Artikel

AI-risicotaxatie: nieuwe kansen en risico's voor statistische voorspellingen van recidive

G.M. (Max) de Vries, mr. dr. J. (Johannes) Bijlsma, prof. mr. dr. A.R. (Anne Ruth) Mackor, prof. dr. F.J. (Floris) Bex en prof. dr. G. (Gerben) Meynen*

1. Inleiding

Met name in de Verenigde Staten worden nieuwe risicotaxatie-instrumenten ontwikkeld die kunstmatige intelligentie (AI) toepassen om recidiverisico te voorspellen. AI-risicotaxatie is net als actuariële risicotaxatie gebaseerd op statistische verbanden tussen voorspellende factoren en recidive, maar door toepassing van *supervised machine learning* worden er meer, nauwkeuriger en nieuwe verbanden gelegd. De vooruitgang waarop wordt gehoopt, is dat AI-risicotaxatie-instrumenten recidive accurater[1] voorspellen dan bestaande actuariële instrumenten.[2] Enerzijds wordt ervoor gepleit om in de rechtspraktijk meer geavanceerde risicotaxatie-instrumenten te gebruiken om tot meer accurate voorspellingen te komen.[3] Anderzijds stuit AI-risicotaxatie op weerstand. De kritiek luidt onder andere dat deze risicotaxatie-instrumenten *biased* (bevooroordeeld) zouden zijn, vooral ten opzichte van etnische minderheden. Ook zouden deze risicotaxatie-instrumenten black boxes zijn, omdat de voorspellingen niet transparant zijn voor de gebruikers en soms zelfs niet voor de makers van de instrumenten, en omdat het algoritme bedrijfsgeheim kan zijn.[4] De komst van AI-risicotaxatie heeft zodoende geleid tot een nieuw maatschappelijk, wetenschappelijk en uiteindelijk ook juridisch debat.

In Nederland is het nog niet zo ver als in de Verenigde Staten: de toepassing van AI-risicotaxatie-instrumenten moet hier nog beginnen. Toch zal de komst van AI-risicotaxatie naar Nederland waarschijnlijk een kwestie van tijd zijn.[5] Voordat deze vorm van risicotaxatie in de Nederlandse rechtspraktijk wordt toegepast, is het verstandig om na te denken over het gebruik van deze risicotaxatie-instrumenten. Het debat over AI-risicotaxatie is in Nederland al wel op gang gekomen. In juni 2020 ontstond discussie over OxRec, een risicotaxatie-instru-

* Max de Vries volgt de Master Rechtswetenschappelijk Onderzoek aan de Rijksuniversiteit Groningen. Johannes Bijlsma is als universitair docent strafrecht verbonden aan het Willem Pompe Instituut voor Strafrechtswetenschappen (WPI) en het Utrecht Centre for Accountability and Liability Law (UCALL), Universiteit Utrecht. Anne Ruth Mackor is als hoogleraar professie-ethiek, in het bijzonder van juridische professies, werkzaam bij de Faculteit Rechtsgeleerdheid, Rijksuniversiteit Groningen. Floris Bex is bijzonder hoogleraar data science en rechtspraak aan het Department of Law, Technology, Markets and Society, 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. Gerben Meynen is als hoogleraar forensische psychiatrie verbonden aan het WPI en UCALL, Universiteit Utrecht en tevens bijzonder hoogleraar ethiek en psychiatrie aan de Vrije Universiteit Amsterdam.

1. De accuratesse (*accuracy*) is het percentage correcte voorspellingen van een risicotaxatie-instrument. Accuratesse is een maatstaf om aan te duiden hoe goed een instrument voorspelt. Er zijn ook andere maatstaven, zie R.J. Verkes, 'Technische begrippen bij risicotaxatie-instrumenten', in: J.W. Hummelen, R.J. Verkes en M.J.F. van der Wolf (red.), *Forensische psychiatrie in de rechtspraktijk*, Amsterdam: De Tijdstroom 2020, p. 307-312.
2. Zie paragraaf 3.
3. In Nederland bijvoorbeeld door C.E. Dettmeijer en L.M.E. Menenti, *Gewogen risico. Deel 2: Behandeling opleggen aan zedendelinquenten*, Den Haag: Nationaal Rapporteur Mensenhandel en Seksueel Geweld tegen Kinderen 2017; *Forensische zorg en veiligheid. Lessen uit de casus Michael P. van de Onderzoeksraad voor Veiligheid*, Den Haag: OVV 2019.
4. Zie paragraaf 4.
5. Nieuwe inzichten uit Noord-Amerika op het gebied van risicotaxatie worden na verloop van tijd ook in Nederland toegepast. Actuariële risicotaxatie is hiervan een voorbeeld. Zie S.M.M. Lammers, 'Risicotaxatie', in: Hummelen, Verkes en Van der Wolf 2020, p. 288-289.

doi: 10.5553/Bsb/266669012021002002004

ment dat de reclassering gebruikt.[6] Het verwijt was dat OxRec kan leiden tot etnisch profileren,[7] iets wat de reclassering weersprak.[8] De discussie over OxRec vertoont overeenkomsten met de discussie die in de Verenigde Staten wordt gevoerd over COMPAS, een veelgebruikt[9] risicotaxatie-instrument dat met name Afro-Amerikanen zou discrimineren.[10]

Deze discussies zijn relevant voor juristen, want voor hen is een belangrijke rol weggelegd bij de beantwoording van de vragen die AI-risicotaxatie oproept.[11] Daarvoor moeten zij echter begrip hebben van wat AI-risicotaxatie is. In deze bijdrage geven we uitleg van AI-risicotaxatie door deze te vergelijken met bestaande vormen van risicotaxatie. We bespreken verschillen en overeenkomsten tussen de verschillende soorten risicotaxatie en laten zien in hoeverre AI-risicotaxatie nieuwe dilemma's met zich brengt voor juristen.

De opbouw van de bijdrage is als volgt. In paragraaf 2 behandelen we bestaande vormen van risicotaxatie. Vervolgens gaan we in paragraaf 3 in op de vragen wat AI-risicotaxatie is, wat verschillen en overeenkomsten zijn met bestaande vormen van risicotaxatie en of AI-risicotaxatie recidive accurater voorspelt. Daarna komen in paragraaf 4 de eerdergenoemde kritiekpunten op AI-risicotaxatie aan de orde, namelijk het gevaar van *bias* en het black box-karakter. Ook bespreken we een derde kritiekpunt, namelijk dat AI-risicotaxatie niet altijd tot behandeladviezen leidt.

In paragraaf 5 sluiten we onze bijdrage af met een korte conclusie en slotbeschouwing. Daarin benoemen we drie fundamentele problemen die al langer spelen op het bredere terrein van het preventieve sanctierecht, maar die door de actuele discussie over het gebruik van (AI-)risicotaxatie in het strafrecht urgenter worden. Het gaat om het verschil tussen het bewijzen van een in het verleden gepleegd strafbaar feit en het voorspellen van een strafbaar feit in de toekomst; om de onduidelijke bewijsmaatstaf bij het voorspellen van toekomstige strafbare feiten en om het ontbreken van een normering van het gevaar in het preventieve sanctierecht.

