Robots die raad weten met morele dilemma's

By **Dorine Schenk**, www.nrc.nl
november 3de, 2017

Een leger boze robots dat zich tegen de mensheid keert. Het is een bekend scenario uit sciencefictionfilms. Maar de razendsnel ontwikkelende technologie laat weinig aan de verbeelding over. De angstaanjagende filmpjes van het robotbedrijf Boston Dynamics tonen robots die zowel op ongelijke als op gladde grond zich snel en soepel bewegen. En nieuwe, zelflerende software: eind mei versloeg Google's computerprogramma AlphaGo weer een wereldkampioen in het voor computers ooit onbereikbare bordspel Go. Zowel fysieke robots als computers worden steeds beter en zelfstandiger.

De actieradius van robots beperkt zich allang niet meer tot een gecontroleerde omgeving zoals een fabriek. Bij zelfrijdende auto's en zorgrobots moeten we rekening gaan houden met computergestuurde machines die tussen en met mensen werken.

Hoe veilig is dat? Je wilt niet dat je zorgrobot per ongeluk je zoontje van de trap gooit omdat zijn software hem opdraagt exact om drie uur de was te doen. Een huiselijke machine *moet* begrijpen dat het welbevinden van je zoontje belangrijker is dan de was. Zelfs in een voor mensen eenvoudig huishouden zijn niet alle mogelijke scenario's die een robot tegen kan komen te voorspellen. Robots zelf moeten de juiste afweging kunnen maken. In de voor machines chaotische wereld waarin wij leven, kunnen robots alleen functioneren als ze, net als de meeste mensen, routinematig en onmiddellijk ethische keuzes maken tijdens het uitvoeren van hun taken. Laat staan als het gaat om 'killer robots' die zelfstandig hun doelen kunnen selecteren en elimineren. In augustus riepen de oprichters van ruim honderd bedrijven wereldwijd op het terrein van robotica en kunstmatige intelligentie daarom op om de ontwikkeling van 'killer robots' te stoppen. Onder hen was Elon Musk, oprichter van autoproducent Tesla.

Het is een oude kwestie die actueel is geworden. Over ethische normen en waarden voor robots dacht de sciencefictionschrijver Isaac Asimov zeventig jaar geleden al na. In zijn boek *Ik, Robot* beschrijft hij de drie wetten van de robotica

die hij bedacht heeft om te voorkomen dat ver ontwikkelde robots zich tegen de mensheid keren. Het is dan ook toepasselijk dat juist dit boek in november gratis uitgedeeld door de openbare bibliotheken in het kader van Nederland Leest.

1 item toegevoegd

Er moet iets gebeuren, want als de maatschappij zich uit angst verzet tegen verdere ontwikkeling van robotica en kunstmatige intelligentie kan deze veelbelovende technologische ontwikkeling wel eens tot een abrupt einde komen. „Dat zou zonde zijn", zegt Alan Winfield, hoogleraar robot-ethica van de University of the West of England in Bristol. „Robots kunnen ons leven makkelijker en veiliger maken."

Azimovs regels

Hoe voorkom je dat intelligente robots zich tegen de mensheid keren? Hier dacht Amerikaanse schrijver Isaac Asimov in 1942 al over na. In zijn sciencefiction boeken loste hij dit probleem op door elke robot drie gedragsregels mee te geven:

Een robot mag een mens geen letsel toebrengen of door niet te handelen toestaan dat een mens letsel oploopt.

Een robot moet de bevelen uitvoeren die hem door mensen gegeven worden, behalve als die opdrachten in strijd zijn met de Eerste Wet.

Een robot moet zijn eigen bestaan beschermen, voor zover die bescherming niet in strijd is met de Eerste of Tweede Wet.

Later kwam daar de Nulde Wet bij:

Een robot mag geen schade toebrengen aan de mensheid, of toelaten dat de mensheid schade toegebracht wordt door zijn nalatigheid.

Sympathieke robots

De wetten van Asimov zijn bedacht als literair instrument, maar niettemin worden ze vaak aangehaald door robotici. Er zijn zelfs onderzoekers die werken aan robots die op basis van juist deze wetten morele beslissingen nemen. Ook Alan

Winfield experimenteert bij het Bristol Robotics Laboratory in Engeland met robots die zelfstandig ethische keuzes kunnen maken. „Een paar jaar geleden dacht ik dat het onmogelijk was om een ethische robot te bouwen”, vertelt hij, „maar inmiddels ben ik 180 graden gedraaid.” De robots in het lab in Bristol waar Winfield werkt, nemen nu beslissingen op basis van simpele ethische regels die vergelijkbaar zijn met die van Asimov.

