens	2	1	3	3	2	3	14
V&G eis 2: Intenties communiceren	3	3	0	3	2	3	14
V&G eis 3: Geen afwegingen schade	2	2	3	3	3	1	14
V&G eis 4: Limitatie nieuw gedrag	2	2	3	3	3	1	14
V&G eis 5: Onderscheid SML en UML	0	0	0	0	1	2	3
V&G eis 6: Transparantie herleiden gedrag	2	1	0	3	3	1	10
V&G eis 7: Transparantie beslissingen	2	3	3	3	2	2	15
V&G eis 8: Pre-registratie onderzoek	1	1	*	*	3	1	6
V&G eis 9: Noodstopfunctie	3	3	3	3	3	1	16

0 = niet belangrijk;
1 = enigszins belangrijk;
2 = belangrijk;
3 = zeer belangrijk;
*Geen score gegeven.

5.1 Definitief voorstel aanvullende essentiële V&G eisen voor machine learning

5.1.1 V&G eis 1: Reageren op de mens

De V&G eis zoals voorgesteld in conceptversie bevatte willekeurige voorbeelden die niet uitputtend zijn en daarom niet noodzakelijk zijn om te noemen voor een doelvoorschrift. Daarom zijn de voorbeelden verwijderd.

Definitieve versie aanvullende V&G eis:
Een machine met machine learning technologie moet in staat zijn om adequaat en passend te kunnen reageren op de mens.

5.1.2 V&G eis 2: Intenties communiceren

Volgens de deelnemers is het communiceren van intentie van de machine (wat gaat het doen), een minimale vereiste die technisch al mogelijk is. Volgens enkele werksessiedeelnemers zullen goed geprogrammeerde systemen er ook een waarom toelichting bij geven, omdat dit een logische verdere uitwerking is van wat het systeem gaat doen. Andere werksessiedeelnemers beargumenteerde juist dat zodra de machine een waarom toelichting gaat geven er geen toepassingen van ‘deep learning’ (gegeven de complexiteit van de methode), en in het verlengde sommige toepassingen van machine learning, meer mogelijk zijn. Het communiceren waarom een machine een bepaalde actie gaat doen, gaat voor deze groep deelnemers technisch te ver om te realiseren.

In dit rapport is veel aandacht besteed aan waarom transparantie van de machine met machine learning een belangrijk element is om de veiligheid van de gebruiker te borgen. Daarom is besloten om de ‘waarom’ vraag in de V&G eis te behouden. Dit zal mogelijk betekenen dat sommige vormen van machine learning nog niet in industriële machine gebruikt kunnen worden totdat deze beter begrepen worden.

Verder bleek uit de discussie dat het woord ‘waarom’ kan leiden tot verwarring. De discussie kwam bijvoorbeeld ook op het punt dat voor machines met machine learning om ook daadwerkelijk te begrijpen waarom zij iets doet er geavanceerde AI nodig is. Deze interpretatie was ook niet bedoeld met de V&G eis. Daarom is de eis aangepast zodat deze niet meer het woord ‘waarom’ bevat, maar expliciet stelt dat de machine moet aangeven op basis van welke informatie deze een actie uitvoert.

Definitieve versie aanvullende V&G eis:
Een machine met machine learning technologie moet aangeven welke acties deze gaat ondernemen en op basis van welke informatie.

5.1.3 V&G eis 3: Geen afweging over schade aan mensen

Als mens en machine elkaar kunnen aanraken dan is het onvermijdelijk dat er afwegingen gemaakt worden. Gevolg van deze eis in oorspronkelijke vorm is dat de machine gescheiden moet blijven van de mens. Echter, het doel van de eis, is om er voor te zorgen dat het niet de machine is die de uiteindelijke beslissing neemt over ethische kwesties. Daarom is de V&G eis geherformuleerd om duidelijk te maken dat deze ethische beslissingen niet aan de machine mogen worden overgelaten.

Wel wordt herkend dat er een dilemma blijft in situaties waarin in principe ‘on the fly’ een beslissing moet worden genomen. Het uitgangspunt zal volgens sommige deelnemers moeten zijn dat de uiteindelijke uitkomst in deze situaties gebaseerd is op wat de maatschappij toelaatbaar vindt en dat dit niet door een algoritme bepaald is.