2. Bestaande vormen van risicotaxatie

Risicotaxaties worden verricht ten behoeve van sanctierechtelijke beslissingen en forensische behandeling.[12] Voor behandeling identificeert risicotaxatie de risicofactoren van de betrokkene, die vervolgens worden behandeld om het recidiverisico te verminderen.[13] In het sanctierecht speelt risicotaxatie een rol bij de oplegging en verlenging van vrijheidsbenemende en -beperkende maatregelen, zoals de terbeschikkingstelling (tbs, art. 37a Sr), de plaatsing in een inrichting voor stelselmatige daders (ISD, art. 38m Sr) en de gedragsbeïnvloedende en vrijheidsbeperkende maatregel (GVM, art. 38z Sr).[14] Om deze maatregelen te kunnen opleggen, moet zijn voldaan aan een gevaarscriterium. De gevaarscriteria van deze maatregelen zijn verschillend, maar hebben gemeen dat recidiverisico de voornaamste grond is voor het vaststellen van het gevaar.[15] De rechter is zelf geen expert wat betreft het inschatten van het recidiverisico.[16] In de rechtspraktijk wordt de rechter daarom veelal door gedragsdeskundigen geadviseerd over het recidiverisico.[17] De deskundige schat het recidiverisico in de regel (mede) in aan de hand van een risicotaxatie-instrument. De rechter kan deze inschatting vervolgens gebruiken voor de beslissing over het gevaar. Rechters blijken het deskundigenoordeel vaak te volgen.[18] Een risicotaxatie kan in de praktijk dus

6. 'Kans op etnische profilering bij algoritme van reclassering', NOS.nl 26 juni 2020; 'Systeem reclassering versterkt mogelijkheid etnisch profileren', *De Telegraaf* 26 juni 2020; 'Algoritmes gebruikt door reclassering zorgen voor etnisch profileren', *Trouw* 26 juni 2020.
7. G. van Dijck, 'Algoritmische risicotaxatie van recidive. Over de Oxford Risk of Recidivism tool (OXREC), ongelijke behandeling en discriminatie in strafzaken', *NJB* 2020, p. 1784-1790.
8. M. Maas, E. Legters en S. Fazel, 'Professional en risicotaxatie-instrument hand in hand. Hoe de reclassering risico's inschat', *NJB* 2020, p. 2055-2059.
9. Sinds de ontwikkeling in 1998 is COMPAS op meer dan één miljoen verdachten toegepast. Zie H. Devlin, 'Software "no more accurate than untrained humans" at judging reoffending risk', *The Guardian* 17 januari 2018.
10. 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. De analyse van Angwin en collega's is later bestreden, zie bijvoorbeeld A.W. Flores, K. Bechtel en C.T. Lowenkamp, 'False positives, false negatives, and false analyses: a rejoinder to "Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks"', *Federal Probation* (80) 2016, p. 38-46.
11. Zie ook J. Bijlsma, F.J. Bex en G. Meynen, 'Artificiële intelligentie en risicotaxatie. Drie kernvragen voor strafrechtjuristen', *NJB* 2019, p. 3313-3319.
12. Zie uitgebreider Lammers 2020, p. 287-288. Voor de sanctietoemeting en de forensische behandeling worden vaak verschillende risicotaxatie-instrumenten toegepast; een instrument bedoeld voor de sanctietoemeting is niet altijd geschikt voor de behandeling en vice versa.
13. Het verminderen van het recidiverisico is het doel van de behandeling in het veelgebruikte *Risk-Need-Responsivity*-model. In dit model hangt de intensiteit van de behandeling af van het recidiverisico. Zie J. Bonta en D.A. Andrews, 'Viewing offender assessment and rehabilitation through the lens of the risk-need-responsivity model', in: F. McNeill, P. Raynor en C. Totter (Eds.), *Offender Supervision: new directions in theory, research and practice*, Oxford: Willan Publishing 2010, p. 19-40.
14. Over de rol van het recidiverisico in het sanctierecht, zie S.G.C. van Wingerden, M. Moerings en J.A. van Wilsem, *Recidiverisico en straftoemeting*, Den Haag: Sdu Uitgevers 2011.
15. S. Struijk en M.J.F. van der Wolf, 'Gevaarscriteria in het strafrechtelijk sanctierecht: een risicovol ratjetoe?', *AA* 2018, p. 938-947.
16. Zie bijvoorbeeld A.R. Mackor, 'De weigerende observandus. Over bewijs, "vrijzwevende feiten", de normatieve aard en de relevantie van de stoornis', in: M.J.F. van der Wolf, F. Koenraadt en C. Kelk (red.), *Van aandoening tot delict, van delict tot sanctie. Ontwikkelingen op het grensvlak van psychiatrie en strafrecht: 2000-2014. Voordrachtenreeks van het Lutje Psychiatrisch Juridisch Gezelschap*, Deventer: Kluwer 2014, p. 353-358.
17. Soms *moet* de rechter worden geadviseerd, bijvoorbeeld bij de oplegging van tbs (art. 37a lid 3 Sr). Deze verplichting verschilt per maatregel, zie Struijk en Van der Wolf 2018, p. 938-947.
18. Dat de rechter meestal de gedragskundige rapportage volgt, blijkt uit kwantitatief onderzocht van J.M. Harte, W.M.C. van den Berg en C. Stroobach, 'De invloed van klinische Pro Justitia rapportage op de rechter', *NJB* 2005, p. 1391-1396, en kwalitatief onderzoek van M. Boone e.a., *De tenuitvoerlegging van sancties: maatwerk door de rechter?*, Den Haag: WODC 2008, p. 49-50. In relatie tot het gevaar,

doi: 10.5553/Bsb/266669012021002002004

doorslaggevend zijn voor de beslissing of vrijheid preventief wordt ingeperkt of ontnomen. Het is van groot belang dat de voorspellingen van de deskundigen correct zijn, om zoveel mogelijk te voorkomen dat ongevaarlijke delinquenten vastzitten (foutpositieven) en gevaarlijke delinquenten vrij rondlopen (foutnegatieven).

We onderscheiden drie vormen van risicotaxatie in het strafrecht: het ongestructureerd klinisch oordeel, actuariële risicotaxatie en het gestructureerd professioneel oordeel.[19] Het ongestructureerd klinisch oordeel is de oudste vorm van risicotaxatie. Dit houdt in dat de deskundige de justitiabele onderzoekt en vervolgens het recidiverisico inschat aan de hand van een eigen (klinisch) referentiekader. De deskundige bepaalt in hoge mate zelf welke factoren in de risicotaxatie worden betrokken en hoe ze worden gewogen. Er kleven nadelen aan het ongestructureerd klinisch oordeel. De voorspellingen blijken weinig accuraat en ook de interbeoordelaarsbetrouwbaarheid ligt laag. Ook zouden vooroordelen van deskundigen over wat voorspellende factoren zijn een rol kunnen spelen.[20]

Bij actuariële risicotaxatie-instrumenten wordt het recidiverisico bepaald aan de hand van statistische verbanden tussen voorspellende factoren en recidivecijfers uit het verleden. Anders dan in het ongestructureerd klinisch oordeel kan de deskundige in een actuariële risicotaxatie niet zelf de factoren en de weging van deze factoren bepalen, die zijn vooraf vastgelegd in het risicotaxatie-instrument. De deskundige hoeft het risicotaxatie-instrument dus 'slechts' in te vullen ('scoren') om het recidiverisico vast te stellen, wat volgens een vaste methode dient te gebeuren. Actuariële risicotaxatie-instrumenten kunnen eenvoudig zijn opgezet, met enkele factoren en een simpele lineaire functie. Meer recent zijn ook zeer complexe actuariële risicotaxatie-instrumenten ontwikkeld, zoals het in de inleiding genoemde OxRec en COMPAS.[21]

