

Betrouwbare AI

De impact van verschillen tussen parketten en maatschappelijke ontwikkelingen op de ontwikkeling en inzet van AI bij het OM

AI Oversight Lab – Marianne Schaaphok, Marissa Hoekstra, Romy van Drie en Lieke Dom

TNO 2023 P12807

Inhoudsopgave



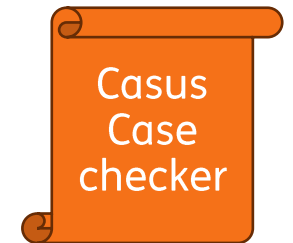
- [Leeswijzer](#)
- [Hoofdstuk 1: Context van het onderzoek](#)
- [Hoofdstuk 2: OM-specifieke eigenschappen bij het toepassen van AI.](#)
- [Hoofdstuk 3: Representativiteit van data](#)
- [Hoofdstuk 4: Eigenschap 1: Invloed van verschil in parketsbeleid op AI ontwikkeling](#)
- [Hoofdstuk 5: Eigenschap 2: Invloed van maatschappelijke ontwikkelingen op AI ontwikkeling](#)
- [Hoofdstuk 6: Conclusie en aanbevelingen](#)
- [Referenties](#)

Leeswijzer

Dit rapport is als volgt opgebouwd:

- **Hoofdstuk 1: Context van het onderzoek.** In dit hoofdstuk wordt het onderzoeksdoel, de onderzoeksvragen, de onderzoeksmethode en de context van de case checker toegelicht.
- **Hoofdstuk 2: OM-specifieke eigenschappen bij het toepassen van AI.** Dit hoofdstuk gaat in op de onderzoeksvraag: *Welke organisatorische en context specifieke eigenschappen van het OM vragen extra aandacht in de verantwoorde ontwikkeling, inzet en monitoring van AI?* Op twee context specifieke eigenschappen wordt verder in dit rapport dieper ingegaan: 1) De invloed van verschillen tussen arrondissementsparketten (AP's) en 2) de invloed van maatschappelijke ontwikkelingen op de representativiteit van data en de ontwikkeling en toepassing van AI.
- **Hoofdstuk 3: Representativiteit van data.** In dit onderzoek ligt de focus op de aspecten van verschil in beleid tussen de parketten en de impact van maatschappelijke ontwikkelingen. De invloed van deze aspecten zijn sterk gelinkt aan het concept van representativiteit, om deze reden wordt in dit hoofdstuk dieper ingegaan op het concept representativiteit van data.
- **Hoofdstuk 4: Invloed van verschil in beleid van parketten op AI ontwikkeling.** In dit hoofdstuk gaan we in op de verschillen in het beleid van de AP's in de context van de Casechecker en mogelijke ontwerpen van de inzet van een AI-toepassing (landelijk versus regionaal).
- **Hoofdstuk 5: Invloed van maatschappelijke ontwikkelingen op AI ontwikkeling.** In dit hoofdstuk gaan we in op de invloed van maatschappelijke ontwikkelingen op werkwijze, beleid en bestuur van organisaties, overwegingen in ontwikkeling, inzet en monitoring en aan de hand van twee casussen wordt de impact geïllustreerd.
- **Hoofdstuk 6: Conclusies en aanbevelingen.** Dit hoofdstuk presenteert de conclusies en aanbevelingen van dit rapport.

NB: In het rapport wordt gebruik gemaakt van de case checker als leercasus. Indien geleerde lessen en observaties specifiek voor de case checker casus gelden dan wordt het volgende oranje symbool op de pagina weergegeven.



Hoofdstuk 1: Context van het onderzoek

Link met het KOM en het AI Oversight Lab

Dit onderzoek wordt gefinancierd vanuit het Kennisopbouwprogramma OM (KOM) en het AI Oversight Lab.

KOM:

- In het KOM werken TNO en het OM samen aan kennisopbouw voor het OM op een aantal transformatiethema's: effectgestuurd werken, strategisch inzicht, sociale innovatie en samenwerking, geduide kennis en kennis-en innovatieproces. Het doel hiervan is om de innovatie en benodigde technologische en organisatorische veranderingen te stimuleren bij het OM.
- Dit onderzoek draagt hieraan bij door te kijken welke OM specifieke eigenschappen een rol spelen in de innovatie met Artificiële Intelligentie (AI) en hoe het OM zich voor kan bereiden op de betrouwbare ontwikkelig en inzet van AI zowel op technisch als organisatorisch vlak.

