Thinking outside-the-black-box

Algoritme en audit

22 juli 2021 Emil van Werven, Gerke Roorda en Pim Smulders

(Publicatiedatum: 22 juli 2021)

Sinds enkele jaren verschijnen regelmatig publicaties over algoritmes en audit. Er bestaat echter nog geen breed geaccepteerde en concrete aanpak voor deze categorie van audits. In dit artikel willen wij bijdragen aan de discussie op dit terrein door te delen hoe wij tot een algemeen bruikbaar raamwerk zijn gekomen.

In de herfst van 2019 waren wij aanwezig op de IT-auditorsdag van NOREA. De aanwezigen werd gevraagd wie van hen al assurance bij een algoritme had verstrekt. Nauwelijks gingen er handen omhoog. Niet vreemd natuurlijk voor zo'n nieuw onderwerp in ons vakgebied. Wel zien we dat NOREA bezig is het nieuwe terrein te verkennen, onder meer via de kennisgroep Algorithm Assurance. Deze kennisgroep heeft in maart 2021 een consultatieversie van een document met uitgangspunten (*guiding principles*) voor onderzoek naar algoritmische systemen gepubliceerd. [BOER21]

Op dit moment bestaat nog geen eenduidig normenstelsel om het gebruik van algoritmes aan te toetsen – de guiding principles van de kennisgroep Algorithm Assurance vormen een eerste stap in deze richting. Ook bezit de gemiddelde IT-auditor nu niet voldoende technische kennis om op eigen kracht normen voor audits op dit terrein te kunnen opstellen. Tegelijkertijd horen we steeds vaker de roep om algorithm assurance. Daarom delen we in dit artikel het resultaat van onze zoektocht naar een aanpak voor audits bij algoritmes.

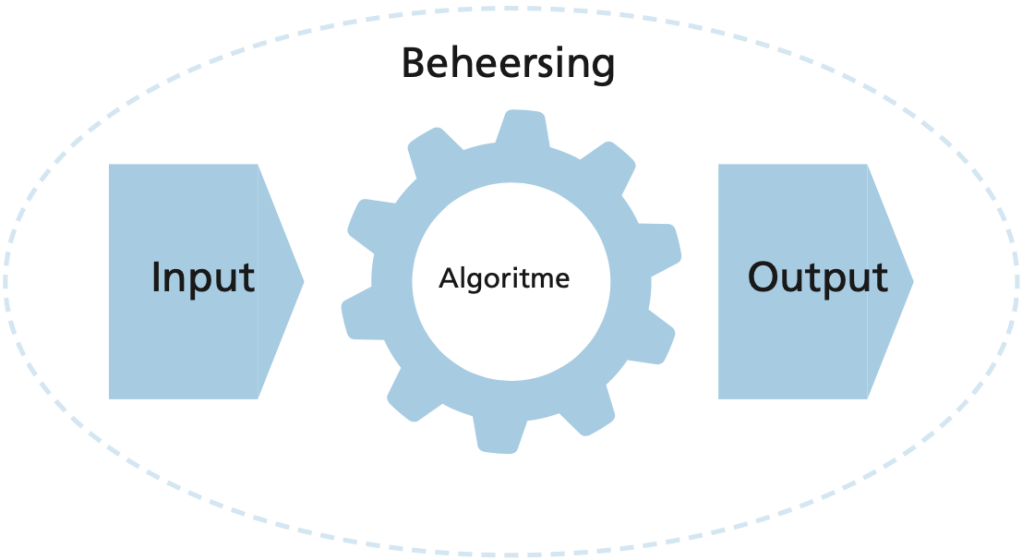
Er zijn veel verschillende typen algoritmes. Wanneer in huidige discussies gesproken wordt over algoritmes in de context van auditing, dan gaat het in de regel over complexe algoritmes die zelfs voor de ontwikkelaars nauwelijks inzichtelijk zijn en waarvan ze niet kunnen uitleggen hoe precies bepaalde input tot bepaalde output leidt. We hebben het dan over *machine learning*-toepassingen: een toepassing van kunstmatige intelligentie om op basis van data en door middel van wiskundige en statistische techniek te komen tot een bepaalde output. Dergelijke algoritmes zijn lastig uitlegbaar en toetsbaar, gerefereerd wordt daarbij aan de term *black box*. Hierdoor is het vertrouwen van het maatschappelijk verkeer in het geding.

Mogelijke benaderingen

Ruwweg zijn twee benaderingen mogelijk om een algoritme te toetsen. De ene benadering is de black box open te breken, de andere benadering is 'outside-the-black-box' denken.