Een actuarieel risicotaxatie-instrument kan verschillende soorten risicofactoren bevatten. Het onderscheid tussen statische en dynamische factoren is hierbij belangrijk. Statische factoren zijn onveranderlijk en kunnen niet door forensische behandeling worden veranderd. Voorbeelden zijn het strafblad en de leeftijd van de justitiabele. De kritiek op eenzijdig gebruik van statische factoren is dat deze geen aanknopingspunt bieden om door interventies (behandeling) het risico te verlagen. Sommige actuariële risicotaxatie-instrumenten bevatten ook dynamische, oftewel veranderlijke factoren, bijvoorbeeld middelengebruik, sociaal netwerk en de mate waarin de justitiabele meewerkt aan een behandeling.

Het gestructureerd professioneel oordeel is in Nederland de meest gebruikte vorm van risicotaxatie.[22] In deze methode wordt een actuariële component in de vorm van een 'checklist' aangevuld met het klinisch oordeel van de deskundige. Anders dan in een actuariële risicotaxatie, waar de inschatting van de deskundige geen invloed heeft op de uitkomst, kan de deskundige bij het gestructureerd professioneel oordeel na kennisneming van de uitkomst van de checklist tot een aangepaste gevaarsinschatting komen, bijvoorbeeld op basis van factoren die in de actuariële risicotaxatie geen rol spelen. De gedachte is dat het gestructureerd professioneel oordeel recidive accurater voorspelt omdat het ook gebruikmaakt van de klinische inschatting van de onderzoeker ten aanzien van de individuele justitiabele.[23]

Tot nu toe hebben actuariële risicotaxatie en het gestructureerd professioneel oordeel op zijn best een matige predictieve validiteit.[24] Met name het aantal foutpositieven is hoog.[25] Uit een omvangrijke meta-analyse van Fazel e.a. van actuariële en gestructureerd professioneel oordeel-risicotaxatie-instrumenten voor het voorspellen van geweldsrecidive bleek dat van de groep van wie recidive werd voorspeld slechts 41% daadwerkelijk recidiveerde, en 59% dus niet.[26]

zie M.J.F. van der Wolf, *TBS – veroordeeld tot vooroordeel* (diss. Rotterdam), Oisterwijk: Wolf Legal Publishers 2012, p. 217.

19. Voor de bespreking van de verschillende vormen van risicotaxatie wordt naar de volgende standaardwerken verwezen: J.L. van Emmerik en E.F.J.M. Brand, 'Risicotaxatie in de forensische psychiatrie', in: H.J.C. van Marle, P.A.M. Mevis en M.J.F. van der Wolf (red.), *Gedragskundige rapportage in het strafrecht*, Deventer: Kluwer 2013, par. 20.4; V. de Vogel, M. de Vries Robbé, E. van den Broek, 'Risicotaxatie in de forensische psychiatrie: fundamenten en praktijk', in: K. Goethals, G. Meynen, A. Popma (red.), *Leerboek forensische psychiatrie*, Amsterdam: De Tijdstroom 2019, p. 501-524; Lammers 2020, p. 287-306; W.J. Smid, 'Risicotaxatie', in: W.J. Smid e.a. (red.), *Zicht op zedendelinquenten: achtergronden, risicotaxatie en behandeling*, Amsterdam: De Tijdstroom 2020, p. 127-151.
20. Zie bijvoorbeeld J. Bonta, 'Risk-needs assessment and treatment', in: A.T. Harland (red.), *Choosing correctional options that work: Defining the demand and evaluating the supply*, Thousand Oaks, CA: Sage Publications 1996, p. 18-32.
21. Dat deze bestaande actuariële risicotaxatie-instrumenten complexer zijn, wil nog niet zeggen dat ze ook accurater voorspellen dan eenvoudige instrumenten. Zo kan een eenvoudig actuarieel instrument even accuraat voorspellen als COMPAS, zie J. Dressel en H. Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* (4) 2018; Z. Lin e.a., 'The limits of human predictions of recidivism', *Science Advances* (6) 2020.
22. W.J. Smid e.a., 'Gestructureerde risicotaxatie: noodzakelijker wijs?', *De Psycholoog* 2014, p. 28-39.
23. Maar dat hoeft volgens een aantal auteurs niet tot een accurater oordeel te leiden, zie bijvoorbeeld R.K. Hanson en K.E. Morton-Bourgon, 'The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies', *Psychological Assessment* (21) 2009/1, p. 1-21.
24. S. Fazel, 'The scientific validity of current approaches to violence and criminal risk assessment', in: J.W. de Keijser, J.V. Roberts en J. Ryberg (Eds.), *Predictive sentencing: Normative and empirical perspectives*, Londen: Bloomsbury 2019, p. 197-212, en voor Nederland W.J. Smid en K. Uzieblo, 'Risicotaxatie: waarheen, waarvoor?', *De Psycholoog* 2020, p. 32-40.
25. Fazel 2019, p. 201-202.
26. S. Fazel e.a., '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, 345:e4692.

doi: 10.5553/Bsb/266669012021002002004

3. AI-risicotaxatie

3.1 Wat is AI-risicotaxatie?

Er is geen vastomlijnde definitie van AI-risicotaxatie. De belangrijkste overeenkomst tussen AI-risicotaxatie en actuariële risicotaxatie is dat ook bij AI-risicotaxatie het recidiverisico wordt bepaald aan de hand van statistiek. Net als actuariële risicotaxatie maakt AI-risicotaxatie gebruik van een algoritme. We kunnen AI-risicotaxatie daarom beschouwen als een vorm van actuariële risicotaxatie, waarbij we de huidige actuariële risicotaxatie-instrumenten in dit artikel verder als bestaande actuariële risicotaxatie aanduiden. Er zijn echter drie belangrijke verschillen tussen bestaande actuariële risicotaxatie en AI-risicotaxatie. Ten eerste gebruikt AI-risicotaxatie meestal geavanceerdere statistische methoden dan bestaande actuariële risicotaxatie. Ten tweede heeft AI-risicotaxatie de potentie om veel meer factoren in de risicotaxatie op te nemen dan bestaande actuariële risicotaxatie-instrumenten. Ten derde zijn de algoritmen van AI-risicotaxatie zelflerend: het algoritme vindt zelf op basis van de ingevoerde data de statistische verbanden tussen voorspellende factoren en recidive. We leggen deze verschillen hierna uit.

