Artificiële intelligentie en risicotaxatie

Drie kernvragen voor strafrechtjuristen

Johannes Bijlsma, Floris Bex & Gerben Meynen[1]

De snelle ontwikkeling van artificiële intelligentie werpt een aantal prangende vragen op voor strafrecht-juristen en forensisch gedragsdeskundigen. Risicotaxatie met behulp van AI heeft de potentie om een relatief betrouwbaar middel te worden om inschattingen van recidivegevaar te maken. Een belangrijk vraagpunt is tot hoeveel vals-positieven (*precision*) en vals-negatieven (*recall*) de inschatting leidt. Door de 'kosten' van verschillende fouten te bepalen en in het algoritme te verwerken, kan het belang van de maatschappelijke veiligheid worden afgewogen tegen het belang dat burgers niet ten onrechte door strafrechtelijk ingrijpen moeten worden getroffen. Een tweede vraagpunt is hoe omgegaan moet worden met *biases*. Er zijn technische manieren om vooroordelen in de algoritmes tegen te gaan, maar daarbij moet wel de normatieve vraag onder ogen worden gezien hoe dat op een rechtvaardige wijze kan gebeuren. Ten derde zijn AI-algoritmes relatief ondoorzichtig, waardoor zij niet altijd mogelijkheden aanwijzen voor interventies die gericht zijn op rehabilitatie van de veroordeelde.

1. Inleiding

Met regelmaat wordt ervoor gepleit om vaker en meer geavanceerde risicotaxatie-instrumenten te gebruiken bij het nemen van strafrechtelijke beslissingen waarvoor een inschatting van recidivegevaar van de verdachte van belang is, bijvoorbeeld bij de oplegging van tbs.[2] Risicotaxatie-instrumenten voor recidive hebben echter (voorals nog) aanzienlijke beperkingen. Een algemeen kenmerk van risicotaxatie-instrumenten is bijvoorbeeld dat ze vrij betrouwbaar kunnen aangeven wie niet gevaarlijk is, maar van degenen die als 'hoog risico' worden aangemerkt recidiveren de meesten niet.[3]

Door de recente ontwikkelingen (en de *hype*) rond artificiële intelligentie ('AI') toont ook het juridisch veld steeds meer interesse in de mogelijkheden die AI biedt,[4] getuige de vele stukken in de populaire en wetenschappelijke literatuur over 'robotrechters' en algoritmes in de rechtspraak.[5] Een vraag die rijst is: zou AI niet ook een bij-

Auteurs

1. Mr. dr. J. Bijlsma is universitair docent strafrecht, Willem Pompe Instituur voor Strafrechtswetenschappen en Utrecht Centre for Accountability and Liability Law (UCALL), Universiteit Utrecht. Prof. dr. G. Meynen is hoogleraar forensische psychiatrie, Willem Pompe Instituut voor Strafrechtswetenschappen en UCALL, Universiteit Utrecht, tevens bijzonder hoogleraar ethiek en psychiatrie, Geesteswetenschappen, Vrije Universiteit Amsterdam. Prof. dr. F.J. Bex is bijzonder hoogleraar data science en de rechtspraak aan het Tilburg Insitute for Law, Society and Technology (TILT), Tilburg University, wetenschappelijk directeur van het Nationaal Politielab AI bij het Innovation Centre for AI (ICAI) en universitair docent AI bij het departement Informatica, Universiteit Utrecht.

Noten

2. Voorbeelden zijn het rapport *Gewogen risico. Deel 2: Behandeling opleggen aan zedendelinquenten* van de Nationaal Rapporteur Mensenhandel en Seksueel Geweld tegen Kinderen, Den Haag: Nationaal Rapporteur 2017 en, recenter, het rapport *Forensische zorg en veiligheid. Lessen uit de casus Michael P.* van de Onderzoeksraad voor Veiligheid, Den Haag: OVV 2019.

3. S. Fazel, J.P. Singh & H. Doll, 'Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24.827 people: systematic review and meta-analysis', *British Medical Journal* 2012 (DOI: 10.1136/bmj.e4692). Zie ook: J. Harte, 'Recidive inschatten met behulp van een empirisch model. Kansen voor de strafrechtspraktijk?', *NJB* 2017/1799, afl. 33.

4. We gebruiken in dit artikel de term AI om te verwijzen naar op statistiek gebaseerde zelflerende algoritmes, ook wel bekend als *machine learning*-algoritmes.

5. Zie bijvoorbeeld Jaap van den Herik in *Mr.* (www.mr-online.nl/in-2030-zullen-computers-rechtspreken/), Folkert Jensma in *NRC Handelsblad* (www.nrc.nl/nieuws/2017/10/28/big-data-kunnen-ook-de-rechter-verdringen-13312623-a1579020), een interview met Mireille Hildebrandt (newsroom.unsw.edu.au/news/business-law/ai-law-how-lawyers-and-scientists-can-avoid-automated-injustice), Corien Prins en Jurgen van der Roest, 'AI en de rechtspraak', *NJB* 2018/206, afl. 4; en Henry Prakken, 'Komt de robotrechter eraan?' *NJB* 2018/207, afl. 4.