De eerste benadering, het openbreken van de black box, houdt in dat de IT-auditor zich specialiseert in data science en zich richt op het doorgronden en beoordelen van het ontwikkelproces en de technische werking van het algoritme. Deze benadering heeft zeker waarde. Tegelijk blijven de nodige vragen over. Is het wel mogelijk een generiek normenkader te ontwikkelen voor deze benadering, gezien de grote variëteit aan algoritmes? Is het wenselijk dat enkel in data science gespecialiseerde IT-auditors dergelijke audits kunnen uitvoeren? Hoeveel zekerheid over de uitkomsten van het algoritme geeft deze benadering feitelijk? Immers, veel algoritmes maken een beslissing of voorspelling op basis van waarschijnlijkheid. Ook bestaan er algoritmes die gebruikmaken van een techniek waarbij de parameters van een algoritme gaandeweg van karakter kunnen veranderen. Het vaststellen van de werking van het algoritme over een periode wordt daarmee bemoeilijkt. Het betekent in elk geval dat je niet kunt volstaan met een technische test en het beoordelen van een algoritme, vergelijkbaar met code-inspectie van een reguliere applicatie.

Bij de tweede benadering, 'outside-the-black-box', zien we het algoritme (in figuur 1 in het midden weergegeven) als black box en richten we de aandacht op de elementen daaromheen: input en output van het systeem en beheersing ervan.



Figuur 1: Afbakening benadering 'outside-of-the-black-box'

We zetten deze elementen even op een rij:

Input. Een proces dat gebruikmaakt van een algoritme begint met input. Deze input kan verschillende vormen hebben, bijvoorbeeld: gestructureerd of ongestructureerd, digitaal of analoog.

Algoritme. Het machine learning-algoritme wordt ontwikkeld, getraind, onderhouden en ingezet.

Output. Het algoritme levert output op die de basis vormt voor, al dan niet, geautomatiseerde beslissingen. Het type beslissingen en het doel waarvoor ze worden gebruikt, bepalen de impact van de output.

Beheersing. Een algoritme heeft een eigenaar en kent verantwoordelijken. Het algoritme is ontwikkeld en wordt gebruikt en onderhouden in een specifieke context. Bovendien kent het algoritme verschillende stakeholders. Dit betekent dat we rondom de inzet van een algoritme een breed scala van interne beheersingsmaatregelen mogen verwachten.

Eisen aan de aanpak

We stelden drie vereisten aan de te ontwikkelen aanpak.
De eerste vereiste was dat een IT-auditor geen data science-expertise hoeft te bezitten om zich een oordeel te kunnen vormen over de inzet van een algoritme. Daarom hebben we ervoor gekozen het algoritme als een gegeven te beschouwen en te onderzoeken hoe we de beheersingsomgeving van een algoritme in kaart kunnen brengen en toetsen. Immers, door te focussen op de interne beheersingsmaatregelen rond de inzet van een algoritme is onderzoek mogelijk voor een niet-data scientist. Tegelijkertijd zijn we niet blind voor de risico's die in het inwendige schuilen, in de techniek van algoritmes. Daarom moet de aanpak wel de ruimte bieden om eventueel een data scientist in te schakelen. Ter vergelijking: bij het toetsen van informatiebeveiliging kan een auditor zich richten op beheersingsmaatregelen én kan deze overwegen om een penetratietest uit te laten voeren door een expert. De auditor hoort dan wel voldoende kennis te hebben om de uitkomsten van de werkzaamheden van de expert te interpreteren.

Een tweede vereiste was een brede maatschappelijke relevantie van de aanpak. Deze eis hebben we ingevuld door vier externe stakeholders te definiëren: consumenten, maatschappij, overheid en auditors. We hebben vanuit deze stakeholders onderzocht welke richtlijnen voor algoritmes of AI gepubliceerd waren. We kwamen tot een zevental nationale en internationale richtlijnen van verschillende gremia (zie kader 'Verwerkte richtlijnen'). Dit vormde ons uitgangspunt voor een inventarisatie van de elementen die in elk geval in een normenkader voor algoritmes thuishoren.

De derde en laatste vereiste was dat de wijze van toetsen breed toepasbaar is voor het bonte scala van algoritmes en hun toepassingen. Daarom wilden we komen tot een generiek toetsingskader met normen die op uiteenlopende situaties zijn toe te snijden.