Oversight Lab:

- In het Oversight Lab werkt TNO samen met verschillende publieke partijen met als doel om hen te ondersteunen bij verantwoorde keuzes in de ontwikkeling en inzet van betrouwbare AI. In het lab worden het technische en het organisatorische perspectief meegenomen om partijen in de praktijk verder te helpen.
- Dit onderzoek draagt bij aan de kennisbasis van het Oversight Lab over de uitdagingen van verantwoorde ontwikkeling en inzet van AI in de praktijk en de impact van maatschappelijke ontwikkelingen en regionale verschillen op de ontwikkeling en inzet van AI.

Onderzoeksdoel

- In 2023 zijn TNO en het OM een samenwerking gestart om te onderzoeken hoe de Trustworthy AI principes kunnen worden toegepast binnen het OM bij de ontwikkeling en inzet van AI binnen het OM.
- Deze samenwerking maakt onderdeel uit van het AI Oversight Lab van TNO. Binnen dit lab wordt voornamelijk gewerkt aan innovaties rondom 'Trustworthy AI', oftewel betrouwbare AI. Zowel inzichten en ontwikkelingen vanuit beleidskaders en wetenschappelijke literatuur, alsook de vertaling daarvan naar de praktijk omtrent het opzetten, toepassen en monitoren van AI staan centraal in dit Lab.
- Vanuit de ervaring in de eerdere jaren van het AI Oversight Lab is gebleken dat het essentieel is om al in een vroeg stadium data (science) teams van een organisatie en de beoogde AI-toepassing te verbinden met de andere onderdelen van een organisatie. De context van de organisatie waarin de AI-toepassing wordt gebruikt en de manier waarop zij interacteert met de maatschappij is namelijk van groot belang. Om deze reden is hier dan ook extra aandacht voor in dit onderzoek.
- Het onderzoek is in 2023 opgedeeld in twee fasen.
 - **Fase 1: voorbeeld van de case checker casus.**
 - In fase 1 lag de focus op het voorbeeld van de casechecker casus, een experimenteel AI model dat kan prioriteren welke zaken als eerste op zitting zouden moeten worden ingepland. In fase 1 van het onderzoek heeft TNO met de casechecker als leercasus workshops en interviews gehouden waarin onderwerpen als doelen, gebruikers, publieke waarden, bias en transparantie aan bod zijn gekomen.
 - Op basis van deze resultaten [1] zijn er onderzoeksvragen naar voren gekomen over hoe je dit kan waarborgen en faciliteren bij nieuwe ontwikkeltrajecten van het OM.
 - **Fase 2: in kaart brengen van OM specifieke factoren en hun relatie tot verantwoorde inzet van AI**
 - In fase 2 van het onderzoek lag de focus van de samenwerking tussen het AI Oversight Lab van TNO en het Testlab OM op het inzichtelijk krijgen wat de belangrijkste aspecten zijn om te behandelen bij de ontwikkeling van AI binnen het OM.
- Dit rapport presenteert de resultaten van fase 2.

[1] zie TNO (2023). Trustworthy AI Resultaten Fase 1.

Onderzoeksvragen

Om inzicht te krijgen in hoe context specifieke eigenschappen van het OM aandacht vragen in verantwoorde ontwikkeling, inzet en monitoring van AI, zijn voor deze studie drie onderzoeksvragen geformuleerd:

- Welke **organisatorische** en **context specifieke eigenschappen** van het OM vragen extra aandacht in de verantwoorde ontwikkeling, inzet en monitoring van AI?
- Hoe krijg je inzicht in de **context** (maatschappelijke ontwikkelingen, beleidskeuzes, historie) waarin de data zijn verzameld en de invloed op de **representativiteit** voor de beoogde toepassing?
- Wat is de **invloed van verschil in beleid van parketten** bij de ontwikkeling, inzet en monitoring van AI binnen het OM?

Onderzoeksmethode fase 2

De beantwoording van de onderzoeksvragen is uitgevoerd aan de hand van desk research en interviews.

1. Desk research

- Literatuuronderzoek naar organisatorische aspecten, context en representativiteit van data
- Literatuuronderzoek naar impact van regionaal beleid op toepassing van data modellen
- Literatuuronderzoek naar impact van maatschappelijke ontwikkelingen op werkwijze en beleidskeuzes

2. Interviews

- Interviews met beleidscoördinatoren vanuit verschillende arrondissementsparketten (AP) en van het Parket Centrale Verwerking OM (CVOM)

3. Uitwerking voorbeelden

- Een drietal voorbeelden is uitgewerkt om de impact van verschillend beleid en maatschappelijke ontwikkelingen concreet te maken.