drage kunnen leveren op het gebied van inschatten van het gevaar dat een verdachte zal recidiveren?[6] AI kan namelijk automatisch complexe verbanden vinden tussen voorspellende factoren en recidivegevaar[7] en zo met voorspellingen komen die mogelijk accurater zijn dan bestaande risicotaxatie-methoden.[8] AI-risicotaxatie is echter ook negatief in het nieuws geweest: het zou vaak bevooroordeeld (*biased*) zijn tegen bepaalde bevolkingsgroepen[9] en de algoritmen zijn *black boxes* die hun redenen voor een bepaalde risicoinschatting niet prijs kunnen geven.[10]

De snelle ontwikkeling van AI werpt een aantal prangende vragen op voor strafrechtjuristen en forensisch gedragsdeskundigen. Hoe werkt algoritmische risicotaxatie met AI-technieken eigenlijk? En, welke kritische vragen moeten we bij gebruik van dergelijke AI in het achterhoofd houden? In dit artikel trachten wij een eerste antwoord te geven op deze vragen.[11]

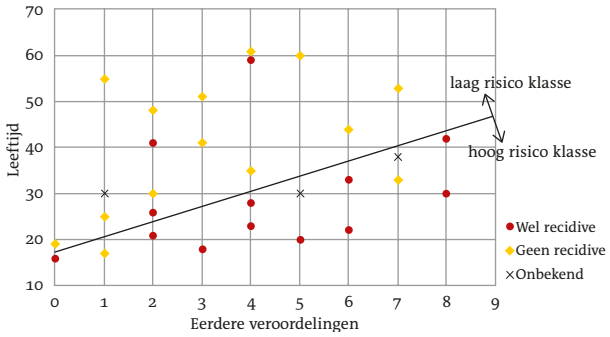
2. Risicotaxatie met AI

Risicotaxatie draait om het inschatten van de kans dat een bepaalde veroordeelde opnieuw een delict pleegt. Het idee van risicotaxatie met AI is dat een algoritme zelf uit bestaande data over recidives patronen vindt en zo de kans leert schatten dat een nieuwe, nog ongeziene veroordeelde recidiveert. Met andere woorden, wat zijn de belangrijke risicokenmerken van (niet-)recidivisten?

Risicotaxatie is een typisch classificatieprobleem. Bij classificatie worden personen waarvan we wéten of ze nadat ze vrijgelaten gerecidiveerd hebben aan de hand van bepaalde kenmerken (bijvoorbeeld geslacht, leeftijd, eerdere veroordelingen) in één van de categorieën of klassen (hoog-risico, laag-risico) ingedeeld, zodat de wiskundige relatie tussen deze kenmerken en klassen vastgesteld kan worden. In figuur 1 zijn bijvoorbeeld 27 personen te zien van wie bekend is of ze wel (rode stip) of niet (gele ruit) opnieuw gerecidiveerd hebben na hun vrijlating. Voor elke persoon is de leeftijd (y-as) en het aantal eerdere veroordelingen (x-as) aangegeven. De doorgetrokken lijn is de zogenaamde *decision boundary* welke de – in dit geval lineaire – wiskundige functie aangeeft die de twee klassen van elkaar scheidt. Deze is zo bepaald dat zo veel mogelijk recidivisten aan de ene kant van de lijn vallen (de klasse 'hoog risico) en tegelijkertijd zo veel mogelijk niet-recidiverende personen aan de andere kan van de lijn vallen (de klasse 'laag risico').

De *decision boundary* kan gebruikt worden om nieuwe, nog ongeziene personen aan de hand van hun kenmerken in een van de klassen in te delen, en zo dus te 'voorspellen' wat de kans is dat zo iemand in de toekomst gaat recidiveren. In figuur 1 zijn drie personen van wie we niet weten of ze weer zullen recidiveren weergegeven als kruisjes. Gegeven onze classificatie kunnen we nu inschatten in welke klasse deze personen zouden moeten zitten, dat wil zeggen: of er een hoog of laag risico op recidive is. De persoon van 30 jaar met één eerdere veroordeling, bijvoorbeeld, valt aan de 'laag risico' kant van de lijn en zal dus als laag risico aangemerkt worden. De twee anderen, één van 30 jaar met vijf veroordelingen en één van 38 jaar met zeven veroordelingen, zullen als hoog risico aangemerkt worden.

Classificatie is nooit volledig accuraat: er zullen altijd vals-positieven en vals-negatieven zijn



Figuur 1: Classificatie van wel en niet recidiverenden

Classificatie is nooit volledig accuraat: er zullen altijd vals-positieven (hoog risico-verdachten die niet hebben gerecidiveerd of zullen recidiveren) en vals-negatieven (laag risico-verdachten die wél hebben gerecidiveerd of zullen recidiveren) zijn. De reden hiervoor is dat het vaak onmogelijk is om de *decision boundary* op zo'n manier te bepalen dat alle niet-recidivisten in de 'laag risico' categorie vallen en alle recidivisten in de 'hoog-risico' categorie. In het voorbeeld van figuur 1 zijn er twee vals-positieven (zeventien jaar, één veroordeling en 33 jaar, zeven veroordelingen) en drie vals-negatieven (26 en 41 jaar, beiden twee veroordelingen en 59 jaar, vier veroordelingen).[12]