Actuariële risicotaxatie leidt tot classificatie,[27] namelijk het indelen van individuele justitiabelen in categorieën, bijvoorbeeld 'hoog risico' of 'laag risico'.[28] Classificatie moet zo accuraat mogelijk zijn, want een verkeerde classificatie is een foutpositief of foutnegatief.[29] Bestaande actuariële risicotaxatie classificeert meestal met behulp van lineaire regressieanalyse, wat – kort gezegd – inhoudt dat het verband tussen voorspellende factoren en recidive in een lineaire wiskundige functie wordt weergegeven.[30] Een beperking aan lineaire regressieanalyse is dat het verband tussen voorspellende factoren en het recidiverisico sterk wordt vereenvoudigd.[31] Daarnaast wordt lineaire regressieanalyse toegepast op niet-lineaire verbanden, waardoor voorspellingen minder accuraat zijn.[32] Bestaande actuariële risicotaxatie loopt tegen grenzen aan naarmate de verbanden tussen voorspellende factoren en recidive complexer zijn.[33]

AI-risicotaxatie maakt meestal gebruik van geavanceerdere statistische methoden dan bestaande actuariële risicotaxatie. Hierbij moet overigens worden opgemerkt dat recente bestaande actuariële risicotaxatie-instrumenten soms ook geavanceerde statistiek toepassen, wat het verschil tussen AI-risicotaxatie en bestaande actuariële risicotaxatie relativeert.[34] AI-risicotaxatie past op het hierboven omschreven classificatievraagstuk *supervised machine learning* toe, een verzamelterm voor zelflerende algoritmen die in de aangeleverde data zelfstandig patronen herkennen en vervolgens classificeren in vooraf aangegeven categorieën.[35] Hoewel het algoritme zelflerend is, bepaalt de mens dus wel de *input* (gegevens over veroordeelden en recidive) alsook de *categorieën van de output* van het algoritme (bijvoorbeeld laag/matig/hoog-risico).

Een algoritme dient eerst te worden getraind aan de hand van trainingsdata: daarin leert het algoritme patronen te herkennen. Hiervoor worden grote datasets gebruikt met kenmerken van (tien)duizenden delinquenten en hun recidivecijfers.[36] De patronen die het algoritme vindt, zijn statistische verbanden tussen kenmerken en recidive. Deze kenmerken zijn dus voorspellende factoren. Het tweede onderscheidende kenmerk van AI-risicotaxatie is de mogelijkheid om in grote hoeveelheden informatie verbanden te herkennen. Big data wordt wel omschreven als een (1) zeer grote en (2) heterogene set data die uit verschillende bronnen afkomstig zijn die, omdat die meestal in digitale vorm bestaat, (3) zeer snel kan worden geanalyseerd met behulp van algoritmen. Big data wordt daarom wel gekarakteriseerd met de termen (1) *volume*, (2) *variety* en (3) *velocity*. Met name de eerste twee eigenschappen leiden enerzijds tot mogelijkheden voor verbetering van de nauwkeurigheid van de voorspellingen, maar anderzijds juist ook tot twijfel daarover. Door het grote volume en de grote variëteit bestaat zowel een risico op niet ontdekte fouten en *biases* als ook op onjuiste dataselectie.[37]

Ook de algoritmen van bestaande actuariële risicotaxatie-instrumenten kunnen zijn gebaseerd op verbanden die zijn gelegd in omvangrijke datasets,[38] maar – en dit is het derde onderscheidende kenmerk – anders dan bij bestaande actuariële risicotaxatie zoekt het algoritme bij AI-risicotaxatie deze verbanden zelf. Dit is de zelflerende aard van *machine learning*. Er zouden nieuwe voorspellende factoren kunnen worden gevonden die nog niet in bestaande actuariële risicotaxatie-instrumenten zijn opgenomen.[39] Het algoritme hanteert bij het ontwikkelen van een bestaand actuarieel risicotaxatie-instrument geen op causale verbanden gebaseerde theorie voor het zoeken naar verbanden tussen voorspel-

27. D.M. Gottfredson, 'Prediction and classification in criminal justice decision making', *Crime & Justice* (9) 1987, p. 1-20.
28. R.A. Berk en J. Bleich, 'Statistical procedures for forecasting criminal behavior. A comparative assessment', *Criminology & Public Policy* (12) 2013/3, p. 519-521. Zie ook Bijlsma, Bex en Meynen 2019, p. 3314-3315.
29. Zie paragraaf 2.
30. N. Tollenaar en 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* (176) 2013/2, p. 566.
31. Berk en Bleich 2013, p. 523-524; T. Brennan, 'An alternative scientific paradigm for criminological risk assessment. Closed or open systems, or both?', in: F.S. Taxman (red.), *Handbook on Risk and Need Assessment: Theory and Practice*, Londen: Routledge 2016, p. 171-172.
32. Berk en Bleich 2013, p. 523-524; Brennan 2016, p. 172.
33. Berk en Bleich 2013, p. 524-525. Zie ook Bijlsma, Bex en Meynen 2019, p. 3316.
34. Zie paragraaf 2.
35. R.A. Berk, *Machine learning risk assessments in criminal justice settings*, Cham: Springer 2019, p. 3-6.
36. Berk 2019, p. 24-25, 27.
37. S. Leonelli, 'Scientific Research and Big Data', in: E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, 2020 (online).
38. Zo is het in de inleiding genoemde OxRec op de data van tienduizenden delinquenten gebaseerd, zie S. Fazel e.a., 'Prediction of violent reofending in prisoners and individuals on probation: a Dutch validation study (OxRec)', *Scientific Reports* (9) 2019, 841.
39. Dan moeten deze factoren wel in de trainingsdata zijn opgenomen, want de mens bepaalt de *input* van het algoritme. Zie Berk 2019, p. 18.

doi: 10.5553/Bsb/266669012021002002004

lende factoren en recidive.[40] Vanwege het geautomatiseerd zoeken naar verbanden worden er bij AI-risicotaxatie ook veel meer variabelen in het algoritme verwerkt dan bij bestaande actuariële risicotaxatie.

De toepassing van *machine learning* bepaalt dus het verschil met bestaande actuariële risicotaxatie. *Supervised machine learning* is een verzamelterm[41] en er zijn ook belangrijke verschillen tussen AI-algoritmen, maar deze algoritmen delen wel een aantal voordelen vergeleken met de algoritmen van bestaande actuariële risicotaxatie. Ten eerste kan AI-risicotaxatie de statistische relatie tussen voorspellende factoren en recidive nauwkeuriger vastleggen dan bij lineaire regressieanalyse. AI-risicotaxatie kan beter omgaan met complexe, niet-lineaire verbanden.[42] Ten tweede kunnen de algoritmen van AI-risicotaxatie vanwege de grotere rekenkracht gemakkelijker in meerdere categorieën classificeren, zoals bijvoorbeeld geen recidive, recidive van een niet-geweldsdelict en recidive van een geweldsdelict.[43] Ten derde kunnen in algoritmen van AI-risicotaxatie gemakkelijker de 'asymmetrische kosten' van een foute voorspelling worden opgenomen. Asymmetrische kosten houden in dat in het algoritme verschillende waarden worden toegekend aan foutpositieven en foutnegatieven, iets wat niet kan bij lineaire regressieanalyse.[44] De kosten van een foute voorspelling worden in het algemeen anders gewaardeerd bij een foutpositief dan bij een foutnegatief. In het eerste geval zit een delinquent ten onrechte vast, in het tweede geval wordt een (ernstig) delict gepleegd dat had kunnen worden voorkomen. De consequentie van een foutnegatief wordt in de Verenigde Staten over het algemeen als ernstiger gezien dan die van een foutpositief: de vrijheidsbeneming van een justitiabele wordt minder zwaar gewogen dan het verlies van leven of lichamelijke integriteit van het slachtoffer.[45]