Definitieve versie aanvullende V&G eis:
Machines met machine learning mogen geen beslissingen of overwegingen maken in relatie tot schade aan mens en omgeving.

5.1.4 V&G eis 4: Limitatie nieuw gedrag

In de concept V&G eis is behalve de mens ook de ‘omgeving’ opgenomen. Dit vereist nog discussie over wat een gedefinieerde werkomgeving minimaal is. Verder zou je volgens de werksessiedeelnemers het woord ‘gedrag’ beter kunnen vervangen door het woord ‘acties’.

Definitieve versie aanvullende V&G eis:
Machine learning mag niet leiden tot nieuwe acties van de machine buiten zijn gedefinieerde taak en werkomgeving.

5.1.5 V&G eis 5: Onderscheid SML en UML

Volgens de werksessiedeelnemers kun je wel onderscheid maken welke beslissingen op basis van SML en welke beslissingen op basis van UML gemaakt zijn, maar is dit niet zinvol. Zelfs als je een goede definitie hebt, heb je er niets aan om te weten of het SML of UML is. Het is niet relevant voor de eindgebruiker. De kern is om vast te stellen in welk deel van het algoritme er iets mis ging om de corrigeerbaarheid mogelijk te maken. Dit wordt in eis 7 opgevangen. Wel is het goed om corrigeerbaarheid dus op te nemen van de beslissingen.

Definitieve versie aanvullende V&G eis:
Machines met machine learning technologie dienen bij verkeerde beslissingen van de machine achteraf corrigeerbaar te zijn om te voorkomen dat het in de toekomst op dezelfde manier fout gaat.

5.1.6 V&G eis 6: Transparantie herleiden gedrag

Om gedrag van de machine te herleiden is methodologische transparantie nodig. Dit wordt door een aantal werksessiedeelnemers als een extreme eis gezien die zelfs in ISO normen voor laboratoria al lastig te definiëren is (vandaar de lage score door sommige experts, Tabel 3). Daarnaast wordt opgemerkt dat als een robot zelfstandig of via reinforcement learning iets leert je de uitkomst (wanneer de data niet gedeeld wordt) ook niet meer kan controleren en/of corrigeren. Het onderzoeksteam heeft besloten dat alhoewel het naleven van de V&G eis moeilijk zal zijn, dit niet afdoet aan het belang van deze eis om machine met machine learning inherent veilig te maken. Daarom heeft het onderzoeksteam besloten deze wel op te nemen. Ook is hier nogmaals het woord ‘gedrag’ door ‘acties’ vervangen.

Definitieve versie aanvullende V&G eis:
De acties van een machine met machine learning technologie moet vooraf en achteraf te herleiden zijn op basis van transparantie in toegepaste datasets, testomgevingen[13] en afwegingskaders of toetsingscriteria voor beslissingen van algoritmen.

[13] Bijvoorbeeld gebruikte scenario’s in trainings- en validatiemodellen van het algoritme

5.1.7 V&G eis 7: Transparantie beslissingen

Geen opmerkingen. Deelnemers waren akkoord.

Dat de machine en zijn onderdelen te auditen zijn is een belangrijke voorwaarde. Door de datasets die de machine gebruikt vast te leggen in een systeem wordt de machine ook te auditen. Het beslisproces van de machines met machine learning moeten bovendien gelogd worden, bewaard blijven om te auditen.

Definitieve versie aanvullende V&G eis:
Het beslisproces van een machine met machine learning technologie moet gelogd worden en bewaard blijven[14].

5.1.8 V&G eis 8: Pre-registratie onderzoek

De deelnemers aan de workshop vinden de intentie van deze eis goed: om via wetenschappelijke onderbouwing aan te laten tonen dat de machine veilig is. Deze eis is echter moeilijk voor een MKB bedrijf. Datasets hoeven vooralsnog niet vrijgegeven te worden, waardoor het intellectuele eigendom van organisaties beschermt blijft. Er zijn ook alternatieve methodologische verificaties mogelijk om de kwaliteit te borgen, hier is echter in dit rapport niet verder op ingegaan.