Verwerkte richtlijnen

Wij gebruikten de volgende richtlijnen:

- Oxford-Munich Code of Conduct.
- Impact Assessment AI door ECP.
- Ethics guidelines for trustworthy AI door High-Level Expert Group on AI.
- IIA Audit Framework for AI door IIA.
- DNB Principes verantwoord gebruik AI door DNB.
- 10 Aandachtsgebieden AI door AFM en DNB.
- AI Governance Frameworks door Personal Data Protection Commission of Singapore.

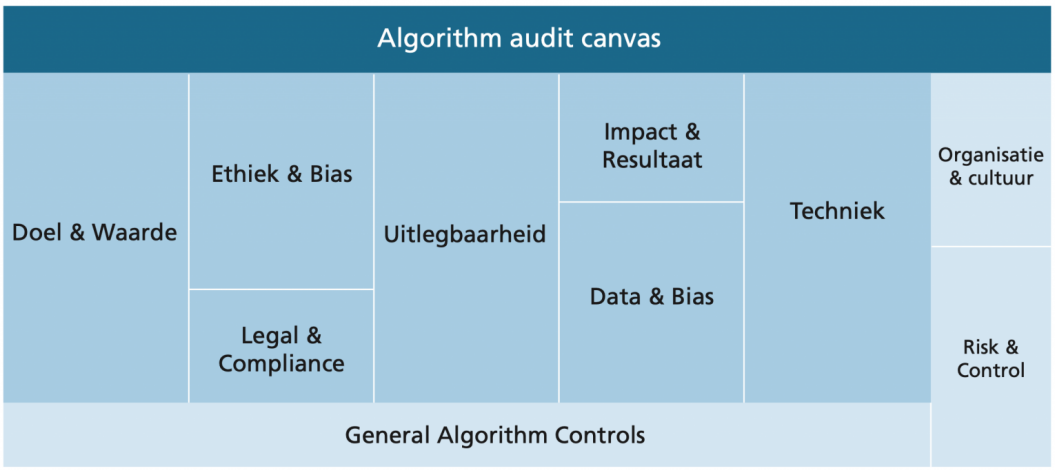
Een generiek normenkader

Algoritmes en de inzet ervan verschillen behoorlijk van elkaar. Ook raken de componenten input, AI, output en beheersing op hun beurt een breed scala aan aandachtsvelden. Hierbij valt te denken aan zaken als techniek, ethiek en compliance. Hoe zijn al deze variabelen en aandachtsvelden te vangen in een generiek normenkader?

Onze benadering was inspiratie te halen uit het Business Model Canvas. [OST10] Dat is de visuele weergave van een bedrijfsomgeving in de vorm van een rechthoek, verdeeld in negen vlakken (zie figuur 2). Door de vlakken in te vullen kun je de levensvatbaarheid van een businessmodel bepalen. Naar analogie daarvan geven we de variabelen voor de risicoanalyse van (de omgeving van) een algoritme weer als een rechthoek, verdeeld in tien vlakken. Daarachter hebben de normen uit verschillende kaders een plek gekregen. Met het invullen van de vlakken gidst het model je door de risicoanalyse heen. Per vlak weeg je af waar het zwaartepunt van de auditwerkzaamheden moet liggen.

Wij hebben uit de zeven verwerkte nationale en internationale richtlijnen (zie tekstkader ‘Verwerkte richtlijnen’) normen afgeleid die relevant zijn vanuit het perspectief van de vier stakeholders, te weten: consumenten, maatschappij, overheid en auditors. Deze normen hebben wij in verschillende stappen gecategoriseerd, geëvalueerd en geconsolideerd.
We hebben ze ook vertaald naar beheersingsdoelstellingen en risico-aandachtsgebieden wanneer dat niet al zo was. Zo kwamen wij uiteindelijk tot tien vlakken (aandachtsvelden}

met per vlak de beheersingsdoelstellingen en bijbehorende risico-aandachtsgebieden. Zie voor een voorbeeld van een beheersingsdoelstelling uit het vlak Ethiek & Bias het kader 'Voorbeeld beheersingsdoelstelling: medewerkers werken conform ethische richtlijnen'.



Figuur 2: Algorithm audit canvas

Het gaat hier te ver om elk aandachtsveld uit te diepen. Wel maken we enkele toelichtende opmerkingen.