Het aantal goede en foute classificaties kan worden gebruikt in verschillende beoordelingsmaten voor classificatie. Figuur 2 is een zogenaamde *confusion matrix* waarin het voorspelde risico (de klasse 'hoog risico' of 'laag risico'), het werkelijke risico (wel of niet gerecidiveerd) en de beoordelingsmaten staan. Een veelgebruikte maat is *accuracy*. De classificatie in figuur 1 is 81% accuraat. Dat wil zeggen: 81% van de personen is correct ingeschat als hoog of laag risico. Dat lijkt misschien een vrij goede inschatting, maar zoals verderop zal blijken kan accuratesse een misleidende maat zijn. Een betere maat is bijvoorbeeld de precisie (*precision*) van de classificatie: als er hoog risico wordt geschat, hoe vaak is dat correct (zijn er veel vals-positieven)? In figuur 1 is dat tien van de twaalf keer (dus twee vals-positieven) wat een *precision* van 83% geeft. Verder is de 'reikwijdte' (*recall*, in de medische diagnostiek ook wel sensitiviteit genoemd) van de classificatie relevant: hoeveel van de werkelijk recidiverende hoog-risicogevallen worden als zodanig aangemerkt (zijn er veel vals negatieven)? In figuur 1 zijn dat tien van de dertien gevallen (drie vals-negatieven), wat een *recall* van 77% geeft.

Het is goed mogelijk om met 'traditionele' technieken uit de statistiek handmatig een classificatie (*decision boundary*) te bepalen als de data niet al te complex zijn.[13] Dit wordt echter lastig als er veel persoonskenmerken zijn die mogelijk ook nog onderlinge relaties hebben, of als er meer dan twee klassen zijn (bijvoorbeeld hoog-middel-laag risico, of een score van 1 tot 10). In zo'n geval kan classificatie beter – of in ieder geval gemakkelijker – automatisch met behulp van een AI-algoritme gedaan worden.[14]

		Werkelijk Risico		
		Wel recidive	Geen recidive	
Voorspeld Risico	Hoog risico	Terecht positief	Vals positief	*Precision* $\frac{TP}{TP+VP}$
	Laag risico	Vals negatief	Terecht negatief	
		Recall $\frac{TP}{TP+VN}$		*Accuracy* $\frac{TP+TN}{TP+VP+TN+VN}$

		Werkelijk Risico		
		Wel recidive	Geen recidive	
Voorspeld Risico	Hoog risico	10	2	$\frac{10}{10+2} = 83\%$
	Laag risico	3	12	
		$\frac{10}{10+3} = 77\%$		$\frac{10+12}{12+15} = 81\%$

Figuur 2: Confusion-matrix voor risicoinschatting met links beoordelingsmaten en rechts cijfers en percentages van figuur 1

6. Een andere toepassing van AI voor risicotaxatie in het strafrecht is in de opsporingsfase (*predictive policing*). Zie daarover R.A. Hoving, 'Verdacht door een algoritme. Kan *predictive policing* leiden tot een redelijke verdenking?', *Delikt en Delinkwent* 2019/41; M.B. Schuilenburg & A. Das, '*Predictive policing*: waarom bestrijding van criminaliteit op basis van algoritmen vraagt om aanpassing van het strafprocesrecht', *Strafblad* 2018/4.

7. R.A. Berk & J. Bleich, 'Statistical procedures for forecasting criminal behavior: a comparative assessment', *Criminology & Public Policy* 2013, p. 513.

8. Zie W. Rhodes 'Machine learning approaches as a tool for effective offender risk prediction', *Criminology & Public Policy* 2013, p. 507-510 voor een inleidende discussie over de toegevoegde waarde van AI bij risicotaxatie. Zie voor een vergelijking tussen AI-gebaseerde methoden en meer klassieke 'actuariële' methoden N. Tollenaar & P.G.M. van der Heijden 'Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2013, p. 565-584; N. Tollenaar & P.G.M. van der Heijden. 'Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes' *PloS one* 14.3, 2019; Berk e.a., 'Fairness in criminal justice risk assessments: the state of the art', *Sociological Methods & Research* 2018, p. 1-42.

9. J. Angwin e.a., 'Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks', *ProPublica* 23 mei 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

10. W. Samek, T. Wiegand & K.R. Müller, 'Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models', *arXiv preprint, arXiv:1708.08296*: 2017.

11. Merk op dat deze vragen en de onderstaande discussie grotendeels van toepassing zijn op *alle* algoritmische risicotaxatiemethoden, dus ook op risicotaxatie met meer klassieke statistische methoden.

12. Van de drie onbekenden weten we pas of ze vals-positieven of -negatieven zijn als we ze in de toekomst volgen om te kijken of ze nog eens recidiveren.

13. Bestaande actuariële methoden maken vooral gebruik van (lineaire) logistische regressie, een standaardtechniek uit de statistiek (Tollenaar & van der Heijden 2013).

14. Andere voordelen van AI-algoritmen zijn dat ze kunnen omgaan met niet-lineaire classificatie, mogelijke uitschieters in de data, verschillende kosten van foutclassificaties en complexe relaties tussen voorspellende kenmerken en recidivegevaar. Zie o.a. Tollenaar & van der Heijden 2013, Berk & Bleich 2013 en Berk e.a. 2018.