3.2 Voorspelt AI-risicotaxatie accurater dan bestaande risicotaxatie?

Een belangrijke vraag is of AI-risicotaxatie recidive accurater voorspelt dan bestaande risicotaxatie.[46] Zo algemeen geformuleerd kan deze vraag moeilijk worden beantwoord. Onder bestaande risicotaxatie en AI-risicotaxatie vallen verschillende risicotaxatie-instrumenten, zodat het maar net is wat en hoe wordt vergeleken. Een eerste benadering is het vergelijken van statistische modellen van AI-gebaseerde methoden met modellen van bestaande actuariële risicotaxatie. Een aantal onderzoeken vond een hogere accuratesse bij AI-risicotaxatie,[47] maar in andere onderzoeken bleek AI-risicotaxatie niet accurater te voorspellen dan bestaande actuariële risicotaxatie.[48] Verklaringen voor de verschillende uitkomsten van deze onderzoeken zijn gezocht in de algoritmen en in de data die de onderzoekers hebben gebruikt.[49] De accuratesse van de algoritmen die voor AI-risicotaxatie kunnen worden gebruikt blijken onderling aanzienlijk te kunnen verschillen. Nieuwere algoritmen voorspellen over het algemeen accurater dan oudere algoritmen.[50]

Een andere benadering om de accuratesse van AI-risicotaxatie met bestaande risicotaxatie te vergelijken is om een in de praktijk gebruikt risicotaxatie-instrument af te zetten tegen een door de onderzoekers ontwikkeld AI-risicotaxatie-instrument. Een voorbeeld is een onderzoek waarin de SAVRY, een risicotaxatie-instrument voor jeugdigen (gestructureerd professioneel oordeel), werd vergeleken met een AI-risicotaxatie-instrument. Het AI-risicotaxatie-instrument bleek significant accurater te voorspellen dan de SAVRY.[51] Twee andere onderzoeken met een soortgelijke onderzoeksopzet leverden eenzelfde resultaat op.[52] Tot een andere uitkomst kwamen Dressel en Farid in een bekend onderzoek waarin zij het in de inleiding genoemde actuariële risicotaxatie-instrument COMPAS met een door henzelf ontwikkeld AI-risicotaxatie-instrument vergeleken. Het AI-risicotaxatie-instrument bleek ongeveer even accuraat te voorspellen.[53]

40. Berk 2019, p. 5. Zie ook, in kritische zin, C. Barabas e.a., 'Interventions over predictions: reframing the ethical debate for actuarial risk assessment', *Proceedings of Machine Learning Research* 2018/81, p. 6.
41. Voor de hier genoemde en andere voordelen van algoritmen, zie bijvoorbeeld Berk en Bleich 2013, p. 526-527; Brennan 2016, p. 177-179; G. Duwe en K. Kim, 'Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism', *Criminal Justice Policy Review* (28) 2017, afl. 6, p. 575-578.
42. Berk en Bleich 2013, p. 523-526.
43. R.A. Berk e.a., 'When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment', *Journal of Experimental Criminology* (6) 2010/2, p. 191-208; Berk en Bleich, 2013, p. 527. De ernst van de voorspelde recidive is van belang voor de beslissing of een preventieve maatregel is aangewezen: gevaar voor een winkeldiefstal verschilt immers van gevaar voor een roofmoord.
44. De asymmetrische kosten van een foute voorspelling kunnen in het algoritme worden opgenomen, waardoor er bijvoorbeeld minder foutnegatieven en meer foutpositieven zijn, zie Berk 2019, p. 32-36, 49-51, 78-79.
45. R.A. Berk, 'Balancing the Costs of Forecasting Errors in Parole Decisions', *Albany Law Review* (74) 2011/3, p. 1071-1085.
46. De accuratesse van een risicotaxatie-instrument is doorslaggevend voor de bruikbaarheid ervan in de rechtspraktijk. Zie R. Hester, 'Risk assessment savvy: The imperative of appreciating accuracy and outcome', *Behavioral Sciences & the Law* (38) 2020/3, p. 248-250.
47. Berk en Bleich 2013, p. 523-524; Duwe en Kim 2017, p. 570-600.
48. Tollenaar en Van der Heijden 2013, p. 565-584; N. Tollenaar en P.G.M. van der Heijden, 'Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes', *PLoS ONE* (14) 2019, e0213245.
49. S.D. Bushway, 'Is there any logic to using Logit: finding the right tool for the increasingly important job of risk prediction', *Criminology & Public Policy* (12) 2013/3, p. 564.
50. Duwe en Kim 2017, p. 570-600.
51. S. Tolan e.a., 'Why machine learning may lead to unfairness: evidence from risk assessment for juvenile justice in Catalonia', *ICAIL Conference Papers* 2019.
52. M.H. Ting e.a., 'Predicting recidivism among youth offenders: Augmenting professional judgement with machine learning algorithms', *Journal of Social Work* (18) 2018/6, p. 631-649; M. Ghasemi e.a., 'The application of machine learning to a general risk-need assessment instrument in the prediction of criminal recidivism', *Criminal Justice and Behavior* 2020.
53. Dressel en Farid 2018.

doi: 10.5553/Bsb/266669012021002002004

De voorlopige conclusie lijkt te zijn dat AI-risicotaxatie op dit moment ongeveer even accuraat voorspelt als bestaande risicotaxatie, en onder omstandigheden accurater voorspelt. Wat die omstandigheden zijn, moet nader worden onderzocht.[54] Waarschijnlijk wordt de potentie van AI-risicotaxatie nog niet ten volle benut. Berk stelt dat er soms fouten zitten in de algoritmen die in AI-risicotaxatie worden gebruikt.[55] Een andere oorzaak van de matige accuratesse van sommige AI-risicotaxaties kan zijn gelegen in het soort algoritme dat is gebruikt, omdat algoritmen verschillend kunnen presteren bij verschillende data. Het is daarom belangrijk dat bij het ontwikkelen van een AI-risicotaxatie-instrument meerdere typen algoritmen worden getest.[56] Tot slot kan de accuratesse van AI-risicotaxaties samenhangen met de *fairness* van het algoritme. De keuzes die hierin worden gemaakt kunnen de accuratesse beïnvloeden. We komen hierop terug in paragraaf 4.1 over AI-risicotaxaties en *bias*.

4. Kritiekpunten

In deze paragraaf gaan we dieper in op de in paragraaf 1 genoemde kritieken op AI-risicotaxatie. Hierbij bespreken we de vraag of deze kritieken ook opgaan voor bestaande actuariële risicotaxatie.

