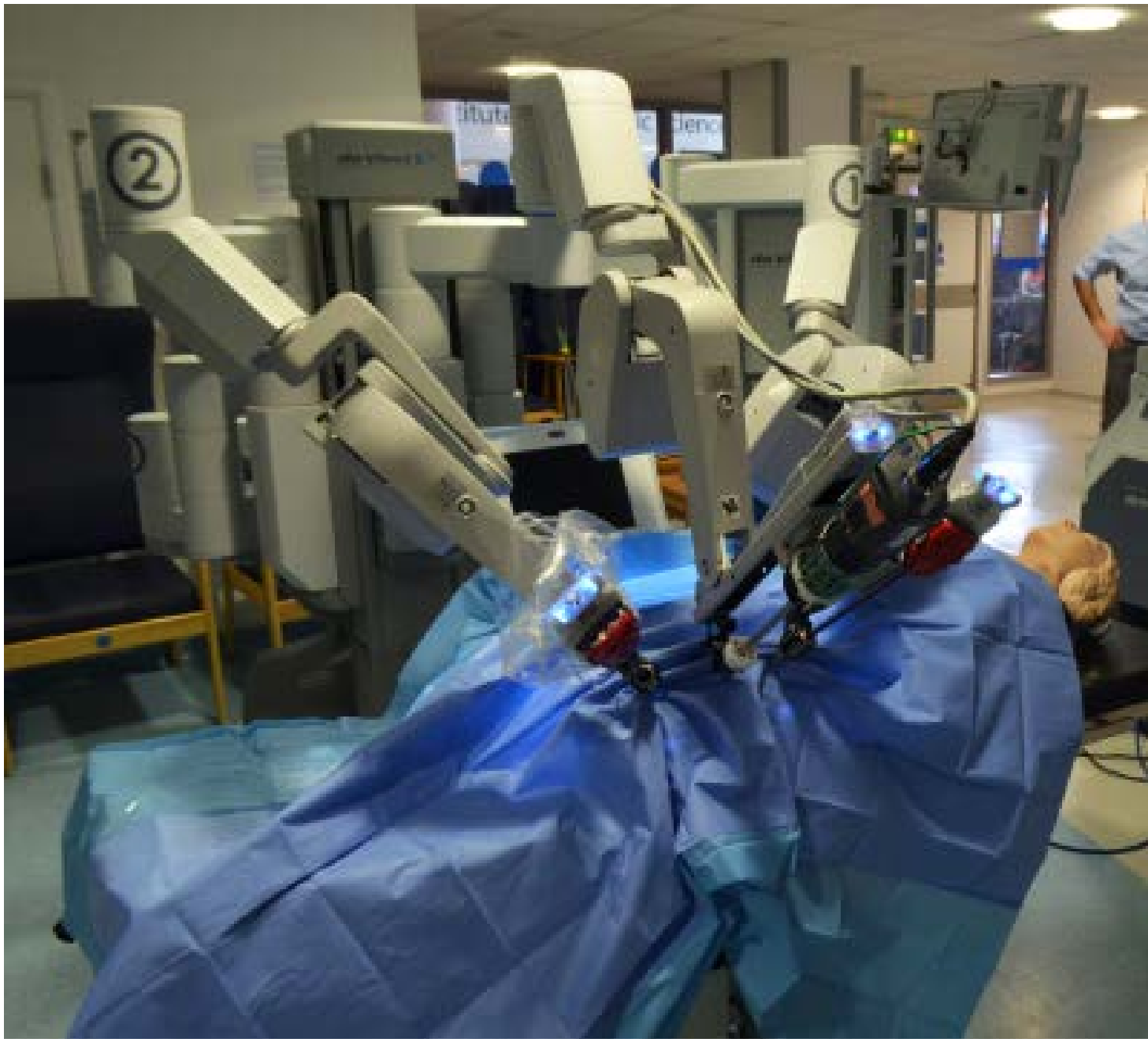
Kunstmatige intelligentie begrijpen

Zoeken naar het waarom



De Da Vinci operatierobot in Cambridge

In sommige taken overtreft kunstmatige intelligentie inmiddels de mens. Classificatie is daar een voorbeeld van. Zeker bij toepassing van deep learning. Maar deep learning is niet rule-based. En dat maakt de resultaten onovertroffen, maar niet interpretabel. Een probleem?

tekst Roy de Kleijn

Jarenlang veiligheidsonderzoek laat zien waar er dingen mis kunnen gaan bij menselijk handelen. Vakgebieden als cognitieve ergonomie en mens-computerinteractie onderzoeken de variabelen die van invloed zijn op de beslissingen die mensen nemen en hun relatie tot het maken van fouten. Maar waar mensen fouten maken, zijn het niet slechts de mensenhanden die schade veroorzaken. Wanneer mensen grote machines besturen kunnen de gevolgen van deze fouten nog groter worden. Want die machines hebben vaak meer kracht of zijn groter, waardoor ze meer schade kunnen aanrichten.