Classificatie met AI is een voorbeeld van *supervised machine learning*: je geeft de computer genoeg voorbeelden van eerder geziene en besliste gevallen (de *training set*), zodat hij automatisch de *decision boundary* kan bepalen.[15] De beoordeling van het classificatiealgoritme gaat aan de hand van een zogenaamde *test set*, die verschilt van de *training set*, en die ook gevallen bevat waarin de uitkomst al bekend is. Wanneer we tevreden zijn over de prestaties van het algoritme (bijvoorbeeld *accuracy*, *precision* en *recall*) kan het algoritme ingezet worden om risico-inschattingen te geven van nieuwe verdachten. De vraag is nu wanneer we, gezien ons normatieve kader, 'tevreden' zijn en wat we met AI-voorspellingen kunnen doen.

3. Drie vragen

I Wat zijn de 'kosten' van foute voorspellingen?

In het strafrecht worden vaak beslissingen op basis van risico-inschattingen genomen. Wordt een verdachte als gevaarlijk ('hoog risico') aangemerkt, dan kan dat bijvoorbeeld leiden tot verlenging van de voorlopige hechtenis, tot oplegging van een andere of hogere straf, van tbs of andere vrijheidsbeperkingen. In het geval iemand ten onrechte als hoog risico-verdachte wordt aangemerkt (vals-positief: hij zou in werkelijkheid niet recidiveren) betekent dat dat hij ten onrechte aan een (meer ingrijpend(e)) dwangmiddel, straf of maatregel wordt blootgesteld. Als iemand ten onrechte als laag risico-verdachte wordt aangemerkt (vals-negatief: hij recidiveert wel), betekent dat dat een delict wordt gepleegd dat (mogelijk) met strafrechtelijk ingrijpen voorkomen had kunnen worden.

Vals-positieven én vals-negatieven zijn beide onwenselijk, maar kunnen niet vermeden worden. De consequenties van een foute inschatting zijn wel anders. Waar het gaat om het voorspellen van ernstige geweldsdelicten zijn de 'kosten' van een vals-negatieve voorspelling (wel recidive) bijvoorbeeld lichamelijk letsel of de dood van een slachtoffer. De kosten van een vals-positieve uitslag (in werkelijkheid geen recidive) bestaan in onterecht strafrechtelijk ingrijpen, bijvoorbeeld door middel van tbs. Dat laatste is problematisch, maar zal in een individueel geval doorgaans als minder ernstig worden beschouwd dan een levensdelict. De kosten van verschillende soorten verkeerde voorspellingen zijn in dit geval *asymmetrisch*.[16]

Die vaststelling heeft gevolgen voor het ontwerp van het algoritme. Daarin moeten de relatieve kosten van beide uitkomsten worden verwerkt.[17] Het volgende voorbeeld illustreert goed hoe dat in zijn werk gaat.[18] Met een algoritme dat in de Verenigde Staten werd ontwikkeld, werd beoogd om het recidiverisico op een (poging tot een) levensdelict van voorwaardelijk in vrijheid gestelden te berekenen.[19] Van de *test set* was bekend dat 2% van de veroordeelden recidiveerde en 98% niet. Hieruit volgt overigens dat vrij eenvoudig een grote mate van accuratesse (*accuracy*) bereikt kan worden. Als immers *altijd* voorspeld zou worden dat *geen* sprake was van recidivegevaar, dan zou de voorspelling in 98% van de gevallen juist blijken te zijn. Op de andere beoordelingsmaten wordt dan echter slecht gescoord: de *precision* en *recall* zijn beide 0%, omdat er voor niemand voorspeld wordt dat hij zal recidiveren. Dit betekent dat op deze wijze enerzijds niemand ten onrechte van zijn vrijheid is beroofd, maar anderzijds dat geen van de levensdelicten voorspeld (en voorkomen) zijn. Juist vanwege de veel hogere kosten van het ten onrechte niet voorspellen van een levensdelict in verhouding tot het ten onrechte niet toekennen van een invrijheidsstelling, is dat onbevredigend. Accuratesse is in dit geval dus geen goede beoordelingsmaat. Het algoritme werd daarom zo ingesteld dat de kosten van een (poging tot een) levensdelict twintig keer zo hoog gewaardeerd werden als de kosten van intrekking van de *probation*.

Het algoritme werd getest op data van 11.157 veroordeelden waarvan al bekend was dat zij wel of niet hadden gerecidiveerd. Het voorspelde correct welke 9.972 veroordeelden niet zouden recidiveren en van 45 van de in totaal 198 recidivisten *ten onrechte* dat zij niet zouden recidiveren (vals-negatieven). De andere 153 recidivisten werden correct geclassificeerd, wat dus een *recall* van 77% gaf. Er was echter een keerzijde: van 153 veroordeelden werd weliswaar terecht voorspeld dat zij zouden recidiveren, maar van maar liefst 987 veroordeelden werd *ten onrechte* voorspeld dat zij zouden recidiveren. De *precision* van het algoritme was dus erg laag, maar 13%. Dit hoge aantal vals-positieven is het directe gevolg van de veel hogere kosten die het algoritme toekent aan het ten onrechte niet voorspellen van een levensdelict ten opzichte van het ten onrechte niet in vrijheid stellen. Kortweg komt het erop neer dat als het algoritme blindelings gevolg zou worden, 987 veroordeelden – bijna 10% van alle 11.157 – ten onrechte van hun vrijheid zouden moeten worden beroofd om 153 (pogingen tot) levensdelicten te voorkomen.