1. de blauwe vlakken hebben direct betrekking op het proces waarin het algoritme functioneert. De roze vlakken zijn randvoorwaardelijk.
2. net als bij alle andere ICT-toepassingen het geval is, moeten ook bij algoritmes de IT General Controls op orde zijn. Niet alleen bij het algoritme zelf, maar ook bij de input en output data. In de door ons geraadpleegde richtlijnen vonden we ze niet terug; we hebben ze toegevoegd.
3. bij de risicoanalyse en daaruit volgende auditwerkzaamheden in het vlak 'Techniek' kan het zinvol zijn een extern deskundige in te zetten, bijvoorbeeld een data scientist.

Voorbeeld beheersingsdoelstelling: medewerkers werken conform ethische richtlijnen

Norm	Beheersingsdoelstelling	#	Risico-aandachtsgebied
EB2	Beheersingsmaatregelen bieden een redelijke mate van zekerheid dat betrokken medewerkers hun verantwoordelijkheden kennen en werken conform ethische richtlijnen.	EB2.1	De ervaring en scholing van mede-werkers is niet actueel.
		EB2.2	Het ontbreekt medewerkers aan een kritische houding ten opzicht van input data, bias, modellen en uitkomsten.
		EB2.3	De organisatie beschikt niet over een ethische code of richtlijn.

Dit is een beheersingsdoelstelling uit het vlak 'Ethiek & Bias'.

Inzet van het Canvas

Onze aanpak van een audit op basis van het Canvas bestaat uit de volgende stappen.

Stap 1. Onderzoeksvraag

Aan de voorkant wordt bepaald wat de onderzoeksvraag is en welke mate van zekerheid de gebruiker van de rapportage verlangt. Vervolgens bepaalt de auditor welke vlakken in scope zijn en op welke normen binnen elk vlak het zwaartepunt moet liggen.

Stap 2. Risicoanalyse

Het generieke normenkader, dus het Algorithm Audit Canvas met het bijbehorende risicoanalysekader met beheersingsdoelstellingen en risico-aandachtsgebieden, wordt in deze stap geconcretiseerd. Het resultaat is een normenkader dat is toegesneden op de specifieke onderzoeksvraag, in de vorm van een overzicht van de af te dekken risico's. Dit gebeurt door een risicoanalyse uit te voeren op de risico-aandachtsgebieden uit het Algorithm Audit Canvas die in scope zijn. De auditor kan daarbij besluiten risico's die als laag of niet van toepassing zijn ingeschat, niet op te nemen.

Stap 3. Specificeren en toepassen normenkader

Het geconcretiseerde normenkader uit de vorige stap maakt duidelijk voor welke beheersingsdoelstellingen en op welke risico-aandachtsgebieden beheersingsmaatregelen nodig zijn. In deze stap wordt gespecificeerd welke maatregelen dat precies moeten zijn. Daarbij is het mogelijk dat één beheersingsmaatregel meerdere risico's afdekt. Ook kan het zijn dat voor één risico meerdere beheersingsmaatregelen nodig zijn. Het framework biedt dus enkel risico's en beheersingsdoelstellingen en het is aan de verantwoordelijke voor het algoritme om invulling te geven aan de beheersingsmaatregelen die daar bij horen.

De benoemde beheersingsmaatregelen vormen het toegepaste normenkader. Als de auditor het toegepaste normenkader als toereikend beoordeelt, vormt dit het uitgangspunt voor de audit. Afhankelijk van de getroffen beheersingsmaatregelen overweegt de auditor of en zo ja hoe een externe deskundige moet worden ingezet. Zo is het denkbaar dat voor het toetsen van beheersingsmaatregelen binnen het aandachtsveld Data & Bias een datascientist wordt ingeschakeld.

Stap 4. Auditresultaten

Zoals gezegd, ondervangen beheersingsmaatregelen in een aantal gevallen meerdere risico's die verspreid kunnen zijn over meerdere vlakken van het Algorithm Audit Canvas. Resultaten van de audit kunnen per vlak inzichtelijk gemaakt worden. Echter, er kan ook sprake zijn van vooraf bepaalde key controls.

Validering van de aanpak

Als toets op de praktische bruikbaarheid hebben we hebben het geschetste normenkader in een *case study* getoetst. Het object van onderzoek was een algoritme binnen een organisatie in de financiële sector. Het doel was te evalueren of het framework praktisch toepasbaar is. Daarbij keken we onder andere naar:

1. **Geschiktheid:** sluiten de gevonden resultaten uit het framework voldoende aan bij de assurance-vraag vanuit de business?
2. **Dekking:** komen alle resultaten uit het framework voldoende naar voren ?
3. **Concreetheid en toetsbaarheid:** zijn de theoretische risico's te vertalen naar praktische, toetsbare maatregelen?
4. **Efficiëntie:** brengen de auditwerkzaamheden geen excessieve inspanningen mee?

