AI moet geen pijn doen

Kunstmatige intelligentie Als geautomatiseerde systemen steeds

slimmer worden, blijven ze dan doen wat we willen?

Bennie Mols 18 september 2020 om 13:47

Leestijd 7 minuten

Tesla's die zelfstandig van rijbaan wisselen en inhalen op de snelweg. Medische diagnosesystemen die vertellen wat een patiënt mankeert. Software die aan de bel trekt bij frauduleuze creditcardbetalingen. Kunstmatige intelligentie (AI) maakt technische systemen steeds autonomer. AI-systemen kunnen zelf nieuwe dingen leren en zijn niet gebaseerd op expliciet uitgeschreven en goed begrepen regels, zoals bij klassieke automatisering.

Idealiter verbeteren door het zelflerende karakter de prestaties, maar dan moet het ontwerp wel goed zijn. En dat is nou net geen sinecure. Is het ontwerp onvoldoende doordacht, dan liggen fouten op de loer, zoals de Tesla die in maart 2018 in Californië tegen een betonnen wegafscheiding knalde. Het officiële onderzoeksrapport gaf de schuld aan technisch falen van zowel de automatische piloot als het automatische remsysteem.

Ook discriminerende of onuitlegbare beslissingen zijn een reëel gevaar. En zelfs de meest geavanceerde AI-systemen hebben meestal weinig gezond verstand en zijn eenvoudig te foppen. Plak een psychedelische sticker naast een banaan en de slimme computer ziet de banaan voor een broodrooster aan.

Privacy en gelijke behandeling

Als geautomatiseerde systemen steeds slimmer worden, steeds meer op eigen houtje leren, hoe kunnen we er dan voor zorgen dat ze ook blijven doen wat ze moeten doen, en dat ze zich blijven houden aan onderliggende menselijke waarden als privacy en gelijke behandeling? Hoeveel controle kan en moet een mens nog hebben over steeds slimmere computersystemen?

Als er één domein is waarin automatisering ver is doorgevoerd, minutieus wordt gecontroleerd en waarin ook nog eens veel mensenlevens op het spel staan, dan is het de luchtvaart. Tot nu toe gaat het daar om klassieke automatisering. Maar om te begrijpen hoe de mens controle kan houden over zelflerende AI-systemen, die zelfs voor de programmeurs deels ondoorgrondelijk zijn, is het goed om eerst de automatisering in de luchtvaart onder de loep te nemen.

Dankzij verregaande automatisering is tussen 1959 en 2018 wereldwijd het aantal vliegtuigongelukken per miljoen vertrekkende vluchten met een factor vijftig gedaald. Toch gaat het ook in de luchtvaart soms faliekant mis. In 2018 en 2019 verongelukten toestellen van Lion Air en Ethiopean Airlines, allebei van het type Boeing 737 MAX.

Hoofdoorzaak van die ongelukken was een haperend automatisch systeem dat het vliegtuig stabiel in de lucht moet houden. Bij het eerste ongeluk wisten de piloten niet eens van de aanwezigheid van het systeem. Bij het tweede ongeluk kreeg de bemanning het toestel

niet meer onder controle nadat het systeem zich onterecht inschakelde na een signaal van een kapotte sensor.

Wat kunnen we uit de automatisering in de burgerluchtvaart leren over menselijke controle in kunstmatig intelligente systemen?

„In de burgerluchtvaart wordt klassieke, regelgebaseerde automatisering gebruikt en geen zelflerende AI. Dat werkt heel goed en heeft het bijgedragen aan een zeer veilig, efficiënt en betaalbaar vervoersmiddel", zegt Max Mulder, hoogleraar mens-machinesystemen in de luchtvaart aan de TU Delft. „Dat de menselijke piloot nu al vervangen kan worden door kunstmatige intelligentie, zoals je vaak hoort beweren, is aperte onzin."

Zijn we veiliger af zonder piloten? Nee, zeker niet

Max Mulder — hoogleraar

Mulder ergert zich aan berichten als dat 70 procent van de vliegtuigongelukken te wijten is aan menselijke fouten en dat dat een argument zou zijn om mensen door AI te vervangen. „Wat er dan niet bij wordt verteld is dat er gemiddeld 2.500 maal meer problemen met de klassieke automatisering zijn die door de piloten worden opgelost dan er ongelukken gebeuren. En elk van die problemen, hoe klein ook, had tot een ongeluk kunnen leiden. Zijn we veiliger af zonder piloten? Nee, zeker niet. En hetzelfde geldt voor zelfrijdende auto's."

