Zonder gedegen validatie van algoritme geen publiek vertrouwen meer

Het uitvoeren van risicoanalyses lijkt een complex proces, maar deze complexiteit komt van origine voort uit menselijk handelen. Om risico's en negatieve gevolgen, zoals onterechte bias, te voorkomen, dient een onafhankelijke toets plaats te vinden. Hiermee worden de kwaliteit en het werken binnen de wettelijke en ethische kaders gewaarborgd. KPMG heeft een methode ontwikkeld voor het uitvoeren van onafhankelijke modelvalidaties. Een uniform toetsingskader vormt hierbij de basis, waarbij een onderscheid wordt gemaakt tussen vier verschillende aspecten. Onderdeel van elke modelvalidatie is ook het uitvoeren van technische testen. De uit te voeren werkzaamheden en technische testen conform het toetsingskader resulteren vervolgens in observaties, bevindingen en aanbevelingen, die de modelvalidator rapporteert aan de organisatie.

INLEIDING

De analyse van risico's door het gebruik van algoritmes in risicomodellen blijft de gemoederen wereldwijd bezighouden. Inhoudelijk lijken discussies langs elkaar heen te lopen. Alleen al over de definitie van een risicomodel lopen de meningen uiteen, evenals over de vraag wanneer algoritmes moeten worden gebruikt en wat geautomatiseerde besluitvorming is.

In de praktijk komen de problemen die gepaard gaan met het gebruik van risicomodellen regelmatig in het nieuws. Voorbeelden zijn er te over, zoals het frauderisicosysteem (FSV) dat de Belastingdienst gebruikte. Algoritmes worden vaak aangeduid als 'black boxes', omdat niet zichtbaar zou zijn waarom een risicomodel een bepaalde uitkomst toont. Deze perceptie voedt – logischerwijs – verdere discussie en een wantrouwen tegen algoritmegebruik.

In rechterlijke uitspraken zien we terugkomen dat risicomodellen regelmatig onvoldoende inzichtelijk zijn. In een zaak omtrent het Systeem Risico Indicatie (SyRI), een wettelijk instrument voor fraudebestrijding, oordeelde de rechter dat de doelstellingen van SyRI niet in de juiste verhouding staan tot de bijkomende schending van de privacy. Bovendien is een dergelijk model volgens de rechter onvoldoende transparant ([Rech20]). Risicomodellen en algoritmes lijken zo complex te zijn dat het lastig is om de interne logica ervan te volgen.

COMPLEXE MATERIE

Maar waar gaat het mis? Het is niet het algoritme of het risicomodel dat per definitie 'fout' is. Het menselijk handelen, zoals bij de keuze voor inputdata of bij het onderzoeken van gegenereerde signalen, kan ook problemen veroorzaken. Als een risicomodel ongrijpbaar of complex is, heeft dat vaak meer te maken met de opzet en inrichting van een risicomodel – door mensen.

> Een risicomodel an sich is niet per se ongrijpbaar of complex

Patrick Özer MSc
is partner bij KPMG Forensic Technology.

Mr. drs. Katja den Bieman RA
is senior manager bij KPMG Forensic Investigations.

Koen van der Krogt MSc
is manager bij KPMG Forensic Technology.

Inputdata zijn data die als input aan een model worden geleverd. Op basis daarvan gaat het model een uitkomst genereren. Om het model te leren of te 'trainen' wordt gebruikgemaakt van bestaande inputdata, waar mogelijk inclusief de uitkomst; dit zijn de trainingsdata. Ten slotte wordt ook een deel van de reeds bestaande data apart gehouden om de effectiviteit van het model te toetsen met andere data dan die zijn gebruikt voor de training van het model.

Het ontwikkelen, beheren en implementeren van een risicomodel binnen technische, wettelijke en ethische kaders is een proces dat niet over een nacht ijs gaat. (H)erkenning van de complexiteit van dit proces is *key*, evenals identificeren waar die complexiteit uit voortkomt. Dit is een belangrijke stap om te bepalen hoe met een risicomodel moet worden omgegaan. Het is dan ook begrijpelijk dat het toepassen en controleren van risicomodellen en algoritmes een uitdaging vormt.

UIT DE PRAKTIJK GEGREPEN

In de praktijk zijn meerdere voorbeelden te vinden waar het toepassen van risicomodellen misgaat. Hieronder is een aantal toegelicht.

Bij het eerder aangehaalde voorbeeld rondom het FSV gebruikte de Belastingdienst data over de dubbele nationaliteit van Nederlanders als indicator in het systeem dat bepaalde aanvragen voor kinderopvangtoeslag geautomatiseerd als risicovol labelde. Zowel KPMG als de Autoriteit Persoonsgegevens stelde na onderzoek vast dat de Belastingdienst deze gevoelige data onterecht jarenlang bewaarde en onrechtmatig gebruikmaakte van een 'zwarte lijst' ([KPMG20], [Auto20], [Auto21]).