Het onderzoeksobject was een algoritme met een behoorlijke impact op consumenten. Daarom moest de rapportage vertrouwelijk blijven. Helaas kunnen we daarom geen nadere details van het onderzoek beschrijven. Wel willen we over de validatie-methodologie opmerken dat we de objectiviteit zoveel mogelijk hebben gewaarborgd door andere IT-auditors te betrekken bij de review op onze case study. Ook hebben we de opdrachtgever het normenkader laten valideren en heeft het interne auditteam de classificatie van bevindingen naar risico's voor haar rekening genomen.

De resultaten van ons onderzoek hebben we opgenomen in een managementrapportage die we aan de opdrachtgever hebben gepresenteerd. Op basis van het opgestelde normenkader hebben wij het geheel aan bevindingen in ogenschouw genomen en zijn wij tot een aantal kernbevindingen gekomen, waarbij wij de oorzaak, het probleem, het risico en de aanbeveling hebben opgenomen. Op deze wijze hebben wij het management een zo compleet mogelijk beeld gegeven hoe men opvolging kan geven aan de bevindingen.

De organisatie heeft de managementrapportage positief ontvangen en als inzicht gevend ervaren. Wij zijn in het onderzoek zowel gestuit op reeds onderkende hiaten in de interne beheersing, als op nieuwe, soms verrassende tekortkomingen. Ook heeft de organisatie beter zicht gekregen op het algemene vraagstuk van beheersing rondom (en audit van) haar algoritmes.

Conclusie

Het normenkader biedt bruikbare handvatten om de risico's bij de beheersing van algoritmes in kaart te brengen. Vanuit het literatuuronderzoek hebben wij een raamwerk opgesteld om 'outside the black box' beheersingsmaatregelen in kaart te brengen om deze risico's voldoende te mitigeren. Over de bruikbaarheid van onze aanpak concluderen we dat die aanpak heeft geleid tot bruikbare inzichten voor het management waarmee ze de interne beheersing van de inzet van algoritmes kan verbeteren. Onze *overall* conclusie is dat het generieke Algorithm Audit Canvas prima geschikt is gebleken voor het opzetten van een toegepast en toetsbaar normenkader voor een audit rondom een algoritme.

Tot slot

Wij hopen dat onze aanpak ook anderen kan inspireren verder te werken aan normenkaders. Zo kunnen we met gezamenlijke inspanningen uiteindelijk komen tot een aanpak die (inter)nationaal wordt gedragen door onze beroepsorganisaties.

Voorbeeld: risico-aandachtsgebied ‘Doel & Waarde’

Het risico-aandachtsgebied ‘Doel & Waarde’ betreft het doel waarvoor het algoritme wordt gebruikt en wat de beoogde toegevoegde waarde van het gebruik hiervan is voor zowel de organisatie als de omgeving waar het algoritme invloed op heeft. Een organisatie kan bij de analyse van dit aandachtsgebied bijvoorbeeld uitgaan van het Product Vision Board van Pichler. [PICH11] toe te passen op het algoritme. Vanuit het Product Vision Board kan een organisatie een inschatting maken van de toegevoegde waarde van een product, zoals de toepassing van een algoritme. Daarin worden een aantal zaken meegenomen die ook naar voren komen in het aandachtsgebied van het Algorithm Audit Canvas:

- Het doel van het algoritme en waarvoor het wordt gebruikt.
- De omgeving waarop het algoritme van invloed is.

Gebaseerd op verschillende normenkaders uit de internationale onderzoeksgemeenschap en uit wet- en regelgeving, hebben we verschillende risico’s geïdentificeerd, zoals:

- De organisatie heeft niet voor ogen wat de doelstelling van het algoritme is en hoe het algoritme functioneert of zou moeten functioneren binnen de bedrijfscontext.
- Het is onduidelijk wat de minimale accuraatheid van het algoritme moet zijn, waardoor de doelstelling van het algoritme niet kan worden gemeten.
- De baten en lasten, waarden en belangen van de algoritmetoepassing zijn niet in kaart gebracht of niet in kaart gebracht voor (alle mogelijke) stakeholders.
- De doelstelling voor het verzamelen en verwerken van datasets die worden ingezet in het algoritme is niet duidelijk.