De burgerluchtvaart is extreem terughoudend om ook maar enige vorm van zelflerende AI in te zetten. „Het enige in een passagiersvliegtuig dat adaptief mag en moet zijn, is de piloot”, zegt Mulder. „Alle software en regelsystemen zijn volledig deterministisch. Stop je een vliegtuig vol zelflerende AI, dan zul je onnoemelijk veel mogelijke praktijksituaties van te voren moeten bedenken, testen en certificeren. Dat is onbetaalbaar. Een verkeersvliegtuig dat in alle mogelijke omstandigheden veilig en zonder menselijke piloten kan vliegen, ligt ver weg in de toekomst.”

Boeing heeft een naam opgebouwd als een uitermate degelijk ingenieursbedrijf. Dat het toch misging met de 737 MAX, maakt Holger Hoos, hoogleraar machine learning aan de Universiteit Leiden, bezorgd over het inzetten van AI-systemen die over mensen beslissen. „Het falende automatische stabiliteitssysteem in een Boeing-vliegtuig is qua software een stuk eenvoudiger dan veel AI-systemen. En als Boeing-ingenieurs hun klassieke automatische systeem al onvoldoende begrepen, dan kan het met AI-systemen in principe nog veel erger worden.”

Toch ligt in het toepassen van kunstmatige intelligentie volgens Hoos juist ook een oplossing. „De enige manier waarop we complexe systemen veilig en goed kunnen ontwerpen, is door AI als instrument te laten gebruiken door slimme mensen die weten hoe ze verstandig met AI moeten omgaan.”

Belastingfraude opsporen

Hij geeft als voorbeeld de fout die in 1994 werd gevonden in de Intel Pentium-chip. In sommige gevallen bleek het delen van twee

getallen tot kleine onnauwkeurigheden te leiden. Tegenwoordig gebruiken hardware-ontwerpers AI-technieken om menselijke experts aan te vullen en ontwerpfouten te voorkomen. Sinds de Pentium-bug hebben we niet meer zulke hardware-fouten gekregen. „Hetzelfde zien we steeds meer gebeuren in het bouwen van software", zegt Hoos. „AI kan en moet juist een deel van de oplossing zijn om complexe computersystemen te ontwerpen."

Een van de nieuwe onderzoekspaden die Hoos en andere AI-onderzoekers exploreren, is het bouwen van zelfmonitorende AI. Hoos illustreert het idee aan de hand van AI die wordt gebruikt om belastingfraude op te sporen. „Zo'n AI-systeem is getraind met data uit het verleden. Maar in de loop van de tijd hoeven de trainingsdata niet meer representatief te zijn voor de actuele situatie, bijvoorbeeld omdat de economie of de bevolkingssamenstelling is veranderd. In dat geval geeft zelfmonitorende AI een seintje: 'De nieuwe data zien er wezenlijk anders uit dan mijn trainingsdata. Ik ga geen voorspellingen meer doen, want dan begeef ik me buiten mijn expertise.'"

Zelfmonitorende AI functioneert zo als een ingebouwd waarschuwingslicht. Staat het licht op groen, dan functioneert het AI-systeem binnen zijn beperkingen, bij oranje bereikt het de grenzen van zijn mogelijkheden, en bij rood is het die grens overschreden. Dan moet het systeem de controle helemaal aan de mens overlaten, of – in het geval dat mensen bijvoorbeeld niet snel genoeg kunnen ingrijpen – andere mechanismen inschakelen die de

situatie oplossen. Hoos verwacht dat zulke systemen over een jaar of vijf in praktijk komen.

Stofzuigen op een feestje

Bij klassieke technische systemen kan de mens vaak met een druk op de knop de controle overnemen van het systeem. Voor AI-systemen lijkt het meer op de controle die een ruiter over haar paard heeft. Het paard handelt deels autonoom, dankzij zijn eigen intelligentie, en wordt deels geleid door signalen van de ruiter, via de teugels en de benen bijvoorbeeld. Op dezelfde manier handelt een AI-systeem deels autonoom, maar staat het deels nog onder controle van gebruikers, ontwerpers en bouwers.

Volgens Catholijn Jonker, hoogleraar interactieve intelligentie aan zowel de TU Delft als de Universiteit Leiden, worden AI-systemen te vaak ontworpen om de mens overbodig te maken, terwijl de toekomst volgens haar veel meer ligt in slimme machines die met mensen samenwerken om zo iets voor elkaar te krijgen wat noch de machine noch de mens alleen voor elkaar krijgt. „Mijn motto is: met machines meer mens”, aldus Jonker.