Een tweede voorbeeld betreft de gemeente Rotterdam, die een algoritme gebruikt met als doel personen een risicoscore toe te wijzen om bijstandsfraude te voorspellen. De Rekenkamer Rotterdam waarschuwde in december 2021 dat er mogelijk gebruik werd gemaakt van inputfactoren als geslacht en woonwijk voor het bepalen van dit risico. Het risicomodel zou hiermee onethisch kunnen zijn. Als reactie op het onderzoek van de Rekenkamer diende SP-lid Renske Leijten Kamervragen in over het gebruik van algoritmes door lokale overheden. Minister van Binnenlandse Zaken en Koninkrijkrelaties Hanke Bruins Slot besloot op centraal niveau geen onderzoek te laten uitvoeren, mede gezien het feit dat de verantwoordelijkheid tot onderzoek bij de gemeentes zelf ligt, niet bij de rijksoverheid ([Team22]). In het door de gemeente Rotterdam gepubliceerde algoritmeregister wordt het risico-inschattingsmodel omtrent uitkeringsonrechtmatigheid nog steeds benoemd. Hierbij wordt gesteld dat het gebruikte algoritme geen data verwerkt die kunnen leiden tot discriminatie ([Geme22]).

Naast voorbeelden uit Nederland zijn er internationale voorbeelden. Ook internationale overheidsorganen en rechterlijke instellingen gebruiken risicomodellen. In de Verenigde Staten werd onder president Trump in 2018 de 'First Step Act' ingevoerd. Het doel van deze wet was het inkorten van onnodig lange straffen. Op basis van risicomodel PATTERN worden gevangenen in de gelegenheid gesteld mogelijk eerder de gevangenis te verlaten als zij een lage kans hebben om weer terug te vallen in crimineel gedrag. Burgerrechtengroeperingen uitten al snel hun zorgen over mogelijke onevenredigheden gebaseerd op ras. Het algoritme zou de kans op het opnieuw vervallen in crimineel gedrag aanzienlijk hoger inschatten als iemand bijvoorbeeld een Afro-Amerikaanse, Spaanse of Aziatische achtergrond heeft ([FBP22]). Het betrekken van iemands criminele historie als risicofactor kan problematisch zijn omdat in de VS etnisch profileren een bekend issue is. Het aanvullend betrekken van educatie als risicofactor kan een indirect versterkend effect hebben. Het risicomodel wordt tot op heden gehanteerd en is terug te vinden op de site van het Federal Bureau of Prisons ([John22]).

Bovengenoemde voorbeelden geven een beeld van de risico's en ongewenste gevolgen bij het gebruik van (discutabele) algoritmes in risicomodellen. De voorbeelden tonen daarmee ook de noodzaak om dergelijke risicomodellen onafhankelijk te valideren, onder andere om zorgvuldigheid in het gebruik van algoritmes en voorspellende modellen af te dwingen.

KENNIS SAMENBRENGEN

Inbreng van voldoende inhoudelijke (domein)kennis is een vereiste om het toepassen en valideren van risicomodellen en algoritmes goed vorm te geven. Deze kennis heeft een aantal aspecten:

- Domeinkennis over het toepassingsgebied is nodig om te bepalen of het doel van het risicomodel haalbaar is en hoe het doel kan worden bereikt. Daarnaast is domeinkennis vereist om de benodigde data, hypothesen en randvoorwaarden te bepalen, en om uiteindelijk te verifiëren of de uitkomsten van het model juist zijn.
- Technische kennis is nodig voor het ontwikkelen van een algoritme voor een risicomodel, voor het bepalen van het meest geschikte algoritme voor het risicomodel, voor het analyseren van de vaak grote hoeveelheden data en voor het uiteindelijke programmeren.
- Juridische kennis is nodig over de wettelijke kaders die direct van toepassing zijn op het model en het gebruik van het model, alsmede over de wettelijke

kaders en richtlijnen op het gebied van dataprivacy en mensenrechten. In het verlengde hiervan zijn de ethische kaders relevant met betrekking tot het voorkomen van mogelijke discriminatie of vooringenomenheid ('bias').

De samenkomst van al deze kennis bij de ontwikkeling en het beheer van een model of algoritme is van cruciaal belang om te bepalen of een model mag worden toegepast.