Dat leidt bijvoorbeeld tot de volgende beheersingsdoelstelling: ‘Beheersingsmaatregelen bieden een redelijke mate van zekerheid dat de doelstelling van het algoritme en hoe het algoritme functioneert of zou moeten functioneren binnen de bedrijfscontext, zijn vastgelegd en worden geborgd.’

Hier stopt het generieke gedeelte van het normenkader en zal het normenkader verder moeten worden geconcretiseerd en toegespitst op de organisatie waar het wordt ingezet. Bijvoorbeeld door risico’s te categoriseren en vervolgens maatregelen te definiëren die risico’s dekken.

Literatuur

[AFM19] AFM & DNB. *Artificiële Intelligentie in de verzekeringssector; Een verkenning,* https://www.afm.nl/nl-nl/nieuws/2019/jul/verkenning-ai-verzekeringssector, geraadpleegd op 4 april 2021.

[BOER21] Mona de Boer en Harry van Geijn. *NOREA Guiding Principles Trustworthy AI investigations Guiding principles for investigations of enterprise artificially intelligent algorithmic systems,* NOREA, maart 2021, https://www.norea.nl/download/?id=9720, geraadpleegd op 23 juni 2021.

[DEIT18] De IT-Auditor (2018). *Algorithm Assurance; Nieuwe werkgroep ingesteld,* https://www.deitauditor.nl/business-en-it/algorithm-assurance-nieuwe-werkgroep-ingesteld, geraadpleegd op 4 april 2021.

[DNB19] DNB. *General principles for the use of Artificial Intelligence in the financial sector,* https://www.dnb.nl/media/jkbip2jc/general-principles-for-the-use-of-artificial-intelligence-in-the-financial-sector.pdf, geraadpleegd op 4 april 2021.

[ECP18] ECP. *Artificial Intelligence Impact Assessment,* https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-Assesment.pdf, geraadpleegd op 4 april 2021.

[EURO19] Europese Commissie. *Ethics guidelines for trustworthy AI*, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, geraadpleegd op 4 april 2021.

[EY18]. *Assurance in the age of AI; The impact of emerging technology on assurance approaches and implications for assurance leaders, EY*. https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/digital/ey-assurance-in-the-age-of-ai.pdf, geraadpleegd op 4 april 2021.

[IIA17]. *The IIA's Artificial Intelligence Auditing Framework*, IIA (The Instute of Internal Auditors)https://na.theiia.org/periodicals/Public%20Documents/GPI-Artificial-Intelligence-Part-II.pdf, geraadpleegd op 4 april 2021.

[OXFO18] Oxford – Munich. *Code of Conduct*, http://www.code-of-ethics.org/code-of-conduct/, geraadpleegd op 4 april 2021.

[PDPC20] PDPC (Personal Data Protection Commission Singapore). *Model artificial intelligence governance framework*, https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf, geraadpleegd op 4 april 2021.

[OST10] A. Osterwalder, Yves Pigneur, Alan Smith. *Business Model Generation; A Handbook for Visionaries, Game Changers, and Challengers.* Wiley, 21010.

[PICH11] R. Pichler. *The Product Vision Board,* https://www.romanpichler.com/blog/the-product-vision-board/ geraadpleegd op 4 april 2021

Emil van Werven RE | Senior Consultant IT Advisory bij *Baker Tilly Nederland*

Emil is vijf jaar werkzaam bij Baker Tilly, binnen de afdeling IT Advisory. Hij houdt zich voornamelijk bezig met IT audits in het kader van de jaarrekeningcontrole, assurance- en advies opdrachten.

Drs. Gerke Roorda | Senior Consultant IT Advisory bij *Baker Tilly Nederland*

Gerke werkt sinds 2016 bij Baker Tilly – IT Advisory. Hij doet in de MKB-sector IT-audit werkzaamheden in het kader van de jaarrekeningcontrole. Daarnaast is hij met name voor de publieke sector betrokken bij assurance- en adviesopdrachten in het kader van informatiebeveiliging en privacy.

Pim Smulders RE CISA | Manager IT Advisory bij *Baker Tilly Nederland*

Pim werkt 5 jaar bij IT Advisory binnen Baker Tilly en houdt zich vanuit daar bezig met IT audits in het kader van de jaarrekeningcontrole, assurance trajecten bij serviceorganisaties in de vorm van ISAE en SOCII en een aantal adviesopdrachten. Zijn achtergrond ligt bij Information Management; het slaan van bruggen tussen Business en IT. Daarnaast is Pim enkele jaren actief als lid van de Young Professionals commissie van NOREA en ISACA.