Deelnemers waren het niet eens of de pre-registratie van datasets als V&G eis dient te worden opgenomen. Te meer omdat hier volgens de deelnemers vooral sprake is van een middelvoorschrift om transparantie van de gebruikte datasets en algoritmen te borgen. Als onderzoeksteam hebben wij daarom besloten dat alhoewel de pre-registratie van onderzoek een belangrijk middel is om de kwaliteit te borgen van nieuwe designs die op de markt komen, het niet per se een V&G eis is. Met andere woorden, de noodzaak of plicht om pre-registratie in te voeren zal op een andere manier overgebracht moeten worden. Deze V&G eis is daarom niet opgenomen in de definitieve set V&G eisen.

5.1.9 V&G eis 9: Noodstopfunctie

Deze eis is volgens de werksessie deelnemers haalbaar want je zet de machine gewoon uit op basis van een ‘hard wired’ veiligheidscircuit. Volgens een aantal deelnemers voegt dit echter niets nieuws toe voor sec machine learning want dit geldt ook al voor niet machine learning gedreven machines (bijv. sensoren) zoals in Bijlage 1, paragraaf 1.2.4.3.) van de Machine Directive al is beschreven. Deze eis kan daarmee ook komen te vervallen in de nieuwe (aanvullende) essentiële V&G eisen.

5.1.10 Opmerkingen met betrekking tot Huidige V&G eisen van het besturingssysteem

De eerste bestaande V&G eis voor het besturingssysteem beschrijft dat het besturingssysteem bestand moet zijn tegen invloeden van buitenaf. In het kader van machines met machine learning wil je juist invloeden van buitenaf. Hier zou een specificering gewenst zijn dat het bij machines met machine learning ook ongewenste invloeden van buitenaf betreft.

[14] Op een zodanige manier dat deze informatie voor een vast te stellen minimum periode beschikbaar blijft om bijvoorbeeld tijdens audits of incidentanalyses gecontroleerd te worden.

Voorstel aanpassing V&G eis: Het besturingssysteem van de machine is bestand tegen de normale bedrijfsbelasting en tegen ongewenste invloeden van buitenaf.
Aanvullend kan in het algemeen gesteld worden, dat wanneer er fouten of onverwachte omstandigheden optreden in het besturingssysteem het gewenst is dat de machine terugvalt in een veilige staat. Hierbij zou dan als aanvullende barrière de verplichting kunnen worden gesteld dat alleen een door de organisatie geautoriseerd persoon (bijv. de onderhoudsmonteur) de machine weer kan vrijgegeven.

5.2 Overzicht definitief voorstel

Huidige V&G eisen van het besturingssysteem (Bijlage 1, paragraaf 1.2.1):
1. Het besturingssysteem van de machine is bestand tegen de normale bedrijfsbelasting en tegen *ongewenste* invloeden van buitenaf.
2. Een storing in het besturingssysteem van de machine mag niet tot een gevaarlijke situatie leiden.
3. Fouten in de besturingslogica mogen niet tot een gevaarlijke situatie leiden.
4. Menselijke fouten gedurende de werking mogen niet tot een gevaarlijke situatie leiden.

Aanvullende V&G eisen:

Ten aanzien van ergonomie van het besturingssysteem (aanvulling op: paragraaf 1.1.6, Bijlage 1 MR)
1. Een machine met machine learning technologie moet in staat zijn om adequaat en passend te kunnen reageren op de mens.
2. Een machine met machine learning technologie moet aangeven welke acties deze gaat ondernemen en op basis van welke informatie.

Ten aanzien van veiligheid en betrouwbaarheid van de besturingssystemen (aanvulling op: paragraaf 1.2.1, Bijlage 1 MR)
3. Machines met machine learning mogen geen beslissingen of overwegingen maken in relatie tot schade aan mens en omgeving.
4. Machine learning mag niet leiden tot acties van de machine buiten zijn gedefinieerde taak en bewegingsruimte.
5. Machines met machine learning technologie dienen bij verkeerde beslissingen van de machine achteraf corrigeerbaar te zijn om te voorkomen dat het in de toekomst op dezelfde manier fout gaat.
6. De acties van een machine met machine learning technologie moet vooraf en achteraf te herleiden zijn op basis van transparantie in toegepaste datasets, testomgevingen en afwegingskaders of toetsingscriteria voor beslissingen van algoritmen.
7. Het beslisproces van een machine met machine learning technologie moet gelogd worden en bewaard blijven.