BIAS IN RISICOMODELLEN

Bias en het voorkomen van discriminatie is een veelbesproken onderwerp bij de toepassing van risicomodellen en algoritmes. *Bias* is een vertekening van onderzoeksresultaten, een vooroordeel dat een objectieve waarneming of beoordeling in de weg kan staan. Een ander begrip dat hier ook vaak onder wordt geschaard is *vooringenomenheid*, waarbij men al een oordeel klaar heeft zonder alle feiten te hebben onderzocht. De maatschappij is terecht kritisch ten aanzien van dit onderwerp. Een belangrijk punt om rekening mee te houden in de publieke discussie en tijdens de ontwikkeling van modellen is dat bias altijd in bepaalde mate onderdeel is van een risicomodel of algoritme. Dit komt doordat bias niet altijd direct hoeft te zijn, maar ook indirect kan zijn. Indirecte bias betekent dat een kenmerk dat zelf geen directe bias bevat, samenhangt met een kenmerk dat wel bias bevat. De lengte van een cv hangt bijvoorbeeld samen met de leeftijd van een persoon, en de postcode van iemands huisadres kan verband houden met zijn of haar opleidingsniveau, etniciteit of leeftijd. Voor bijna elk kenmerk is er wel een verwant kenmerk te bedenken waarbij er sprake is van bias. Dit betekent echter niet dat risicomodellen per definitie discriminerend zijn, maar wel dat moet worden stilgestaan bij mogelijke bias en de ethiek die komt kijken bij de toepassing van het model en het effect van bias. De eerste afweging hiervoor zit al in de stap van het bepalen van welke criteria al dan niet een plek in een risicomodel gaan krijgen. Het is van belang deze directe en indirecte bias tijdens de ontwikkeling van een risicomodel te onderkennen en vast te leggen. Dan kunnen de invloed en het gebruik van deze informatie in een algoritme bewust worden geaccepteerd of gemitigeerd.

RISICOMODELLEN TOETSEN

Om ervoor te zorgen dat bij de ontwikkeling de juiste kennis wordt gebruikt en dat voldoende rekening wordt gehouden met bias, is het van belang goede kaders vast te stellen waarbinnen moet worden geopereerd. Het vaststellen van de kaders van een risicomodel of algoritme is een belangrijke stap om te kunnen beoordelen of een model of algoritme kan worden toegepast en later (nog steeds) passend is. Voor deze beoordeling is het waardevol om een nieuwe set ogen met een frisse blik naar het risicomodel te laten kijken. Een onafhankelijke toets van een risicomodel of algoritme ter onderbouwing van het oordeel is een goed instrument om een solide besluit te nemen om een risicomodel te gaan gebruiken.

KADERS VOOR TOETSEN

De te toetsen kaders kunnen harde eisen bevatten, zoals 'Is er een Data Privacy Impact Assessment (DPIA) uitgevoerd?' of 'Zijn alle gebruikte persoonsgegevens vastgelegd?', maar ook minder rigide eisen, zoals 'Is het gekozen type algoritme passend voor ons specifieke doel?' Om deze kaders te toetsen is een 'professional judgement' van een auditor of validator vereist. Dit maakt een onafhankelijke toets, ook wel kwaliteitscontrole of modelvalidatie genoemd, complex maar minstens zo belangrijk als het ontwikkelproces van het risicomodel. Daarom is het van belang dat een auditor of validator kennis heeft van en ervaring met het ontwikkelen en controleren van modellen en algoritmes, het toepassingsgebied en de juridische en ethische kaders.

Ook de kaders voor de toepassing of de controle van risicomodellen of algoritmes zijn onderwerp van maatschappelijke discussie. Belangrijke 'richtlijnen' die hieruit naar voren zijn gekomen:

- In 2021 heeft het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties het Impact Assessment voor Mensenrechten bij de inzet van Algoritmes (IAMA) uitgebracht ([MBZK21]). Hierin wordt stilgestaan bij de keuze voor het toepassen van algoritmes en de verantwoorde ontwikkeling en implementatie.
- De Algemene Rekenkamer heeft in 2021 een toetsingskader ([Alge21]) uitgebracht voor kwaliteitscon-

> Een onafhankelijke toets moet plaatsvinden om de kwaliteit van een risicomodel te waarborgen

trole van algoritmes en heeft deze toegepast op negen algoritmes bij de overheid ([Alge22]).
- Het NIST werkt aan een AI Risk Management Framework, dat eind 2022, begin 2023 zal worden gepubliceerd. Een conceptversie is reeds beschikbaar ([NIST22]).
- NOREA heeft ook een set van principes gepubliceerd voor het onderzoeken van algoritmes ([NORE21]).
- Vanuit de VS zijn er richtlijnen van het Office of the Comptroller of the Currency (OCC) over Model Risk Management ([OCC21]), die al vele jaren in de financiele wereld op financiële- en compliancerisicomodellen worden toegepast.