4.1 *Biases* en discriminatie

AI-risicotaxatie is omstreden wegens *biases*. Mehrabi e.a. omschrijven *bias* als: 'any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics.'[57] *Biases* kunnen leiden tot discriminatie.[58] Twee vormen van *bias* kunnen worden onderscheiden.[59]

Ten eerste kan *bias* besloten liggen in de data waarmee algoritmen worden getraind.[60] Algoritmen kunnen bijvoorbeeld worden getraind met politiegegevens over aanhoudingen.[61] Politieagenten kunnen relatief vaker tot aanhouding overgaan van personen van een bepaalde etniciteit dan personen die deze etniciteit niet hebben. Het algoritme zal dan personen met deze kenmerken eerder aanwijzen als hoog-risico. Als het relatief hoge aantal aanhoudingen niet (alleen) komt doordat de etnische groep om wie het gaat daadwerkelijk vaker recidiveert, maar (ook) door bewuste of onbewuste vooroordelen van politieagenten, zullen personen die tot deze groep behoren door het algoritme ten onrechte vaker als hoog-risico worden geclassificeerd. Menselijke *biases* planten zich op deze manier voort in risicotaxatie. Voor de ontwikkeling van zowel AI-risicotaxatie-instrumenten als bestaande actuariële risicotaxatie-instrumenten kunnen gegevens zijn gebruikt die bevooroordeeld zijn.[62]

Een tweede bron van *bias* is inherent aan risicotaxatie. Risicotaxatie met en zonder AI draait om het opsporen van kenmerken van personen die een hoger risico vormen dan anderen. Personen die deze kenmerken bezitten worden als hoog-risico geclassificeerd.[63] Zolang een bepaald kenmerk voorspellend is voor recidive, kan het als risicofactor worden gebruikt.[64] Leeftijd en geslacht zijn voorbeelden van bekende risicofactoren. Vaak liggen deze kenmerken (of *proxies* daarvan[65]) buiten de controle van de betrokkenen en is discriminatie op basis van deze kenmerken doorgaans niet toegestaan.[66] AI-risicotaxatie wordt daarom wel *quintessentially discriminatory* genoemd,[67] maar hetzelfde geldt voor bestaande actuariële risicotaxatie.

Beide vormen van *bias* zijn dus niet uniek voor AI-risicotaxatie. Een voordeel van AI is dat het kan worden ingezet om naar mogelijke oplossingen te zoeken. In literatuur over AI-risicotaxatie vindt momenteel een interessante discussie plaats over de wijze waarop op een rechtvaardige manier met *bias* kan worden omgegaan en in hoeverre het mogelijk is om algoritmen daarop aan te passen.[68] *Biases* in datasets kunnen mogelijk met behulp van AI worden opgespoord, hoewel dit vanwege het gebrek aan transparantie van AI-risicotaxatie moeilijk kan zijn.[69]

4.2 Transparantie

Een veelgehoorde kritiek op zowel bestaande actuariële risicotaxatie als AI-risicotaxatie is het black box-karakter van deze instrumenten: er worden kenmerken van de

54. W. Rhodes, 'Machine learning approaches as a tool for effective offender risk prediction', *Criminology & Public Policy* (12) 2013/3, p. 509; T. Brennan en W.L. Oliver, 'The emergence of machine learning techniques in criminology: implications of complexity in our data and in research questions', *Criminology & Public Policy* (12) 2013/3, p. 555.
55. R.A. Berk, 'Artificial intelligence, predictive policing, and risk assessment for law enforcement', *Annual Review of Criminology* (4) 2021, p. 4.21-4.22.
56. Duwe en Kim 2017, p. 595.
57. N. Mehrabi e.a., 'A survey on bias and fairness in machine learning', *arXiv* 2019, 1908.09635v2.
58. Mehrabi e.a. 2019 onderscheiden zes verschillende vormen van discriminatie. Over het probleem om tot een goede definitie van ongerechtvaardigde discriminatie te komen, zie R. Binns, 'Fairness in machine learning: lessons from political philosophy', *Proceedings of Machine Learning Research* (81) 2018, p. 149-159.
59. L. Eckhouse e.a., 'Layers of bias: A unified approach for understanding problems with risk assessment', *Criminal Justice and Behavior* (46) 2019/2, p. 185-209.
60. Mehrabi e.a. 2019 onderscheiden 23 soorten *biases* in data.
61. Dit is bijvoorbeeld het geval bij het in de inleiding genoemde COMPAS-algoritme, zie T. Brennan, W. Dieterich en B. Ehret, 'Evaluating the predictive validity of the Compas risk and needs assessment system', *Criminal Justice and Behavior* (36) 2009/1, p. 21-40.
62. T. Douglas e.a., 'Risk assessment tools in criminal justice and forensic psychiatry: the need for better data', *European Psychiatry* (42) 2017, p. 134-137.
63. Zie paragraaf 3.1.
64. Berk 2019, p. 17-18; De Vogel, De Vries Robbé en Van den Broek 2019.
65. B. Davies en T. Douglas, 'Learning to discriminate: the perfect proxy problem in artificially intelligent criminal sentencing', in: J. Roberts en J. Ryberg (red.), *Principled Sentencing and Artificial Intelligence*, Oxford, VK: Oxford University Press 2021 (online prepublicatie).
66. Douglas e.a. 2017, p. 134-137.
67. Binns 2018, p. 149-159.
68. Dit gaat gepaard met complexe *trade-offs*. Zie Bijlsma, Bex en Meynen 2019, p. 3317-3318 voor een bespreking en nadere literatuurverwijzingen.
69. P. Saleiro e.a., 'Aequitas: A bias and fairness audit toolkit', *arXiv* 2019, 1811.05577v2.

verdachte ingevoerd en er komt een risico-inschatting uit. Omdat de berekening van het risiconiveau in de black box plaatsvindt, blijft onduidelijk hoe de kenmerken en het recidiverisico zich tot elkaar verhouden. De justitiabele en de rechter weten niet precies waarom een bepaald resultaat uit de risicotaxatie is gekomen; is het de justitiële voorgeschiedenis van de justitiabele of is het zijn thuissituatie? Dit is strafrechtelijk relevant: de oplegging en verlenging van tbs moet begrijpelijk zijn gemotiveerd,[70] en voor de behandelaar is het nodig om te weten hoe het risico kan worden verlaagd.[71] Daarnaast kan het black box-karakter de beoordeling van de accuratesse van de voorspelling[72] en de toetsing van de rechtmatigheid van de risicotaxatie bemoeilijken.[73] Ook kan het black box-karakter ertoe leiden dat fouten in het risicotaxatie-instrument onontdekt blijven.[74]

Er zijn verschillende betekenissen waarin een algoritme een black box kan zijn. Een algoritme kan een black box zijn omdat de ontwerper van het algoritme geen of onvoldoende uitleg over of toegang tot het algoritme geeft.[75] Er kan sprake zijn van intellectueel eigendom. Dit is bijvoorbeeld het geval bij het in de inleiding genoemde COMPAS-algoritme, waarvan de precieze werking bedrijfsgeheim is.[76] Een andere reden om de werking van een algoritme niet publiek te maken is dat burgers het systeem doelbewust zouden kunnen gaan 'bespelen' of misleiden door hun gedrag aan te passen. Dit was het argument van de Staat om de werking van het fraudeopsporingssysteem SyRI niet prijs te geven.[77]

Een algoritme kan ook een black box zijn omdat het voor de mens te ingewikkeld is om te begrijpen: de statistische verbanden en berekeningen zijn zo complex dat het moeilijk is inzicht te krijgen in zowel de factoren en verbanden die van invloed zijn op de risico-inschatting als de risico-inschatting die het algoritme in een individueel geval heeft gegeven.