5.3 Overwegingen

Hieronder volgen nog enkele overwegingen die tijdens dit project naar voren zijn gekomen maar die niet direct relevant zijn voor het beantwoorden van de onderzoeksvraag in dit

rapport. Wel geven deze overwegingen nog belangrijke context waarbinnen bovenstaande V&G eisen moeten worden beschouwd.

5.3.1 Overweging 1

Indien de hierboven genoemde nieuwe V&G eisen voor machines met machine learning zouden worden opgenomen in de Machinerichtlijn, moet ook bepaald worden welke conformiteitsbeoordelingsprocedure(s) van toepassing is (zijn). Dit onderdeel (keuze conformiteitsbeoordelingsprocedure) maakt geen deel uit van onderhavig onderzoek.

De verschillende conformiteitsbeoordelingsprocedures, acht modules, zijn terug te vinden in Besluit 768/2008/EG[15]. In de kern gaat het daarbij om drie soorten procedures.
1. Uitgangspunt daarbij is dat de fabrikant zoveel mogelijk zélf de conformiteitsbeoordeling m.b.t. het ontwerp als de productie van de machine uitvoert (Module A – Interne productiecontrole).
2. Er kan ook gekozen worden voor een conformiteitsbeoordelingsprocedure op basis van een typekeur (de modules B, aangevuld modules C, D, E en F.). Hierbij wordt enerzijds het ontwerp van de machine beoordeeld op conformiteit met de V&G eisen en wordt anderzijds ook het vervaardigde product beoordeeld op conformiteit met de V&G eisen. Bij dit soort conformiteitsbeoordelingsprocedures moet een geaccrediteerde conformiteitsbeoordelingsinstantie worden betrokken.
3. Tot slot kan er gekozen worden voor een conformiteitsbeoordeling op basis van kwaliteitsborging (Modules D, E. en H.) Daarbij moet de fabrikant gebruik maken van kwaliteitsborgingssystemen, waarbij aangesloten kan worden bij de relevante internationale kwaliteitsnormen (ISO 9000 en ISO 9001). Ook hier is de betrokkenheid van een conformiteitsbeoordelingsinstantie verplicht.

5.3.2 Overweging 2

In dit rapport, hebben wij pre-registratie als voorbeeld gegeven van een middel om de kwaliteit van de machine learning software te borgen. Ook kan overwogen worden om in plaats van pre-registratie, de machine learning software te laten voorzien van een typekeur. Zoals bij overweging 1 (hfst. 5.3.1) is aangegeven, is het de bedoeling van een typekeur om het ontwerp van de machine te beoordelen op conformiteit met de V&G eisen. Vervolgens moet ook het vervaardigde product worden beoordeeld op conformiteit met de V&G eisen. De software van de machine met machine learning kan ook worden voorzien van een typekeur. Een typekeuring is dan een mogelijkheid om de transparantie van de software te borgen (zie: Besluit 768/2008/EG). Het typekeur van de software kan daarna in het verdere productieproces worden gebruikt om overeenstemming aan te tonen van de geïnstalleerde software met de software waar een typekeur voor is afgegeven of om bij de eenheidskeur (bij aflevering) te kunnen vaststellen of de software overeenkomt met de software waar een typekeur voor is afgegeven.

Nauw verbonden met bovenstaande is blockchain. Een blockchain is een gedistribueerd systeem van computers wat de deelnemers een soort gedeelde database geeft.

[15] Besluit nr. 768/2008/EG van het Europees Parlement en de Raad van 9 juli 2008 betreffende een gemeenschappelijk kader voor het verhandelen van producten en tot intrekking van Besluit 93/465/EEG van de Raad, Publikatieblad van de Europese Unie, 13 augustus 2008, L-218, p. 82-128.

Alle aanpassingen in deze database (de transacties) worden vastgelegd en alle deelnemers kan de integriteit controleren. Zolang de meerderheid van de deelnemende computers te vertrouwen is, zal enige manipulatie van eerder toegevoegde gegevens gedetecteerd kunnen worden. Op deze manier zou toepassing van blockchain eraan kunnen bijdragen meer transparantie te geven over welke aanpassingen er zijn geweest in de software tijdens de overgang van ontwerp naar eindproduct.