Deze stukken leveren waardevolle informatie voor het vaststellen van kaders voor de ontwikkeling van modellen en algoritmes en de controle of validatie daarvan.

METHODE VOOR MODELVALIDATIE

Wij zijn van mening dat een onafhankelijke toets moet plaatsvinden om de kwaliteit en het volgen van wettelijke en ethische kaders van een risicomodel en het beoogd gebruik ervan te waarborgen. Daarom heeft KPMG een methode ontwikkeld voor het uitvoeren van onafhankelijke modelvalidaties, op basis van de benoemde bestaande kaders en opgedane kennis en ervaring in het uitvoeren van modelvalidaties in verschillende sectoren.

In een modelvalidatie dient gekeken te worden naar verschillende aspecten die samen de context van het risicomodel vormen. De governance van een risicomodel geeft inzicht in ingerichte rollen en verantwoordelijkheden. Als onderdeel van de governance dient ook de doelstelling van het risicomodel – in het grotere geheel van een organisatie – te zijn vastgesteld, want een risicomodel hoort immers niet geheel op zichzelf te staan. Op basis van deze inrichting heeft een organisatie het concept van het risicomodel ontwikkeld en dit vervat in een technisch ontwerp. Uit het conceptuele model en het technische ontwerp moet blijken of de opzet conform de initiële doelstelling is. De vervolgstap is om te analyseren of niet alleen de opzet maar ook de (technische) werking van het risicomodel past bij de doelstelling die is geformuleerd. Het kan bijvoorbeeld voorkomen dat een risicomodel wel een voorspellende waarde heeft, maar vanuit een andere hypothese dan de bedoeling was. Als het risicomodel werkt zoals beoogd, is het laatste onderdeel de (periodieke) evaluatie en het afleggen van verantwoording over het risicomodel. Ook hierbij is het van belang te onderkennen dat een risicomodel in zijn context moet passen en dat rekening wordt gehouden met veranderende omstandigheden zoals nieuwe wet- en regelgeving of door de tijd heen veranderende inputdata, zogenaamde 'data drift'.

Op basis van het bovenstaande hebben we een toetsingskader ingericht dat toeziet op vier aspecten:
1. governance en inrichting;
2. conceptueel model en technisch ontwerp;
3. werking van risicomodel;
4. evaluatie en verantwoording.

Hieronder volgt een meer gedetailleerde beschrijving van deze aspecten. Om het concreet te maken zijn tevens voorbeelden van uit te voeren werkzaamheden opgenomen voor het specifieke aspect.

GOVERNANCE EN INRICHTING

De ontwikkeling van een risicomodel wordt in het aspect 'Governance en inrichting' getoetst aan relevante wet- en regelgeving en andere vigerende normenkaders. Daarbij wordt onder andere bekeken of de doelstelling van het risicomodel duidelijk is geformuleerd en of binnen de relevante wettelijke en ethische kaders is gewerkt. Dit is de basis van het risicomodel. Tevens is relevant of het risicomodel zijn doel niet voorbijschiet en of het ontwikkelen van een risicomodel wel past bij de doelstelling die het moet nastreven. Zijn proportionaliteit (het gebruikte middel moet in verhouding staan tot het te bereiken doel) en subsidiariteit (het gebruikte middel moet het minst zware middel zijn om het doel te bereiken) voldoende afgewogen? Met andere woorden, is het doel van het risicomodel ook op een effectieve manier te bereiken met risicoselectie met minder persoonsgegevens of met minder op de privacy ingrijpende of alternatieve middelen? Daarnaast is inrichting van de governancestructuur voor het risicomodel van belang. Het moet helder zijn wie welke verantwoordelijkheden draagt vanuit welke rol op verschillende lagen binnen de organisatie; van projectteam tot management. Een vraag die de modelvalidator bijvoorbeeld dient te beantwoorden is of er voldoende kennis van zaken aanwezig is in de organisatie vanuit de vastgelegde rollen en verantwoordelijkheden, waarbij het aantal jaren relevante ervaring of relevante opleiding een rol speelt. Er kan dan bijvoorbeeld worden geconcludeerd dat er voldoende of juist te weinig expertise aanwezig is in het team dat zich bezighoudt met het risicomodel. Het is tevens van belang dat er voorafgaand aan de ontwikkeling een gedegen risicoanalyse is uitgevoerd.

De modelvalidator kijkt bij dit aspect ook naar het DPIA. Dit is relevant om inzicht te krijgen in binnen de organisatie bestaande evaluaties rondom geïdentificeerde en beschreven wettelijke richtlijnen en in het voorkomen van bias en discriminatie. De afwegingen omtrent de ethische kaders en de risicoanalyse dienen te zijn vastgelegd tijdens de ontwikkeling van een risicomodel.