70. Vgl. art. 13-15 Algemene Verordening Gegevensbescherming waarin het recht op 'nuttige informatie over de onderliggende logica' van automatische besluitvorming, waaronder profilering, ligt besloten. Wat deze nuttige informatie precies inhoudt, is onderwerp van discussie, zie bijvoorbeeld A.D. Selbst en J. Powles, 'Meaningful information and the right to explanation', *International Data Privacy Law* (7) 2017, afl. 4, p. 233-242.
71. Zie paragraaf 4.3.
72. A.M. Carlson, 'The need for transparency in the age of predictive sentencing algorithms', *Iowa Law Review* (103) 2017, afl. 1, p. 322-324.
73. L. Wisser, 'Pandora's algorithmic black box: the challenges of using algorithmic risk assessments in sentencing', *American Criminal Law Review* (56) 2019/4, p. 1811-1832.
74. F. Palmiotto, 'The black box on trial: the impact of algorithmic opacity on fair trial rights in criminal proceedings', in: M. Ebers en M. Cantero Gamito (Eds.), *Algorithmic governance and governance of algorithms. Data science, machine intelligence, and law*, Cham: Springer 2021, p. 51-58.
75. Carlson 2017, p. 321-322.
76. Zie *Loomis v. Wisconsin*, 881 N.W.2d 749 (Wis. 2016), waarin aan de orde was of het bedrijfsgeheime karakter van COMPAS in strijd is met het recht op *due process*. Het Hooggerechtshof van Wisconsin oordeelde van niet, zie nader R. Wexler, 'Life, liberty, and trade secrets: Intellectual property in the criminal justice system', *Stanford Law Review* (70) 2018/5, p. 1343-1430.
77. Rb. Den Haag 5 februari 2020, ECLI:NL:RBDHA:2020:865, *NJ* 2020/386, m.nt. E.J. Dommering.

In veel gevallen zijn bestaande actuariële en AI-risicotaxatie-instrumenten lastig te begrijpen voor mensen die niet of weinig statistisch onderlegd zijn, zoals de meeste burgers en juristen. Bestaande actuariële risicotaxatie-instrumenten zijn vaak wel inzichtelijk voor statistisch onderlegde onderzoekers: zij kunnen de door het instrument gelegde verbanden uitleggen en het model aan een validatiestudie onderwerpen.[78] De meer complexe *machine learning*-technieken voor AI-risicotaxatie kunnen daarentegen zelfs voor de makers van het algoritme onbegrijpelijk zijn, omdat tijdens het leerproces het algoritme impliciete aannames maakt over (patronen in) de data om zo accuraat mogelijk te kunnen voorspellen. Het algoritme leert zelf de niet-lineaire verbanden tussen de factoren en recidive en deze verbanden zijn mogelijk te subtiel om door de mens te worden opgemerkt.[79] Er is echter hoop op verbetering: de begrijpelijkheid en uitlegbaarheid van AI is onderwerp van onderzoek in de zoektocht naar *explainable AI*, waarvoor technieken worden ontworpen om complexe algoritmen inzichtelijk te maken.[80]

4.3 Risicotaxatie en resocialisatie

Idealiter biedt risicotaxatie niet alleen zicht op de hoogte van het risico, maar ook handvatten om het risico te verminderen. Daartoe moeten de risicofactoren bekend zijn en moet geprobeerd worden die te verlagen, bijvoorbeeld door behandeling. Hierboven onderscheidden we al tussen statische en dynamische factoren. Bij die laatste kan het bijvoorbeeld gaan om psychotische symptomen; die zijn – in beginsel – te behandelen, evenals iemands impulscontrole.[81] Actuariële risicotaxatie-instrumenten zijn echter veelal opgebouwd uit statische factoren.[82] Voorbeelden van statische factoren zijn criminele voorgeschiedenis en leeftijd. Geen behandeling kan die feiten te veranderen. Een probleem is dat statische factoren vaak naar voren komen als krachtige voorspellers van recidive. Met andere woorden: juist de factoren die helpen bij de voorspelling, blijken voor de behandeling geregeld niet van waarde.[83]

Hierbij is het goed om te bedenken dat risicotaxatie-instrumenten factoren gebruiken die risico voorspellen, maar dat deze risicofactoren geen causale factoren hoe-

78. Zo is het in de inleiding genoemde risicotaxatie-instrument OxRec met Nederlandse data gevalideerd, zie Fazel e.a. 2019, p. 841.
79. Zie paragraaf 3.1.
80. Zie bijvoorbeeld F. Doshi-Velez en B. Kim, 'Towards a rigorous science of interpretable machine learning', *arXiv* 2017, 1702.08608. Onderzoekers als Rudin betogen daarentegen dat we niet moeten proberen complexe modellen uit te leggen, maar dat we ons beter kunnen richten op het ontwikkelen van begrijpelijke *machine learning* algoritmen, zie C. Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* (1) 2019/5, p. 206-215.
81. L. Gijs en J. Henriks, 'Adolescenten die seksueel geweld plegen', in: Goethals, Meynen en Popma (red.) 2019, p. 127.
82. De Vogel, De Vries Robbé en Van den Broek 2019.
83. Zo ook J.S. Wormith, 'Automated offender risk assessment. The next generation or a black hole?', *Criminology & Public Policy* (16) 2017/1, p. 293-294; B.L. Spivak en S.M. Shepherd, 'Machine learning and forensic risk assessment: new frontiers', *The Journal of Forensic Psychiatry & Psychology* (31) 2020/4, p. 571-581.

doi: 10.5553/Bsb/266669012021002002004

ven te zijn. Het kan zijn dat een risicofactor samenhangt met een causale factor, maar dat deze causale factor lastiger te meten is, waardoor deze niet goed wordt geïdentificeerd of niet bruikbaar is als risicofactor. Kortom, risicofactoren en de uitkomst van een actuariële risicotaxatie maken niet altijd duidelijk waarop een interventie zich kan of moet richten.[84]

Dit is eveneens bij AI-risicotaxatie het geval. Ook hier kan risicotaxatie plaatsvinden op grond van onveranderlijke factoren en/of factoren die geen oorzakelijke factor hoeven te zijn.[85] Onderscheidend voor AI-risicotaxatie is dat het algoritme in de trainingsdata zelf relevante factoren ontdekt alsmede de relevante samenhang ertussen – dit is precies waar een algoritme zijn 'intelligentie' voor gebruikt.[86] Hierbij kan, zoals besproken, sprake zijn van gebrek aan transparantie: het is onbekend wat het algoritme precies doet. Daarmee kan het (nog) onduidelijker zijn waar de behandeling zich op moet richten. Concluderend: als er een AI-risicotaxatie-instrument is dat beter voorspelt dan bestaande actuariele risicotaxatie maar waarbij niet duidelijk is welke factoren op welke wijze worden gewogen, dan is het moeilijk om een justitiabele met hoog-risico perspectief te geven in een behandeling. Immers, waar moet aan worden gewerkt?[87]

5. Conclusies en slotbeschouwing

In dit artikel hebben we onderzocht in hoeverre AI-risicotaxatie overeenkomt met en verschilt van bestaande actuariële risicotaxatie. De overeenkomst is dat beide methoden statistische verbanden leggen tussen kenmerken van categorieën van individuen en recidive. Een belangrijk verschil tussen bestaande risicotaxatie en AI-risicotaxatie is, ten eerste, dat AI-risicotaxatie meestal geavanceerdere statistische methoden gebruikt. Ten tweede kunnen in potentie veel meer factoren worden meegenomen in de risicotaxatie. Ten derde zijn de algoritmen van AI-risicotaxatie zelflerend: het algoritme vindt zelf op basis van de ingevoerde data de statistische verbanden tussen voorspellende factoren en recidive. Met behulp van algoritmen kunnen nieuwe risicofactoren worden opgespoord. Het is denkbaar dat AI-risicotaxatie in de toekomst tot betere voorspellingen zal leiden.