		Werkelijk Risico: Wel recidive	Werkelijk Risico: Geen recidive	
Voorspeld Risico	Hoog risico	153	987	*Precision* $\frac{153}{153 + 987} = 13\%$
	Laag risico	45	9972	
		Recall $\frac{153}{153 + 45} = 77\%$		*Accuracy* $\frac{153 + 9972}{153 + 45 + 9972 + 987} = 91\%$

Figuur 3: Confusion-matrix met beoordelingsmaten voor het Amerikaanse voorbeeld

In werkelijkheid wordt de som al gauw complexer, bijvoorbeeld doordat veroordeelden in meer dan twee risicogroepen worden ingedeeld of doordat meer dan twee uitkomsten worden voorspeld, bijvoorbeeld drie: geen delict, geweldloos delict of geweldsdelict. De 'kostenberekening' wordt nu nog

De kosten van verschillende soorten verkeerde voorspellingen zijn in dit geval *asymmetrisch*

moeilijker: wat zijn de relatieve kosten van geen delict, een geweldloos delict en een geweldsdelict? En welke 'kosten' mogen worden meegenomen in de kostenberekening? De politiek-maatschappelijke 'kosten' van een vals-negatieve risico-inschatting kunnen bijvoorbeeld zeer hoog zijn, zo blijkt (opnieuw) uit het debat over de recidive van Michael P.

AI stelt het dilemma alleen maar onaangenaam scherp, doordat het algoritme ons dwingt tot een expliciete keuze

Het klassieke strafrechtelijke adagium luidt dat het beter is om tien schuldigen vrijuit te laten gaan dan één onschuldige te bestraffen.[20] De kosten van een onterechte veroordeling worden door juristen dus hoog ingeschat. In het Amerikaanse voorbeeld is bijna sprake van een omgekeerde verhouding vergeleken met het adagium. Is het aanvaardbaar om negen mensen ten onrechte van hun vrijheid te beroven om daarmee één ernstig misdrijf te voorkomen? Duidelijk is dat dit niet een puur 'technische' kwestie is: het gaat om een *normatieve waardering* van foute inschattingen. Het ontwikkelen van een algoritme kan alleen al daarom niet zonder de *input* van juristen. De dilemma's zijn ongemakkelijk: hoe gaan we met de vals-positieven (*precision*) en vals-negatieven (*recall*) van het AI-algoritme om? De basishouding van de strafrechtjurist is immers dat vergissingen in het voordeel van de verdachte moeten uitvallen – *in dubio pro reo*. Maar het gaat dan om het bewezen verklaren van een strafbaar feit – een oordeel over iets wat al dan niet in het verleden is gebeurd. Risicotaxatie gaat om gevaar in de toekomst – daar bestaat geen Latijns adagium voor. Op dit nog grotendeels onontgonnen terrein moeten nieuwe juridische principes worden ontwikkeld,[21] althans het toepassen van AI voor risicotaxatie vraagt dat. Deze ongemakkelijkheid is echter geen doorslaggevend argument om AI als risicotaxatiemiddel te verwerpen. AI stelt het dilemma alleen maar onaangenaam scherp, doordat het algoritme dwingt tot een expliciete keuze.

II Hoe om te gaan met bias?

Het gevaar van *bias* – vooroordeel – is een belangrijk thema in de AI. In een geruchtmakend artikel met de titel *'Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks'* werd betoogd dat het zogenaamde *Compas*-algoritme *African-Americans* vaker ten onrechte als hoog-risico aanwees, terwijl 'blanke' veroordeelden vaker ten onrechte als laag-risico werden aangemerkt.[22] De producent van Compass, Northpointe, reageerde hierop door er (onder andere) op te wijzen dat het algoritme voor beide groepen even accuraat was. Voor beide groepen lukte het om rond de 70% van de delicten goed te voorspellen.[23]

Het interessante is dat *beide* stellingen tegelijkertijd waar kunnen zijn: het percentage foute voorspellingen kan per persoonskenmerk (etniciteit, leeftijd, geslacht) variëren, maar de accuratesse van de voorspelling van recidive *per groep* kan toch hetzelfde zijn. Dat is het gevolg van verschillende *base rates*: recidive is niet evenredig verdeeld over verschillende categorieën van persoonskenmerken. Zo recidiveren vrouwen minder vaak dan mannen, waardoor het percentage vals-positieve voorspellingen van vrouwelijke recidive hoger is dan van mannelijke recidive (de *precision* van het algoritme is lager voor vrouwen). Dezelfde lagere *base rate* zorgt er echter voor dat het aantal terechte negatieven voor vrouwen hoger is dan voor mannen, waardoor het percentage *correcte* voorspellingen van recidive (de *accuracy*) voor mannen en vrouwen gelijk kan zijn.[24]