Jonker onderzoekt onder andere hoe mens en machine elkaar beter kunnen begrijpen. Daarvoor moeten ze elkaars beperkingen en mogelijkheden kennen. Jonker: „Neem de huishoudrobot van de toekomst. Stel, je geeft een feestje en ineens komt je robot binnen en begint te stofzuigen. Dat wil je niet. Je wilt ook niet een discussie met hem aangaan waar alle gasten bij staan. Je wilt dat hij leert dat hij beter op een rustig moment kan gaan stofzuigen. Daarvoor zal de robot onze taal en onze denkwereld moeten begrijpen.”

Illustratie Timber Sommerdijk

Zelf werkt ze al jaren aan een automatische pocketonderhandelaar, een AI-systeem dat mensen helpt bij het onderhandelen over bijvoorbeeld de aankoop van een huis. „Ook hier moeten mens en machine elkaar begrijpen”, zegt Jonker. „Ons AI-systeem moet leren kijken naar wat voor type onderhandelaar de gebruiker is, bijvoorbeeld voorzichtig of juist agressief. Daarnaast gaat onderhandelen over meer dan prijs alleen. Kom je de verkoper later nog eens tegen of niet? Dat maakt uit in de onderhandeling. En welke andere waarden dan prijs alleen vindt de gebruiker belangrijk: of de verkoper aardig is of niet, maar ook aspecten rondom het huis: wie betaalt de controle over de kwaliteit van de grond waarop het huis staat?”

Het ideaal dat Jonker voor ogen staat, lijkt meer op de samenwerking tussen mens en hulphond dan tussen ruiter en paard, vertelt ze. „Bij een paard kun je aan de teugels trekken wat je wilt, maar als het paard op hol slaat, heb je een probleem. Uiteindelijk gaat het ook om wederzijds vertrouwen, respect en begrip van elkaar. De hond is zich door evolutie meer en meer gaan richten op de mens. Neem de schaapsherder en zijn hond. Ze hebben elk een eigen taak, maar zijn perfect op elkaar afgestemd. Zo zou het met mens en machine ook moeten zijn."

Politieke beïnvloeding

Om te garanderen dat mens en machine samen zich ook houden aan onderliggende waarden als privacy en gelijkheid, is het onvoldoende om alleen te kijken naar hoe mens en machine samenwerken. Normen en waarden moeten al bij het ontwerp van een AI-systeem worden meegenomen, vindt Filippo Santoni de Sio, universitair docent ethiek en filosofie van technologie van de TU Delft. Samen met Delftse collega's heeft hij het concept van 'betekenisvolle menselijke controle' ontwikkeld.

Hij illustreert 'betekenisvolle menselijke controle' aan de hand van het Cambridge Analytica-schandaal bij Facebook uit 2018. „Het ging om een externe app waar Facebook-gebruikers wel of niet gebruik van konden maken. Maar om er gebruik van te maken, moesten ze akkoord gaan met het delen van hun data met een derde partij. Ze hadden echter geen idee dat hun data gebruikt konden worden voor politieke beïnvloeding van de Amerikaanse verkiezingen en het Brexit-referendum. Je kunt zeggen dat Facebook-gebruikers

controle hadden over hun data, maar dan toch alleen oppervlakkige controle. Wat ze zouden moeten hebben is betekenisvolle controle: hun doelen, intenties en waarden moeten weerspiegeld zijn in het systeem dat ze gebruiken."

We moeten AI-systemen volgens Santoni de Sio altijd beschouwen in een bredere context: het gaat niet alleen om de rol van gebruikers, maar ook om de macht van bedrijven, om wet- en regelgevers, om de verantwoordelijkheid van ontwerpers, en om de rechten van alle mensen die geraakt kunnen worden door de technologie. Hij vindt de burgerluchtvaart een goed voorbeeld van succesvolle betekenisvolle controle over automatisering: „De luchtvaart combineert een hoge mate van automatisering met een heel complexe sociale en maatschappelijke omgeving. In de luchtvaart worden veiligheid en verantwoordelijkheid niet als een last gezien, maar als integraal onderdeel van de professie. Zo zou het ook moeten zijn met AI-systemen in de toekomst."

Het meenemen van normen en waarden in het ontwerpen van slimme systemen zou innovatie hinderen, krijgt Santoni de Sio af en toe te horen. Zijn reactie: „Zeggen dat het inbouwen van betekenisvolle controle te kostbaar of te inefficiënt is, is hetzelfde als zeggen dat je geen democratie wilt omdat dat te inefficiënt werkt. We moeten ophouden ethiek als een obstakel voor technologie te zien. Het is juist een voorwaarde om onze normen en waarden democratisch te verankeren in technologie."