Een voorbeeld van het niet voldoen aan wettelijke kaders inzake DPIA's, uit maart 2022, is een geschil tussen het door pakketleverancier USPS beheerde Internet Covert Operations Program (iCOP) en het Electronic Privacy Information Center (EPIC). iCOP zette gezichtsherkenning in om tijdens het monitoren van social media posts mogelijke dreigingen te identificeren. Het gebruik van soortgelijke gezichtsherkenning brengt grote risico's en ethische bezwaren met zich mee en daarnaast zou het gebruikte programma volgens EPIC illegaal zijn wegens het gebrek aan een DPIA ([Hawk22]).

CONCEPTUEEL MODEL EN TECHNISCH ONTWERP

Inzake het aspect 'Conceptueel model en technisch ontwerp' voert de modelvalidator activiteiten uit om te valideren dat het conceptuele model en het technische ontwerp aansluiten bij de doelstelling van het risicomodel en de bijbehorende kaders zoals geïdentificeerd in het aspect 'Governance en inrichting'. Hierbij analyseert de modelvalidator de beschikbaarheid en kwaliteit van de beschrijving van het conceptuele model en het technische ontwerp in onder andere documentatie. De modelvalidator toetst onder andere de duidelijkheid en uitlegbaarheid van de beschrijving en onderliggende hypothesen.

Onder dit aspect valt ook het identificeren van een beschrijving en onderbouwing van de aannames en keuzes die ten grondslag liggen aan het technische ontwerp. Belangrijk is of er voldoende sprake is van onderbouwing voor de gekozen trainingsdata, algoritmes en programmeertaal voor de ontwikkeling van het risicomodel.

Voor dit aspect is een relevant onderdeel het gebruik van variabelen met indirecte bias. Een voorbeeld waaruit blijkt dat niet altijd voldoende aandacht is geschonken aan het voorkomen van (indirecte) bias, heeft betrekking op de slagingskans van sollicitanten bij e-commercebedrijf Amazon. In 2015 kwamen machine-learningspecialisten bij Amazon erachter dat het ontwikkelde algoritme voor het selecteren van nieuwe kandidaten bevooroordeeld was. Het algoritme observeerde patronen in ingediende cv's over een periode van tien jaar, en gezien de mannelijke dominantie in de technologische industrie, leerde het algoritme zichzelf om cv's van mannen voor te trekken – op basis van historie. Als een cv woorden als 'women' of 'women's chess club captain' bevatte, werd de kandidaat van de lijst met kandidaten geschrapt ([Dast18], [Logi19]).

WERKING VAN RISICOMODEL

Werkzaamheden onder het toetsingsaspect 'Werking van risicomodel' zijn gericht op toetsen of de werking van het risicomodel aansluit bij de doelstelling en het ontwerp van het risicomodel uit de vorige twee aspecten. Is het ontwerp op de juiste wijze doorvertaald in de technische implementatie en levert het risicomodel daarmee de resultaten op zoals beoogd?

De modelvalidator analyseert onder andere of bij de ontwikkeling van een risicomodel technische testen zijn uitgevoerd om de werking en output van het model te beoordelen. Hierbij dient de modelvalidator te analyseren welke testen zijn uitgevoerd en wat de daaruit voortgekomen resultaten zijn (geweest). Daarnaast wordt gekeken of er testen zijn uitgevoerd om de datakwaliteit van de inputdata te controleren en of sprake is van gedegen versiebeheer. De modelvalidator voert tevens onafhankelijk technische toetsen uit (zie de subparagraaf hierna).

Tevens relevant bij dit aspect is een toets op de beschrijving van het geprogrammeerde risicomodel. Hierbij moet ten minste een aantal punten voldoende zijn toegelicht, zoals een overzicht van de gebruikte input, de werking van het risicomodel, het toegepaste algoritme, eventuele onzekerheden en beperkingen, en een toelichting waarom het risicomodel geschikt is voor het beoogde gebruik.

Een mogelijke bevinding van de modelvalidator ten aanzien van dit aspect kan betrekking hebben op de kwaliteit van de gebruikte data. Wanneer de trainingsdata van slechte kwaliteit zijn, zal dit waarschijnlijk betekenen dat de kwaliteit van de genereerde resultaten tevens onvoldoende sterk is. Een voorbeeld hiervan zien we bij een door de NS gebruikt algoritme. Een klant kon geen abonnement afsluiten doordat de postcode van aanvragers onder meer werd gebruikt als input voor hun kredietcheck. Als, zoals in dit geval, bleek dat een voormalige bewoner van het adres een wanbetaler was, kreeg de klant een negatieve kredietscore en werd de aanvraag geweigerd ([Voll22]).