Het is duidelijk dat grote verschillen in percentages vals-positieven tussen categorieën veroordeelden oneerlijk zijn. Dit leidt er immers toe dat bijvoorbeeld vrouwen of bepaalde etnische groepen een relatief veel grotere kans hebben om *ten onrechte* door strafrechtelijk ingrijpen te worden getroffen. Er zijn verschillende technische mogelijkheden om algoritmes in dit opzicht eerlijker te maken. Zo is het bijvoorbeeld denkbaar dat het algoritme de opdracht krijgt om de *precision* te maximaliseren door bijvoorbeeld de kosten van een vals-positieve voorspelling van de ene categorie (vrouwen, etnische groep) zwaarder te wegen dan een vals-positief van een andere categorie, waardoor het percentage vals-positieven voor beide categorieën vergelijkbaar is. Er is echter altijd een *trade off* tussen *precision* en *recall*. Omdat veel meer 'bewijs' nodig is om een 'beschermde categorie' als hoog-risico aan te merken, resulteert dit niet alleen in minder vals-positieven van de bewuste groep, maar zal het algo-

15. E. Alpaydin, *Introduction to Machine Learning*, MIT Press, 2010.

16. R. Berk, *Machine learning risk assessments in criminal justice settings*, Springer 2019, p. 32-36.

17. Berk 2019, p. 34.

18. Beschreven in Berk 2019, p. 10-13.

19. Daaronder viel ook slachtofferschap van de veroordeelde zelf (Berk 2019, p. 10). Vanwege de overzichtelijkheid wordt in de hoofdtekst van recidive gesproken.

20. Dit beginsel is na 11 september 2001 in het licht van de bestrijding van terrorisme door verschillende politici gerelativeerd (E. van Sliedregt, *Tien tegen één. Een hedendaagse bezinning op de onschuldpresumptie* (oratie VU Amsterdam), Den Haag: Boom 2009, p. 11-13). Het is afkomstig van de beroemde achttiende-eeuwse Engelse jurist Blackstone. Deze (W. Blackstone, *Commentaries on the laws of England in four books. Volume II*, Callaghan 1899, p. 1414) heeft het tien-tegen-één-adagium niet bedoeld als een verbod op preventief strafrechtelijk ingrijpen: 'preventive justice is, upon every principle of reason, of humanity, and of sound policy, preferable in all respects to punishing justice.'

21. A. Ashworth & L. Zedner (*Preventive justice*, Oxford: Oxford University Press 2014, p. 269) ontwikkelen beginselen waaraan een op preventie geënt strafrecht moet voldoen en die ook in de discussie over AI en risicotaxatie bruikbaar kunnen zijn, bijvoorbeeld een 'onschadelijkheidspresumptie': 'In principle, every citizen has a right to be presumed harmless, and this presumption of harmlessness can be rebutted only in exceptional circumstances.'

22. Angwin e.a. 2016, voetnoot 9.

23. W. Dieterich, C. Mendoza & T. Brennan, *COMPAS risk scales: demonstrating accuracy equity and predictive parity* (rapport Northpointe), 2016 (www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html), p. 3. (Verwezen is naar een door ProPublica geannoteerde versie.)

24. Berk e.a. 2018, p. 3-9.

ritme ook een groter aantal verdachten 'missen' die zullen recidiveren: het aantal vals-negatieven neemt toe in de 'beschermde categorie'. In dit geval bestaat dus een *trade off* tussen *precision* (rechtvaardigheid) en *recall* (veiligheid van de maatschappij).

Wanneer wordt het maken van onderscheid (iets wat op zich acceptabel is) discriminatie (onacceptabel)?

Ook bestaat in dit voorbeeld een *trade off* tussen verschillende concepten van rechtvaardigheid. De zojuist beschreven aanpassing is 'eerlijker' waar het gaat om het vermijden van vals-positieven (*outcome fairness*), maar vereist voor de ene categorie meer 'bewijs' voor een classificatie als hoog-risico dan voor de andere. De vraag kan worden opgeworpen of deze verschillende beoordeling eerlijk is tegenover de categorie die op basis van minder 'bewijs' als hoog-risico wordt aangemerkt (*treatment fairness*).[25] Je zou kunnen zeggen dat mensen in een groep met lage *base rate* (bijvoorbeeld vrouwen) op deze manier bevoordeeld worden ten opzichte van groepen met een hogere *base rate* (mannen). Voor een vrouw immers is meer bewijs nodig om haar als hoog risico aan te merken dan voor een man – die laatste krijgt veel makkelijker het predicaat: hoog-risico.

Bias kan zich al in de data bevinden op basis waarvan het algoritme is ontwikkeld of zich verder ontwikkelt. Etnisch profileren kan er bijvoorbeeld toe leiden dat personen uit een bepaalde etnische groep veel vaker met politie in aanraking komen dan verwacht zou mogen worden op basis van het werkelijke aandeel van deze groep in de misdaadcijfers. Vanzelfsprekend leidt dat ertoe dat het algoritme dan te veel personen uit die groep als hoog-risico zal aanmerken. In zekere zin klopt dat: ze worden vaker aangehouden, maar dat is niet het gevolg van meer strafbaar gedrag, maar van een hogere 'pakkans' door het etnisch profileren. In dat geval ligt een correctie van het algoritme zoals hiervoor beschreven voor de hand. We zien hier overigens dat de *bias* niet zozeer door het algoritme wordt gecreëerd, maar dat de *bias* die al in het rechtssysteem bestaat veeleer door het algoritme wordt *gereflecteerd*. In die zin houdt het algoritme de maatschappij een – opnieuw ongemakkelijke – spiegel voor.