5.3.3 Overweging 3

Het uitgangspunt van de Europese Commissie is dat robotica en AI vooralsnog onder de richtlijn 2006/42/EG vallen. Echter zodra machines in de toekomst dankzij AI menselijke intelligentie kunnen evenaren (ofwel, High Machine Level Intelligence) vindt er een categorische verschuiving plaats, waarna deze 'autonome robots' mogelijk niet meer simpel als machine te duiden zijn. Er kunnen dan ook nieuwe ethische vraagstukken gaan spelen over bijvoorbeeld de rechten van een robot (heeft de robot het recht om niet uitgezet te worden). Deze overweging heeft verder geen aandacht gehad in het rapport, omdat deze ontwikkelingen pas op veel langere termijn wordt verwacht dan de beschreven ontwikkelingen op het gebied van machine learning in dit rapport.

5.3.4 Overweging 4

Er kan worden gedacht aan het maken van onderscheid tussen verschillende toepassingen en de rol die de machine learning algoritmes hebben, bij het bepalen van de acties van de machine. Naarmate de acties van de machine in hogere mate door machine learning worden bepaald, zal deze machine in een hogere veiligheidscategorie vallen. Vervolgens kan men voorstellen dat voor machine learning toepassing in een hogere veiligheidscategorie, er andere of extra V&G eisen van toepassing zijn dan machine learning toepassingen in een lagere categorie. Om een zinvolle invulling van de categorieën te maken zal eerst een inventarisatie van machine learning toepassingen op de werkvloer nodig zijn.

5.3.5 Overweging 5

We zijn ons ervan bewust dat de V&G eisen in de machinerichtlijn van toepassing zijn op fabrikanten en ontwikkelaars van machines voor zowel de professionele context als de consument. De verwachting is dat ook machines voor de consument (denk aan stofzuigers en grasmaaiers) zullen in toenemende mate met machine learning worden uitgerust. Hier spelen dezelfde thema's een rol, bijvoorbeeld de noodzaak van transparantie van de acties van de machine naar de consument toe. De hier voorgestelde V&G eisen zijn dus ook van toepassing voor deze machines ten behoeve van de consument. Wel dient opgemerkt te worden dat de voorgestelde aanvullende essentiële V&G eisen zijn opgesteld op basis van onderzoek naar machine learning toepassingen alleen bij industriële machines.

6 Referenties

EESC (2017). Advies van het Europees Economisch en Sociaal Comité over kunstmatige intelligentie — De gevolgen van kunstmatige intelligentie voor de (digitale) eengemaakte markt, de productie, consumptie, werkgelegenheid en samenleving. *Initiatiefadvies* (2017/C 288/01). https://eur-lex.europa.eu/legal-content/NL/TXT/?uri=CELEX%3A52016IE5369.

Eurofound (2017). Advanced industrial robotics: Taking human-robot collaboration to the next level. Working paper *The Future of Manufacturing in Europe (FOME)* project. https://www.eurofound.europa.eu/sites/default/files/wpfomeef18003.pdf

Eurofound (2018), *Automation, digitalisation and platforms: Implications for work and employment*, Publications Office of the European Union, Luxembourg.

Europese commissie (2018). Commission outlines European approach to artificial intelligence. https://ec.europa.eu/growth/content/commission-outlines-european-approach-artificial-intelligence_en

Future of Life Institute (2017). Asilomar AI principles. https://futureoflife.org/ai-principles/

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts, *Artificial Intelligence,* https://arxiv.org/abs/1705.08807.

Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., & Jennions, M.D. (2015). The extent and consequences of P-Hacking in science. *PLOS BIOL, 13*(3). https://doi.org/10.1371/journal.pbio.1002106

Hofman, J.M., Sharma, A., & Watts, D.J. (2017). Prediction and explanation in social systems, *Science*, *355*, 486-488.

Jansen, A., van der Beek, D., Cremers, A., Neerincx, M., & van Middelaar, J. (2018). Opkomende risico's voor arbeidsveiligheid: werken in dezelfde ruimte als een cobot. TNO report, TNO 2017 R11463.

OSHA (z.d.). Industrial Robots and Robot System Safety. Section IV: Chapter 4 van OSHA Technical manual. https://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_4.html

Robotics VO (2013). A roadmap for U.S. robotics: From internet to robotics. http://www.roboticscaucus.org/Schedule/2013/20March2013/2013%20Robotics%20Roadmap-rs.pdf

Russell, S. & Norvig, P. (2010). *Artificial Intelligence: A modern approach* (third edition). Pearson.