Technische toetsen

Onderdeel van onze methode is onder andere het technisch toetsen van het risicomodel. Een van de technische toetsen die we inzetten om te valideren of de technische implementatie van het risicomodel aansluit bij het conceptuele model en het technische ontwerp, is het doen van een codereview. Hierbij wordt – deels geautomatiseerd en deels handmatig – de geprogrammeerde code in detail beoordeeld en opnieuw 'uitgevoerd' om na te gaan of deze aansluit bij het ontwerp en daarmee de beoogde resultaten genereert.

Daarnaast kunnen stabiliteitstoetsen worden uitgevoerd. De modelvalidator voert hierbij de geprogrammeerde code van het risicomodel meerdere keren opnieuw uit met dezelfde en minimaal aangepaste inputdata om de impact hiervan op de resultaten te analyseren. Ook kunnen de inputdata worden aangepast met extreme waarden om de impact en adequate afhandeling van deze extreme waarden op het risicomodel te bepalen. Het doel hiervan is om na te gaan of het risicomodel voldoende stabiel is.

Ten slotte kunnen prestatietoetsen worden uitgevoerd om na te gaan of het model in voldoende mate effectief is in het 'voorspellen' van de beoogde uitkomsten.

EVALUATIE EN VERANTWOORDING

Onder het aspect 'Evaluatie en verantwoording' toetst de modelvalidator specifiek dat het model wordt gebruikt conform doelstelling en richtlijnen en toetst de modelvalidator het ingerichte evaluatiemechanisme van de organisatie, waarbij het tevens belangrijk is te analyseren hoe verantwoording wordt afgelegd. Dit aspect is het sluitstuk van de eerdere drie aspecten en kijkt overkoepelend terug op de verschillende eerder beschreven elementen.

Onderdeel van dit aspect is een 'lekentoets': zijn de resultaten van het risicomodel logisch, bezien vanuit een relatief ongeïnformeerde persoon? Dit ziet toe op de logica van de uitkomsten van het risicomodel voor niet-betrokkenen en dient als spiegel. De modelvalidator analyseert of de gegenereerde resultaten te verwachten zijn op basis van logische redenen, wat mogelijke verklaringen kunnen zijn voor afwijkingen hierop en of de uitkomsten uitlegbaar zijn. Het is een toets vanuit een helikopterblik: is, alles overziend, het risicomodel voldoende begrijpelijk?

Binnen dit aspect beziet de modelvalidatie tegelijkertijd of (in de documentatie) eventuele aanbevelingen uit uitgevoerde toetsen en eerdere evaluaties zijn opgevolgd en of het (toekomstige) gebruik van het risicomodel staat beschreven.

Een bestaand risicomodel waar de waarborging van het beoogde gebruik te betwisten valt, betreft de algoritmes die schuilgaan achter de bepaling van de kredietwaarde van individuen bij banken in de VS. De impact van een achterstallige betaling op de kredietwaarde is groter voor klanten met een hogere (in de regel 'betere') kredietwaarde dan voor klanten met een lagere kredietwaarde. Dit effect wordt veroorzaakt door het achterliggende algoritme: wanneer iemand een hogere score heeft, is deze score sensitiever voor negatieve gebeurtenissen, waardoor dezelfde gebeurtenis een andere impact kan hebben op verschillende individuen ([Sing22]).

KAN EEN RISICOMODEL GEBRUIKT WORDEN?

Ons uitgangspunt bij elke modelvalidatie is een uniform toetsingskader op basis van bovenstaande aspecten. Het toetsingskader vormt dus de basis voor de uitvoering van de modelvalidatie. Hierbij is het van belang dat de modelvalidator een professioneel-kritische instelling behoudt.

Op basis van het toetsingskader heeft de modelvalidator observaties en bevindingen om te rapporteren aan de organisatie. Hierbij kan de modelvalidator aanbevelingen doen, indien nodig. De rapportage van de modelvalidator is een gedegen basis voor het bepalen of een risicomodel in gebruik kan en mag worden genomen.

CONCLUSIE

Het draait allemaal dus, zoals vaak, om de verhouding tussen mens en techniek. Heeft de ontwikkelaar een goed proces opgezet om tot een goed risicomodel te komen? Is er een gedegen risicoanalyse, die breder is dan alleen een DPIA, uitgevoerd voorafgaand aan de ontwikkeling van het risicomodel? En sluit vervolgens de techniek aan bij het ontwerp, door de mens, en levert het risicomodel de beoogde resultaten op? Deze vragen en overige belangrijke vragen die opkomen tijdens het proces om tot een risicomodel te komen, en het (technische) risicomodel zelf, zijn verwerkt in een onafhankelijke toets.