Maar hoe moet erover worden gedacht als een bepaalde etnische groep relatief veel jonge mannen kent? In dat geval zal de etnische groep waarschijnlijk óók vaker als recidivegevaarlijk worden aangewezen, maar dan kan dat het gevolg zijn van het disproportionele aandeel jonge mannen in de groep – van wie bekend is dat deze vaker delicten plegen. Mogelijk zal men hier minder moeite mee hebben dan wanneer sprake is van een overrepresentatie als gevolg van etnisch profileren als oorzaak van de *bias*. De wijze waarop de *bias* jegens een groep tot stand komt, lijkt dus van belang te zijn voor de acceptatie ervan. Tegelijkertijd kan gesteld worden dat leeftijd en geslacht net zo goed kenmerken zijn die de veroordeelde niet kan aanpassen als zijn etniciteit.[26] Met andere woorden: is leeftijds- en seksediscriminatie bij risicotaxatie wel (altijd) acceptabel? Risicotaxatie *is* – heel letterlijk – het 'discrimineren' tussen individuen. Dat wil zeggen: het onderscheiden van minder risicovolle van meer risicovolle veroordeelden op basis van kenmerken van groepen individuen. Dat onderscheid is vaak afhankelijk van factoren die de betrokkene niet in zijn macht heeft. Wanneer wordt het maken van onderscheid (iets wat op zich acceptabel is) discriminatie (onacceptabel)?[27]

Op allerlei momenten in de 'strafrechtketen' worden ook nu risico-inschattingen door politie, officieren van justitie, reclasseringsmedewerkers, gedragsdeskundigen en rechters gemaakt. Soms op basis van een meer of minder complex risicotaxatie-instrument, maar vaak ook niet. Het is niet onwaarschijnlijk dat het effect van (on)bewuste menselijke *biases* op dit moment groter is dan met goed werkende AI kan worden bereikt. Het verminderen van *biases* gaat noodzakelijkerwijs ten koste van de *recall* van de voorspelling (meer vals-negatieven), maar dat kan een prijs zijn die betaald moet worden om een (meer) rechtvaardige werking van het algoritme te bereiken. (Hoewel de 'kosten' van het verlies aan juiste voorspellingen weer moeten worden afgewogen tegen de belangen van potentiële slachtoffers, die mogelijk vaker door recidive worden getroffen.) Ook dit zijn geen louter technische kwesties. Wat hier als rechtvaardig geldt, zal in het algemeen aan discussie onderhevig zijn: welke typen van onderscheid die het algoritme maakt zijn acceptabel, en welke niet?

III Hoe kom je van een voorspelling naar rehabilitatie?

Een AI-algoritme probeert zo optimaal mogelijk te voorspellen, gebruikt daarvoor alle beschikbare informatie en kan meer en vooral complexere verbanden leggen en patronen zien in de gegevens dan klassieke risicotaxatie-instrumenten of menselijke beslissers. Het algoritme gebruikt daarbij geen theoretisch model voor het ontstaan van recidive, het is in die zin 'theorievrij'.[28] Met andere woorden, het heeft geen kennis van recidivebepalende factoren, maar maakt gebruik van alles wat de voorspelling beter maakt, ook schijnbaar niet-relevante kenmerken of factoren: 'if an offender's shoe size is a useful predictor, it can be productively included.'[29]

Het algoritme geeft dus slechts een voorspelling door te kijken naar *correlaties* tussen factoren en recidivegevaar, het geeft geen *oorzaken* van het risico op recidive en heeft daar ook geen 'weet' van. Dat betekent dat er op grond van de voorspelling op zich ook geen *aanknopingspunten* voor vermindering van het risico zijn – de factoren die helpen het risico te voorspellen hoeven immers helemaal geen causale factoren te zijn. We weten dus niet hoe we het risico, middels behandeling bijvoorbeeld, kunnen verlagen. Dat is een probleem. Met de huidige risicotaxatie-instrumenten kunnen we vaststellen op welke risicofactoren de verdachte of veroordeelde 'scoort'. Vaak zijn dat 'statische' factoren als leeftijd en justitieel verleden, in andere gevallen 'dynamische' factoren, zoals psychotische verschijnselen of middelengebruik. Dynamische factoren zijn factoren die vatbaar zijn voor verandering en dan

weten we dat als we de psychotische verschijnselen of het middelengebruik effectief behandelen, het risico – althans volgens het gebruikte instrument – daalt.

Nu kan een AI-algoritme ook gebruik maken van louter dynamische factoren om het risico te voorspellen – als we schoenmaat niet in de *training set* opnemen zal het algoritme het ook niet in de voorspelling meenemen. Complexe AI-algoritmen geven over het algemeen echter geen inzicht in *welke* factoren nu van grote invloed op de risico-inschatting zijn geweest. Hoewel resultaten van sommige AI-technieken als – bijvoorbeeld – een simpele beslisboom weergegeven kunnen worden, zijn juist de complexere en nauwkeurigere *machine learning*-technieken veelal een *black box* die niet aangeeft *waarom* een persoon nu als hoog of laag risico is aangemerkt.[30] De voorspelling van het algoritme geeft dan ook geen aanknopingspunten voor een behandeling. Het is echter nu juist een veronderstelling van strafrechtelijke maatregelen als tbs dat je recidivegevaar kunt verminderen door 'in te grijpen'. In het geval van tbs door de stoornis te behandelen, gepaard aan andere interventies. Weet je niets over de factoren die de hoge kans op recidive bij iemand bepalen, dan kun je misschien niet veel anders dan mensen opsluiten.