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210-229.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge university press.

Silver, D., Schrittwieser, J., Simonyan, K. Antonoglou, I., Huang, A. Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature, 550,* 354– 359.

Steijn, W., Luiijf, E., & van der Beek, D. (2016). Opkomend risico voor arbeidsveiligheid door inzet van robots op de werkvloer. TNO report, TNO 2016 R10643.

Steijn, W., van der Vorm, J., Luiijf, E., Gallis, R., van der Beek, D. (2016). Opkomende risico's voor arbeidsveiligheid als gevolg van IT-koppelingen van en tussen arbeidsmiddelen. TNO report, TNO 2016 R10096.

A Appendix: Interviewprotocol

A.1 Protocol 1. Praktijkdeskundige

- Introductie.
- Inleiding.
  - Ruimte voor geïnterviewde om over zijn eigen achtergrond met betrekking tot het onderwerp uit te wijden.
  - Toelichting op praktijktoepassing.
- Wat wil men van de robot van de toekomst?
  - Qua mobiliteit.
  - Qua men-machine interactie.
  - Qua autonomie.

  - Welke functies.
    - Interactie – *toelichten.*
    - Autonomie – *toelichten, evt. benadrukken dat we beslissingsautonomie bedoelen, robots bepalen hun gedrag.*
    - Taak complexiteit – *toelichten.*
- Wat gebeurt er als machines worden uitgerust met Machine learning (ook wel deep learning) technologie.
- Wat is het effect van toename in robots, die in (een zelfde) (werk)omgeving worden ingezet?
- *Kunt u daarbij onderscheid maken in risico's op technisch falen en menselijk falen, denk ook aan mens-ontwerp, en ethische bezwaren/risico's voor de veiligheid en gezondheid.*

- Verwachte effecten?
- Wat zijn de verwachte risico's met betrekking tot de veiligheid van de mens?
  - *Kunt u daarbij onderscheid maken in risico's op technisch falen en menselijk falen, denk ook aan mens-ontwerp, en ethische bezwaren/risico's voor de veiligheid en gezondheid.*
  - Wat zijn de beheersmaatregelen?
- Wat zou er veranderen als een dergelijke robot door een consument wordt gebruikt?

- Wettelijke verplichtingen?
  - Machine richtlijnen?

- Incidenten?

- Is het aansturingssysteem van de robot/ centraal aansturingssysteem beschermd tegen externe invloeden (denk aan virussen of hackers).
  - Zo ja, hoe?

- Voornaamste reden om voor een robot te kiezen.
- Afsluiting.

A.2 Protocol 2. Wetenschappelijke expert

Introductie

- Inleiding.
  - Ruimte voor geïnterviewde om over zijn eigen achtergrond met betrekking tot het onderwerp uit te wijden.

- Toelichting op Expertise terrein.
  - Bestaande toepassingen op robots?
  - Toekomstige toepassingen:
    - Korte termijn (binnen 10 jaar).
    - Lange termijn (Na 10 jaar).
- Risico's op het gebied van arbeidsveiligheid (als gevolg van toenemende mobiliteit, mens-machine interactie, autonomie en machine learning).
  - *Kunt u daarbij onderscheid maken in risico's op technisch falen en menselijk falen, denk ook aan mens-ontwerp, en ethische bezwaren/risico's voor de veiligheid en gezondheid.*
- Wat zou er veranderen als de robot door een consument wordt gebruikt?
- Welke maatregelen om de veiligheid te verbeteren zijn hierbij denkbaar/noodzakelijk?

- In hoeverre zijn de toepassingen al geborgd in wetgeving (als de Machinerichtlijn)?
  - Welke richtlijnen of wetten gebruiken jullie?
  - Waar liggen spanningsvelden binnen gebruikte richtlijnen?
  - Waar bestaat nog geen regelgeving of richtlijn voor, maar moet die wel komen?

- Afsluiting.

TNO.NL

Healthy Living
Schipholweg 77-89
2316 ZL Leiden
Postbus 3005
2301 DA Leiden

www.tno.nl

T +31 88 866 90 00
infodesk@tno.nl

Handelsregisternummer 27376655