Of het nu een onafhankelijke toets, kwaliteitsreview of modelvalidatie wordt genoemd, het beestje moet een naam hebben. KPMG heeft deze methode ontwikkeld om gestructureerd risicomodellen te kunnen valideren, van het ontstaansproces, de techniek, tot de uitkomsten. De methode is ontwikkeld op basis van het IAMA, het toetsingskader van de Algemene Rekenkamer, de OCC Model Risk Management-richtlijn en meer dan tien jaar ervaring met het uitvoeren van modelvalidaties in verschillende sectoren. KPMG maakt inmiddels vier jaar gebruik van deze methode.

Helaas functioneren risicomodellen, zoals aan het begin van dit artikel is opgemerkt, niet altijd zoals bedoeld. Dit schaadt het maatschappelijk vertrouwen dermate dat het voordeel van het gebruik van een dergelijk risicomodel, namelijk verhoogde effectiviteit en efficiency, snel teniet wordt gedaan. Dit ondanks het feit dat het risicomodel juist zou moeten bijdragen aan het maatschappelijk vertrouwen, mede omdat het objectiever zou moeten zijn dan een volledig handmatige controle. De techniek van een risicomodel of algoritme is echter ook gebaseerd op menselijk handelen, en het is goed dat er gedegen *checks and balances* worden ingebouwd om te komen tot een zo betrouwbaar mogelijke prestatie van mens en techniek samen.

Literatuur

[Alge21] Algemene Rekenkamer (2021, 26 januari). *Aandacht voor algoritmes.* Geraadpleegd op: https://www.rekenkamer.nl/publicaties/rapporten/2021/01/26/aandacht-voor-algoritmes

[Alge22] Algemene Rekenkamer (2022, 18 mei). *Algoritmes getoetst.* Geraadpleegd op: https://www.rekenkamer.nl/publicaties/rapporten/2022/05/18/algoritmes-getoetst

[Auto20] Autoriteit Persoonsgegevens (2020, 17 juli). Werkwijze Belastingdienst in strijd met de wet en discriminerend. Geraadpleegd op: https://autoriteitpersoonsgegevens.nl/nl/nieuws/werkwijze-belastingdienst-strijd-met-de-wet-en-discriminerend

[Auto21] Autoriteit Persoonsgegevens (2021, 7 december). Boete Belastingdienst voor discriminerende en onrechtmatige werkwijze. Geraadpleegd op: https://www.autoriteitpersoonsgegevens.nl/nl/nieuws/boete-belastingdienst-voor-discriminerende-en-onrechtmatige-werkwijze#:%7E:text=In%20de%20zomer%20van%202020,de%20bestrijding%20van%20georganiseerde%20fraude.

[Boer21] Boer, A., & Van Meel, M. (2021). Algoritmes en mensen moeten blijven leren. KPMG. Geraadpleegd op: https://home.kpmg/nl/nl/home/insights/2021/06/algoritmes-en-mensen-moeten-blijven-leren.html

[Dast18] Dastin, J. (2018, 11 oktober). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.* Geraadpleegd op: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[FBP22] Federal Bureau of Prisons (z.j.). BOP: First Step Act, Resources. Geraadpleegd op: https://www.bop.gov/inmates/fsa/pattern.jsp

[Geme22] Gemeente Rotterdam (2022). Algoritmeregister. Geraadpleegd op: https://www.rotterdam.nl/bestuur-organisatie/algoritmeregister/

[Hawk22] Hawkins, S. (2022, 29 maart). USPS Escapes Claims Over Its Facial Recognition Technology. *Bloomberg Law.* Geraadpleegd op: https://news.bloomberglaw.com/tech-and-telecom-law/usps-escapes-claims-over-its-facial-recognition-technology

[Hiji22] Hijink, M. (2022, 12 januari). Wie zet z'n tanden in de foute algoritmes? *NRC.* Geraadpleegd op: https://www.nrc.nl/nieuws/2022/01/12/wie-zet-zn-tanden-in-de-foute-algoritmes-a4077972

[John22] Johnson, C. (2022, 26 januari). Flaws plague a tool meant to help low-risk federal prisoners win early release. *NPR.* Geraadpleegd op: https://choice.npr.org/index.html?origin=https://www.npr.org/2022/01/26/1075509175/justice-department-algorithm-first-step-act?t=1650375204324&t=1652358333950