Zowel bestaande actuariële risicotaxatie als AI-risicotaxatie kennen problemen op het terrein van *biases* en transparantie en bieden niet altijd aanknopingspunten voor een op vermindering van recidiverisico gerichte behandeling. Deze problemen worden deels versterkt door het gebruik van AI, al dan niet in combinatie met big data. Tegelijkertijd biedt AI ook mogelijke oplossingen voor deze problemen.

Het gebruik van AI voor risicotaxatie leidt niet tot fundamenteel andere problemen dan het gebruik van de bestaande risicotaxatie-instrumenten. Wel verleent het (toenemende) gebruik van (AI-)risicotaxatie grotere urgentie aan de beantwoording van reeds langer bestaande vragen over de normering van het preventieve sanctierecht. Wij bespreken drie juridische problemen die samenhangen met de specifieke aard van voorspellingen van recidive. Hiermee willen we een aanzet doen tot een fundamenteel juridisch debat over de grenzen van het preventieve strafrecht.

Ten eerste, het voorspellen van toekomstige feiten verschilt fundamenteel van het verklaren en bewijzen van een strafbaar feit dat in het verleden heeft plaatsgevonden. Bij strafrechtelijk bewijzen dat een verdachte het tenlastegelegde feit heeft begaan, kan een bewezenverklaring worden gebaseerd op specifiek en onderscheidend bewijs, zoals een vingerafdruk, DNA of een getuigenverklaring,[88] al speelt algemene kennis van de wereld – zoals algemene inzichten over de betrouwbaarheid van getuigenverklaringen – hierbij ook een rol. Bij het voorspellen van recidiverisico hebben strafbare feiten zich nog niet voorgedaan en is dergelijk onderscheidend bewijs met betrekking tot zo'n feit per definitie niet voorhanden. Een fundamentele vraag is daarom in hoeverre vrijheidsbenemende en -beperkende sancties op basis van statistiek zonder onderscheidend bewijs toegelaten zijn, terwijl dat in het algemeen onvoldoende wordt geacht voor een veroordeling wegens een strafbaar feit.

Ten tweede, wil een maatregel opgelegd mogen worden, dan is vereist dat recidivegevaar 'aannemelijk' is. Onduidelijk is echter welke bewijsmaatstaf hiermee van toepassing is op de vaststelling van recidivegevaar.[89] Voor de vaststelling van een strafbaar feit geldt de maatstaf van bewijs 'buiten redelijke twijfel'.[90] Deze hoge bewijsmaatstaf drukt uit dat de tolerantie voor onterechte veroordelingen (foutpositieven) zeer beperkt is.[91] Ook voor sancties gebaseerd op recidiverisico is een belangrijke vraag hoe groot de tolerantie voor foutpositieven

84. Zie ook Bijlsma, Bex en Meynen 2019, p. 3318-3319; Smid en Uzieblo 2020, p. 37.
85. De huidige AI-risicotaxatie-instrumenten voorspellen voornamelijk met statische factoren, zie paragraaf 3.1 met verwijzing naar Wormith 2017, p. 293-294.
86. Zie paragraaf 3.1.
87. Zo ook Barabas e.a. 2018.
88. A.R. Mackor, 'Veroordelen met "naakte" statistieken?', *RMThemis* 2019, p. 93-96; A.R. Mackor, 'Different ways of being naked. A scenario approach to the naked statistical evidence problem', *IfCoLog Journal of Logics and their Applications* (te verschijnen). In bayesiaanse termen: 'onderscheidend' houdt in dat de *likelihood ratio* hoger of lager is dan 1. Zie E.K. Cheng, 'Reconceptualizing the burden of proof', *The Yale Law Journal* (122) 2013/5, p. 1254-1279; M. Di Bello, 'Trial by statistics: is a high probability of guilt enough to convict?', *Mind* (128) 2019/512, p. 1045-1084.
89. Mackor 2014; Bijlsma en Meynen 2018.
90. Formeel geldt naar Nederlands recht de maatstaf 'wettig en overtuigend bewijs' (art. 338 Sv). Bemelmans wijst op enkele verschillen tussen beide maatstaven, maar concludeert dat het Nederlandse bewijsrecht een veroordeling bij redelijke twijfel niet toelaat, zie J.H.B. Bemelmans, *Totdat het tegendeel is bewezen. De onschuldpresumptie in rechtshistorisch, theoretisch, internationaalrechtelijk en Nederlands strafprocesrechtelijk perspectief* (diss. Nijmegen), Deventer: Wolters Kluwer 2018, p. 356-361. In art. 14 lid 2 IVBPR wordt deze maatstaf ook gelezen, zie Bemelmans 2018, p. 234-235.
91. Bemelmans 2018, p. 95.

doi: 10.5553/Bsb/266669012021002002004

mag zijn.[92] Het relatief grote aantal foutpositieven bij de voorspelling van recidivegevaar maakt echter dat hanteren van de bewijsmaatstaf 'buiten redelijke twijfel' ertoe zou leiden dat gevaar (vrijwel) nooit bewezen kan worden. Kunnen ten aanzien van de vaststelling van gevaar meer foutpositieven worden aanvaard dan voor het bewijs van een strafbaar feit? Naar analogie van de onschuldpresumptie kan invoering van een 'ongevaarlijkheidspresumptie' worden overwogen en een daarmee samenhangende bewijsmaatstaf die uitdrukking geeft aan de mate waarin foutpositieven aanvaardbaar zijn in het preventieve sanctierecht.[93]

Een derde probleem is dat de wettelijke gevaarscriteria zoals die gelden voor tbs en andere beveiligingsmaatregelen niet duidelijk zijn (afgebakend).[94] Dit is problematisch, want zolang geen heldere materiële criteria voor recidivegevaar bestaan en dus onduidelijk is wat het bewijsthema precies is, kan niet worden vastgesteld of het recidiverisico dat door een bepaald risicotaxatie-instrument wordt gegeven voldoet aan de juridische norm voor gevaar.

Het juridische debat is volgens ons daarom niet alleen gebaat bij een kritische analyse van verschillende vormen van risicotaxatie, maar eerst en vooral bij de systematische doordenking van de normering van het gevaar. Een dergelijke doordenking kan de wetgever en de rechter helpen bij hun afbakening van de grenzen van het preventieve sanctierecht.

92. Bijlsma, Bex en Meynen 2019, p. 3316-3317.
93. Ashworth en Zedner 2014, p. 259; J. Bijlsma, 'Het is tijd voor een ongevaarlijkheidspresumptie. Beschouwing naar aanleiding van J.H.B. Bemelmans, *Totdat het tegendeel is bewezen'*, *RMThemis* 2021, p. 81-92.
94. Struijk en Van der Wolf 2018, p. 938-947; J. Bijlsma e.a., 'Stoornis en gevaar. Een aanzet tot onderzoek naar een alternatief voor tbs', *DD* (25) 2020/5, p. 365-366.

doi: 10.5553/Bsb/266669012021002002004