Het handelen van deze apparaten valt onder de verantwoordelijkheid van de operator, hoe eng dit ook is. Die bepaalt welke handelingen en bewegingen het apparaat uitvoert. Een voorbeeld is de Da Vinci robot, gebruikt voor laparoscopische chirurgie. De Da Vinci robot wordt volledig bestuurd door mensen. Bij een fout is het vaak duidelijk wie die heeft gemaakt: de operator. Maar met de vooruitgang in kunstmatige intelligentie (KI) is het niet langer vanzelfsprekend dat dit soort apparaten domweg door mensen wordt bestuurd. Wat nu als robots zelf beslissingen nemen? Naarmate kunstmatige intelligentie slimmer wordt, is het onvermijdelijk dat een deel van de autonomie van de operator verplaatst naar de KI zelf. Dat zou op het eerste gezicht geen probleem moeten zijn als deze kunstmatige intelligentie betere beslissingen neemt dan een menselijke operator.

Superioriteit KI

Classificatie is één van de taken waar moderne kunstmatige intelligentie de mens vaak overtreft. Een goed voorbeeld is het classificeren van plekjes op de huid, om te bepalen of het gaat om een onschuldige moedervlek of om een melanoom (een vorm van huidkanker). Wanneer een dermatoloog zo'n plekje bekijkt, is hij het die deze beslissing moet nemen. KI-oplossingen kunnen deze taak van hem overnemen. De eerste generatie KI-technieken voor dit soort classificatie was gebaseerd op regels die worden toegepast op een afbeelding van de huid. Die regels zijn op hun beurt gebaseerd op de regels die dermatologen gebruiken om een plekje op de huid te beoordelen. Bevat de afbeelding asymmetrische gebieden of »

Bij deep learning-technieken zijn prestaties gebaseerd op niet te interpreteren factoren

Een zelfrijdende auto die een voetganger doodrijdt en dan niet kunnen uitleggen waarom? Dat ligt gevoelig

gebieden met een grillige rand? Dan gaat het om een melanoom. Bij classificatie als melanoom is het vervolgens mogelijk om te bekijken waarom het algoritme deze beslissing heeft genomen.

De nieuwste generatie learning-based classificatie-algoritmen presteren veel beter dan deze rule-based technieken. Maar de algoritmen volgen daarbij een andere benadering. Technieken zoals *deep learning* worden vaak toegepast in de vorm van *supervised learning*. Hierbij krijgt een systeem correct geclassificeerde voorbeelden gepresenteerd, waarna het nieuwe gevallen ook kan classificeren. Maar een deep learner die wordt getraind om melanomen te herkennen, krijgt geen regels ingebouwd. In plaats daarvan krijgt het 10.000 afbeeldingen van vlekken of bulten op de huid te zien. Daarvan zijn 5.000 een melanoom en 5.000 niet. De computer krijgt in deze leerfase te horen wat het juiste antwoord is – de afbeeldingen zijn bijvoorbeeld al gediagnosticeerd door een dermatoloog. "Kijk, dit is een afbeelding van een melanoom. En dit is er een van een onschuldige moedervlek." Na genoeg voorbeelden kan dit systeem bij het zien van een nieuwe afbeelding zeer nauwkeurig bepalen of er een melanoom op de afbeelding staat of niet. Daarin is het nauwkeuriger dan ervaren dermatologen.
Dit systeem is niet langer rule-based in de zin zoals wij dat verwachten. Oftewel: er zijn geen duidelijke regels die uitleggen waarom een afbeelding classificeert als melanoom. Er is sprake van een zogenoemde black box. Maar is dat een probleem? Het algoritme presteert beter dan ervaren dermatologen en dat is wat we willen, toch?