Item toegevoegd

Filmpjes op Winfield’s website tonen schattige Nao robots, van het Franse bedrijf Aldebaran Robotics, die voor een ethisch dilemma staan: zo snel mogelijk naar hun “doel” lopen of afwijken naar rechts en voorkomen dat een mens (gerepresenteerd door een andere robot, genaamd H-robot) in een gat valt. Het “gat” is een op de vloer getekend vierkant, om schade aan de robots te voorkomen. Je ziet de robot, met het formaat van een kleuter, zonder te twijfelen van zijn pad afwijken om de H-robot voor het gevaar te behoeden.

> Over ethiek voor robots dacht schrijver Asimov 70 jaar geleden al na

Voor mensen lijkt dit gedrag simpel, maar er gaat een complexe actie aan vooraf. „De robot moet de cognitieve vaardigheid hebben om zich voor te stellen wat er gebeurt als de ander doorloopt”, zegt Winfield.

Daarnaast moet de robot voorspellen wat hij zelf kan doen en welke van die opties ervoor zorgt dat de H-robot niet in het gat valt. Voor het nabootsen van die cognitieve vaardigheid wordt contact gelegd met een computer, die razendsnel alle mogelijke scenario’s nagaat.

Het nemen van de beslissing op basis van alle mogelijk scenario’s ziet er in computercode zo uit, volkomen Asimoviaans:

IF for all robot actions, the human is equally safe

THEN (* default safe actions *)

output safe actions

1 Item toegevoegd

ELSE (* ethical action *)

output action(s) for least unsafe human outcome(s)

De situatie van Winfield is relatief simpel: de robot hoefde niet te herkennen wat een mens is en wat gevaar. En zijn missie was duidelijk: de H-robot mag niet in het verboden vierkant lopen. Dat is niet te vergelijken met de complexe, manier waarop mensen zulke keuzes maken.

Maar toch: „Met de Nao robots hebben we aangetoond dat het mogelijk is om een ethische robot te bouwen in de veilige en simpele omgeving in het lab", vertelt Winfield.

InMoov, Open Source, 3D-Printed Robot, 2016.

Ethische dilemma's

Het begin is er. En in complexere situaties blijken Winfields robots zich zelfs verrassend menselijk te gedragen.

Want in een tweede experiment stelde Winfield de ethische robot, zonder de programmeercode aan te passen, bloot aan een nieuw dilemma: twee "mensen" bewegen zich in de richting van een gat. De robot kan ze niet allebei redden. Welke keus maakt de arme robot wanneer hij voor dit duivelse dilemma komt te staan?

In 16 van de 33 gevallen lukt het hem om één "mens" te redden en drie keer had hij het geluk zo snel te reageren dat hij beiden kon behoeden voor het gevaarlijke gat. Maar 14 keer raakte hij zó van slag dat hij niet op tijd een keus kon maken, waardoor beiden in het gat vielen.

„Dat komt doordat de robot telkens opnieuw afweegt wat de beste keus is", vertelt Winfield. „Terwijl hij onderweg is naar de ene, ziet hij dan de andere en begint opnieuw na te denken. Dan kiest hij ervoor toch die andere te redden." Doordat

het robotje te lang aarzelend tussen de twee heen en weer liep, redde hij niemand.

1 Item toegevoegd

Winfield: „Een mens zou de keus maken om één van de twee te redden en vervolgens bij die beslissing blijven. De robot bleek dat niet te doen.”

Om de beslissing te nemen, moest de robot zich voorstellen welke acties hij kon ondernemen. Hij heeft dus een zelfbeeld. Maar dat betekent niet dat hij zelfbewust is. Winfield: „Het zijn ethische zombies. Ze voeren de ethische regels die wij geprogrammeerd hebben uit, maar denken daar niet zelf over na.” De volgende stap is een robot die bewust een ethische keus maakt en achteraf kan verantwoorden waarom hij die keus maakt. Die stap is volgens Winfield nog een flinke sprong verwijderd van de ethische zombies.

Leer mij het zelf te doen

De robots in Bristol doorliepen keurig de ingeprogrammeerde stappen. Daarin zijn de ethische regels precies gedefinieerd. Daardoor kunnen de robots alleen omgaan met situaties die van te voren uitgedacht zijn.

Mensen leren op een andere manier ethische beslissingen te nemen. Van kleins af aan leren we welke keuzes goed of fout zijn en op basis daarvan maken we een keus in een nieuwe situatie. Robots kun je ook op deze manier laten leren, met zogenoemde *machine learning* algoritmes; je geeft ze dan een aantal ethische problemen en vertelt welke oplossingen daar volgens jou bij horen. Die informatie categoriseren ze en op basis daarvan besluiten ze hoe ze in een nieuwe situatie reageren.