[KPMG20] KPMG Advisory N.V (2020, 10 juli). *Rapportage verwerking van risicosignalen voor toezicht: Belastingdienst.* Geraadpleegd op: https://open.overheid.nl/repository/ronl-42970d17-d8b9-41d6-aa8f-docc52ab97c8/1/pdf/kpmg-rapport-fsv-onderzoek-belastingdienst.pdf

[Logi19] Logically (2019, 30 juli). 5 Examples of Biased Artificial Intelligence. Geraadpleegd op: https://www.logically.ai/articles/5-examples-of-biased-ai

[MBZK21] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2021, juli). *Impact Assessment Mensenrechten en Algoritmes.* Geraadpleegd op: https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes

[NIST22] National Institute of Standards and Technology (2022). NIST Artificial Intelligence Risk Management Framework (AI RMF). Geraadpleegd op: https://www.nist.gov/itl/ai-risk-management-framework

[NORE21] NOREA (2021, december). *NOREA Guiding Principles Trustworthy AI Investigations.* Geraadpleegd op: https://www.norea.nl/download/?id=10966

[OCC21] Office of the Comptroller of the Currency (2021, 18 augustus). Model Risk Management: New Comptroller's Handbook Booklet. *OCC Bulletin* 2021-39. Geraadpleegd op: https://www.occ.treas.gov/news-issuances/bulletins/2021/bulletin-2021-39.html

[Pols21] Pols, M. (2021, 15 december). Privacywaakhond AP krijgt nieuwe taak als algoritmetoezichthouder en meer geld. *FD.nl.* Geraadpleegd op: https://fd.nl/bedrijfsleven/1424029/privacywaakhond-ap-krijgt-nieuwe-taak-als-algoritmetoezichthouder-en-meer-geld-nle2cacLRhoP

[Rech20] De Rechtspraak (2020, 5 februari). SyRI-wetgeving in strijd met het Europees Verdrag voor de Rechten voor de Mens. Geraadpleegd op: https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-wetgeving-in-strijd-met-het-Europees-Verdrag-voor-de-Rechten-voor-de-Mens.aspx

[Sing22] Singer, M. (2022, 22 maart). How To Guard Against Sudden, Unexpected Drops In Your Credit Score. *Forbes.* Geraadpleegd op: https://www.forbes.com/sites/theyec/2022/03/22/how-to-guard-against-sudden-unexpected-drops-in-your-credit-score/?sh=2f89893e6bf5

[Team22] Team Stadszaken.nl (2022, 21 maart). BZK: gemeenten moeten zelf bepalen hoe ze met algoritmes omgaan. *Stadszaken.nl.* Geraadpleegd op: https://stadszaken.nl/artikel/4169/bzk-gemeenten-moeten-zelf-bepalen-hoe-ze-met-algoritmes-omgaan

[Voll22] Vollebregt, B. (2022, 18 januari). Zo werd Myrthe Reuver de dupe van haar data: 'Ze geloven in eerste instantie het systeem'. *Trouw.* Geraadpleegd op: https://www.trouw.nl/economie/zo-werd-myrthe-reuver-de-dupe-van-haar-data-ze-geloven-in-eerste-instantie-het-systeem~b17af53f/

[VVD21] VVD, D66, CDA en ChristenUnie (2021, 15 december). *Omzien naar elkaar, vooruitkijken naar de toekomst: Coalitieakkoord 2021 – 2025.* Geraadpleegd op: https://www.kabinetsformatie2021.nl/documenten/publicaties/2021/12/15/coalitieakkoord-omzien-naar-elkaar-vooruitkijken-naar-de-toekomst

Over de auteurs

Patrick Özer MSc is partner bij KPMG Forensic Technology en is verantwoordelijk voor het Forensic Technology-team. Hij houdt zich bezig met fraude- en compliancegerelateerde projecten in verschillende sectoren en richt zich daarbij op het uitvoeren van digitaal forensische onderzoeken en het bestrijden van financieel-economische criminaliteit. Als onderdeel daarvan focust hij zich op het valideren van compliancemodellen en risicoscans.

Mr. drs. Katja den Bieman RA is senior manager bij KPMG Forensic Investigations. Zij houdt zich als forensisch accountant voornamelijk bezig met fraudepreventie vanuit frauderisicomanagement en met frauderespons vanuit (persoonsgerichte) feitenonderzoeken naar financiële administraties, processen en gedragingen. Zij heeft zich gespecialiseerd in forensische dienstverlening aan de publieke sector.

Koen van der Krogt MSc is manager bij KPMG Forensic Technology. Hij houdt zich bezig met digitaal onderzoek en (forensische) data-analyse en is gespecialiseerd in het valideren van modellen en systemen ten behoeve van compliance. Hij helpt zijn klanten om snel inzicht te krijgen in grote hoeveelheden data en faciliteert effectieve reviews en onderzoeken.