Het is dus vaak onbekend hoe een algoritme tot een voorspelling komt en de factoren die bijdragen aan een goede voorspelling hoeven geen *causale* factoren te zijn voor recidive. Het evidente gevaar hiervan is dat mensen opgesloten worden op basis van recidiverisico, zonder perspectief op hoe dit risico te verlagen. Risicotaxatietechnologie kan op deze wijze, zo is wel gesteld, massa-opsluiting stimuleren.[31] De technologie wijst de weg naar de gevangenis (incapacitatie), maar geen weg naar buiten (rehabilitatie). Barabas e.a. pleiten dan ook voor een diagnostisch perspectief, waarbij AI juist wordt ingezet om factoren waarop interventie kan plaatsvinden te identificeren. Een gerichte 'interventie' kan dan tot een verlaging van het risico – en een weg naar buiten – leiden. Identificatie van dergelijke factoren vereist ook kennis van gedragswetenschappelijke mogelijkheden tot interventie en juridische mogelijkheden tot rehabilitatie.

De technologie wijst de weg naar de gevangenis (incapacitatie), maar geen weg naar buiten (rehabilitatie)

Een nieuwe ontwikkeling is overigens *explainable AI*, waarmee beoogd wordt om de uitkomst van *machine learning*-algoritmen uit te leggen door de factoren te geven die van invloed waren op bijvoorbeeld de risico-inschatting.[32] Verder moet de transparantie van meer klassieke actuariële methoden ook niet overschat worden: waar een statisticus wellicht de rekenmodellen gebruikt en deze methoden kan doorgronden, zullen deze voor de gemiddelde strafrechtjurist net zo'n *black box* zijn als *machine learning*-algoritmen.

Conclusie

In dit artikel hebben we laten zien wat algoritmische risicotaxatie met behulp van AI inhoudt, en wat de belangrijke rol voor juristen in dat verband is. Risicotaxatie met behulp van AI – een techniek die nog in de kinderschoenen staat – heeft de potentie om een relatief betrouwbaar middel te worden om inschattingen van recidivegevaar te maken. Een belangrijk vraagpunt is tot hoeveel vals-positieven (*precision*) en vals-negatieven (*recall*) de inschatting leidt. Door de 'kosten' van verschillende fouten te bepalen en in het algoritme te verwerken, kan het belang van de maatschappelijke veiligheid worden afgewogen tegen het belang dat burgers niet ten onrechte door strafrechtelijk ingrijpen moeten worden getroffen. Een tweede vraagpunt is hoe omgegaan moet worden met *biases*. Er zijn technische manieren om vooroordelen in de algoritmes tegen te gaan, maar daarbij moet wel de normatieve vraag onder ogen worden gezien hoe dat op een rechtvaardige wijze kan gebeuren. Ten slotte zijn AI-algoritmes relatief ondoorzichtig, waardoor zij niet altijd mogelijkheden aanwijzen voor interventies die gericht zijn op rehabilitatie van de veroordeelde. Een nieuwe ontwikkeling, *explainable AI*, zou meer zicht kunnen geven op de factoren die van invloed waren op de risico-inschatting.

AI voor risicotaxatie kan dus een belangrijk instrument worden in de toekomst. Het is van groot belang dat ontwikkelaars van (nieuwe) AI-technieken en juristen sámen optrekken. AI is geen waardenvrije techniek, maar vraagt juist ook om normatieve *input* van juristen. •

25. Berk 2019, p. 123-126. Hierachter gaat een zeer gecompliceerd debat schuil over rechtvaardigheid in algoritmes. Berk e.a. (2018, p. 13-15) onderscheiden bijvoorbeeld zes verschillende concepten van *fairness* met elk eigen *trade offs*. Dit debat vertoont overeenkomsten met klassieke filosofische, ethische en juridische discussies over rechtvaardigheid en discriminatie. Betwijfeld wordt of de essentie van gecompliceerde en contextafhankelijke noties als rechtvaardigheid en discriminatie wel in algoritmes kan worden gevangen (R. Binns, 'Fairness in machine learning: Lessons from political philosophy', *Journal of Machine Learning Research* 2018/81, p. 1-11). Joseph e.a. ('Rawlsian fairness for machine learning', *3rd Workshop on fairness, accountability and transparency in machine learning* 2016) hebben echter een technische definitie van het beginsel 'fair equality of opportunity' van Rawls in een algoritme opgenomen.

26. Vgl. Binns 2018, p. 7.

27. Binns (2018, p. 4-5) noemt risicotaxatie in dit opzicht 'quintessentially discriminatory', maar laat ook zien dat het probleem van discriminatie door algoritmes filosofisch complex is.

28. C. Barabas e.a., 'Interventions over predictions: reframing the ethical debate for actuarial risk assessment', *Proceedings of machine learning research* 2018/81, p. 6; Berk 2019, p. 5.

29. Berk 2019, p. 18.

30. Aplaydin 2010; Samek et al. 2018.

31. Barabas e.a. 2018, p. 10.

32. F. Doshi-Velez & B. Kim, 'Towards a rigorous science of interpretable machine learning', arXiv preprint, 2017 (arxiv.org/abs/1702.08608).