Interpretabiliteit

Een hoge sensitiviteit en specificiteit is zeker wat we zoeken in dit soort technieken. Maar wat kunnen we doen met de onvermijdelijke gevallen waarin een verkeerde diagnose wordt gesteld?
Bij de rule-based technieken kunnen we onderzoeken waarom dat is gebeurd. Was de melanoom misschien symmetrisch van vorm, waardoor die over het hoofd is gezien? Komt dit vaker voor, dan zouden we kunnen overwegen deze regel weg te laten of aan te passen. We zouden eventueel zelfs de opleiding voor dermatologen kunnen aanpassen. Technieken zoals deep learning bieden deze mogelijkheid niet. En hoewel ze beter presteren dan de meeste menselijke experts, maakt ook een deep learner fouten. De gemaakte fouten (maar ook de juiste classificaties) zijn gebaseerd op factoren die we niet kunnen interpreteren. Dit biedt mogelijkheden tot generalisatie: een melanoom dat toevallig symmetrisch is kan correct als zodanig worden geclassificeerd. Maar hierdoor ontstaat ook een groot nadeel: we kunnen het probleem niet makkelijk herkennen en oplossen.

Op korte termijn moeten we daarom een belangrijke afweging maken. Geven we de voorkeur aan technieken die beter presteren, maar niet interpretabel zijn? Of nemen we genoegen met een slechtere prestatie die valt uit te leggen?

Niet-interpretabele KI

Het voorbeeld van dermatologische classificatie is slechts één van de vele gebieden waarop deep learning-achtige technieken een plaats hebben gekregen. Zelfrijdende auto's en militaire drones werken met soortgelijke technieken en lijden aan dezelfde ondoorzichtigheid. U kunt zich voorstellen dat mensen moeite hebben met een zelfrijdende auto die een voetganger doodrijdt, zonder dat uit te leggen valt waarom dat is gebeurd. Zelfs als dit met een zelfrijdende auto veel minder vaak gebeurt dan met een menselijke bestuurder.
Een ander voorbeeld is *preventative policing*: het gebruik van kunstmatige intelligentie door politiekorpsen om te voorspellen waar en vooral door wie misdaden gepleegd gaan worden. En welke kenmerken de verwachte dader heeft. Als een KI daarin een beslissing neemt, willen we kunnen uitleggen waarom. Een oppervlakkige uitleg "dat het systeem het nu eenmaal zo heeft geleerd" is dan niet genoeg.

Wat doen we eraan?

Ik probeer geen rampzalige toekomst te schetsen waarin we slachtoffer worden van ingewikkelde, ondoorzichtige, bijna Kafkaëske *black boxes* die beslissingen voor ons nemen zonder dat duidelijk is waarom. Integendeel, het vakgebied van *explainable* AI dat zich bezighoudt met de interpretabiliteit van dit soort algoritmen is razend populair. We spreken van een interpretabele KI als inzichtelijk is hoe een algoritme tot een resultaat is gekomen. Verschillende onderzoeksgroepen houden zich bezig met het ontwikkelen van technieken om inzicht te krijgen in hoe deep learners tot een classificatie komen, met toenemend succes. Wel is het belangrijk om de discussie over interpretabiliteit alvast in gang te zetten. Al passen we de genoemde technieken nog niet op grote schaal toe, hun mogelijke gevolgen zijn groot.

Hoe kunnen we dit probleem aanpakken? Dat is geen eenvoudige taak. Maar een suggestie is afgekeken van een sector waar veiligheid hoog in het vaandel staat. Totdat we goede manieren hebben om het gedrag van deze technieken te interpreteren, is het belangrijk dat we het gedrag van geautomatiseerde systemen goed observeren en registreren. Denk aan de black box in de luchtvaart, die nauwkeurig bijhoudt welke invoer naar het systeem gaat, wat de resulterende staat van het systeem is en wat het uiteindelijke resultaat. In het geval van de melanoomdetectie dus het opslaan van de afbeelding, de staat van het leeralgoritme en de classificatie moedervlek/melanoom. De opgeslagen informatie kunnen we dan in de toekomst gebruiken om te ontdekken wat er precies is misgegaan en waar we onze KI-technieken moeten bijschaven of bijleren. We hebben dus nog een black box nodig, maar dan een goede. «

Roy de Kleijn is universitair docent in cognitieve psychologie en kunstmatige intelligentie aan de Universiteit Leiden.