Het nadeel daarvan is dat het mis kan gaan als de leermeester van de robot geen goede bedoelingen heeft.

Een voorbeeld waarbij het mis ging is Tay van Microsoft. Begin 2016 werd deze chatbot losgelaten op Twitter. Het idee was dat Tay zou leren van andere twitteraars. En dat hebben we geweten. Na nog geen 24 uur verkondigde de chatbot dat Hitler niets verkeerd gedaan had en dat Bush achter 9/11 zat. Kwaadaardige online gebruikers (*trolls*) hadden duidelijk hun best gedaan om Tay’s wereldbeeld behoorlijk te verpesten.

Hoogleraar informatica Michael Anderson en zijn vrouw, hoogleraar filosofie Susan Anderson, beiden van de University of Hartford in de Verenigde Staten, pakken het anders aan. Zij werken aan zorgrobots, die ouderen bijstaan zodat ze langer zelfstandig kunnen blijven wonen.

Het computerprogramma dat hun Nao robots aanstuurt gaat in de leer bij genuanceerde ethici. Ze leggen de robot een aantal kwesties voor die duidelijke, ethisch verantwoorde oplossingen hebben en vertellen de robot wat die oplossingen zijn en waarom dat zo is. Aan de hand daarvan vormt de machine zelf ethische principes, die hij in nieuwe situaties toe kan passen.

„De robot leert volgens een oude *machine learning* techniek, waarvoor geen grote hoeveelheden data nodig zijn. Enkele voorbeelden zijn genoeg", vertelt Michael Anderson.

„De basis van het ethische computerprogramma zijn de ideeën die de filosoof David Ross had over ethiek", vertelt Susan Anderson. Volgens Ross gaat het in de ethiek om het balanceren van verschillende doelen. Daar bestaat geen duidelijke rangorde voor; soms is een bepaald doel belangrijker dan een ander.

Susan Anderson: „We hebben een programma gemaakt dat aan elke actie een getal hangt dat aangeeft hoe belangrijk het is om iets te bereiken of juist te voorkomen." Bij de zorgrobot gaat het bijvoorbeeld om hoe belangrijk het is dat iemand zijn medicijnen op een bepaald tijdstip inneemt. Het is daarbij ook belangrijk dat de patiënt zijn of haar zelfbeschikkingsrecht houdt en de medicatie kan weigeren of uitstellen.

De robot kijkt naar de totale waarde van elke actie die hij kan ondernemen en kiest of hij de patiënt eerst een tijdje met rust laat en later helpt herinneren, of dat het nodig is om aan te dringen of zelfs contact op te nemen met een arts.

„Dankzij de hulp van Vincent Berenz van het Max Planck Instituut in Duitsland zijn we nu zover dat we willen gaan testen hoe de robot samenwerkt met mensen", vertelt Michael Anderson. „Maar daarvoor moeten we eerst mensen vinden die gek genoeg zijn om onze robots te willen testen", voegt Susan Anderson daar aan toe.

Hoe maken mensen keuzes?

[1] Item toegevoegd

Uiteindelijk zullen machines dilemma's oplossen die ingewikkelder zijn dan gaten in de vloer en medicijnregulatie.

Welke keus moet een zelfrijdende auto bijvoorbeeld maken als er ineens een zesjarig meisje de weg op rent en de auto alleen maar naar rechts kan uitwijken, waar een tachtigjarige man loopt?

En wat moet een medische robot doen die enkel genoeg tijd heeft om één gewonde soldaat te redden, terwijl er twee op het slagveld liggen?

De eerste stappen zijn gezet met de minimaal ethische robots, zoals die van Alan Winfield en Susan en Michael Anderson. Die maken in veel situaties al keuzes waar mensen het mee eens zijn. Maar het zal nog even duren voordat robots in staat zijn goed te reageren op complexe ethische dilemma's en hun acties achteraf kunnen verklaren, voorspelt Winfield.

Bang voor een robotleger dat de mensheid omlegt zijn de onderzoekers in elk geval niet. „De huidige robots worden door de ontwerpers erg beperkt in wat ze kunnen", zegt Susan Anderson. „Ze zijn niet in staat zelf initiatief te nemen. Dat is jammer, want ethiek gaat niet alleen over wat je niet moet doen, maar ook over wat je wel moet doen."

Winfield: „Vergeet niet dat robots en kunstmatige intelligentie ons leven makkelijker en veiliger kunnen maken. We zouden optimistisch moeten zijn over de snelle ontwikkelingen in plaats van bang."