Overzicht geïnterviewde rollen:

- | | |
|--|--|
| - Beleidsadviseur CVOM | - Coördinator / teamleider AP West-Brabant / Zeeland |
| - Accountmanager CVOM | - Beleidscoördinator AP Amsterdam |
| - Plaatsvervangend Officier van Justitie AP Oost-Nederland | - Officier van Justitie AP Amsterdam |
| - Coördinator zittingsvoorbereiding Oost-Nederland | - Business controller AP Noord-Holland |

Context Casechecker

- Voor het onderzoek werd gebruik gemaakt van de casus de Casechecker, een experiment van het Testlab OM. Deze casus is in fase 1 breder verkend op de aspecten gebruikers, doel, transparantie, bias en publieke waarden. De Casechecker wordt in de volgende slide toegelicht.
- De interviews zijn gehouden met rollen gerelateerd aan deze casus.
- De resultaten in het kader van de casechecker worden vertaald naar algemene inzichten en aanbevelingen voor de inzet van AI-toepassingen.

Context Casechecker

Inleiding: De Casechecker casus is een experiment waarin een algoritme is ontwikkeld door het Testlab OM om de kansen van AI bij prioritering van politierechterzaken (PR-zaken) te kunnen onderzoeken. Het algoritme wordt dus niet ingezet. We gebruiken de casechecker als leercasus in dit onderzoek concreet en praktisch te maken en lessen te trekken over de verantwoorde ontwikkeling en inzet van AI-toepassingen binnen het OM.

Organisatie doel: In fase 1 zijn meerdere mogelijke doelen van het systeem geïdentificeerd.¹ Voor het vervolg is de aanname gemaakt dat de casechecker het doel heeft om te prioriteren welke zaken als eerste op zitting moeten worden gebracht bij de PR.

Model doel: Het model voorspelt de hoogte van de strafeis in dagen celstraf. Op basis van deze voorspelde strafeis wordt een ranking gemaakt welke zaken als eerste moeten worden ingepland. Dat betekent dus dat (in dit geval) de zaken met de hoogste strafeisen de hoogste prioriteit krijgen.

Data: Beschikbare data vanuit het geïntegreerd processysteem strafrecht (GPS) over de periode 2019-2021 vanuit alle AP's.

Model: Machine learning model dat getraind is op deze historische data

Opmerkingen:

De hoogte van de strafeis wordt hier nu als belangrijkste voorspellende factor gezien voor welke zaken de meeste prioriteit moeten krijgen. Er zijn vaak ook andere contextfactoren van belang, zoals de staat van gesteldheid van verdachte (zie slide 8), die niet in het model worden meegenomen.

[1] zie slide X voor het overzicht uit de rapportage TNO, Trustworthy AI, Resultaten Fase 1, 2023

Welke factoren spelen mee bij de prioritering?

- Hoogte van de strafeis, waardoor o.a. impliciet ook meespelen:

- Aanwezigheid van slachtoffer
- Gegevens over verdachte
- Type feit
- Maatschappelijke klasse
- Recidive
- Schade

Prioritering o.b.v.
casechecker

Prioritering o.b.v. huidige werkwijze

Praktisch

- Impact op slachtoffer
- Gesteldheid van de verdachte
- Ouderdom van de zaak
- Quota die gehaald moeten worden
- Evenement-gestuurde zittingen

- Beschikbaarheid van advocaten en rechters
- Documentatie niet op orde
- Ruimte op zitting

Merk op:

- De casechecker voorspelt de hoogte van de strafeis. Als de prioritering puur op basis van de casechecker gemaakt wordt, dan speelt alleen de voorspelde strafeis mee voor de prioritering.
- In de huidige werkwijze voor het bepalen van prioritering spelen meer factoren mee
- Praktische zaken spelen niet mee voor de prioritering, maar wel voor de uiteindelijke planning van de zaken. Wees hiervan bewust bij evaluaties van de casechecker.

Hoofdstuk 2: OM specifieke eigenschappen bij het toepassen van AI

Bevinding onderzoeksvraag 1: overzicht context specifieke eigenschappen OM

De inzet van AI moet altijd worden bekeken in relatie tot de eigenschappen van de organisatie, de context waarin de organisatie opereert en de context van de specifieke toepassing.

In de interviews en workshops zijn een aantal eigenschappen naar boven gekomen die relevant zijn om in gedachten te houden bij het ontwikkelen en inzetten van AI:

- **Verskil in beleid tussen parketten:** het OM bestaat uit verschillende AP's verspreid over het land. Deze AP's kunnen (deels) hun eigen beleid vormgeven wat zorgt voor verschillen in werkwijzen en uitvoering.
- **Impact van maatschappelijke ontwikkelingen:** Maatschappelijke ontwikkelingen kunnen rechtstreeks een effect hebben op het werk van het OM. Het werk van het OM is in sommige aspecten 'gevoelig' voor ontwikkelingen die op korte termijn kunnen plaatsvinden in de maatschappij.
- **Ketensamenwerking:** Het OM opereert in directe samenwerking met andere partners in de keten, zoals de politie, NFI en de rechtbank.
- **Normerende rol van het OM in de maatschappij:** Naast dat ontwikkelingen in de maatschappij invloed hebben op het werk van het OM, is de aanpak van het OM ook een manier om de maatschappij te beïnvloeden en normen te stellen. Met als doel dat een bepaalde vorm van straffen de overtredingen vermindert.
- **Rechtvaardigheid, transparantie en veiligheid** zijn kern publieke waarden voor het OM. Alle publieke waarden zijn natuurlijk van belang, maar ook hier zullen afwegingen in moeten worden gemaakt. Bepaalde publieke waarden moeten hierbij prioriteit krijgen.

■ Focus van dit onderzoeksrapport

■ Geen onderdeel van de scope van het huidige project

■ Focus van fase 1 van het onderzoek

Hoofdstuk 3: Representativiteit

De link met representativiteit

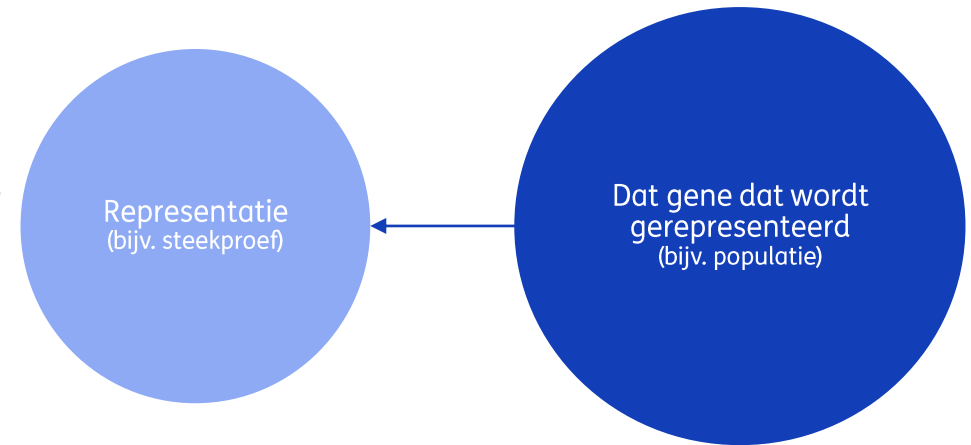
In dit onderzoek ligt de focus op de aspecten van verschil in beleid van parketten en de impact van maatschappelijke ontwikkelingen. De invloed van deze aspecten zijn sterk gelinkt aan het concept van representativiteit.

Deze sectie schetst de achtergrond van representativiteit, zodat de inzichten in het rapport beter geduid kunnen worden:

- Wat is representativiteit?
- Hoe creëer je een representatieve dataset?
- Hoe hangen representativiteit en bias samen?
- Uitdagingen bij representativiteit

Wat is representativiteit?

- Representativiteit heeft betrekking op het vermogen van het ene om voor het andere te staan. Zo is een willekeurige steekproef een representatie van een populatie (Chasalow & Levy, 2021)
- In de context van deze casus draait representativiteit om de mate waarin de data of het model een afspiegeling is van het grotere geheel, in dit geval de echte wereld.
- Een representatie kan in meer of mindere mate passend zijn voor datgene wat je wil representeren
- Er moet dus voor gezorgd worden dat de data representatief is voor de toepassing waarvoor het wordt ingezet, d.w.z. dat de data aansluit bij de populatie en klassen in de context van de toepassing.



Voorbeeld (Bergman et al., 2023)

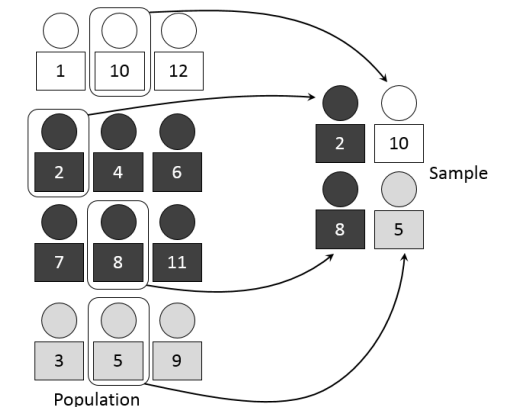
- Stel dat je een AI-model maakt om haat in Portugese Twitterberichten te herkennen, en je wilt dit model wereldwijd inzetten
- Dan is het van belang dat je data niet alleen voorbeelden bevat van Portugees uit Portugal, maar ook uit Brazilië, Angola, Mozambique, en alle andere regio's waar Portugees gesproken wordt.
- Als het Braziliaans Portugees verschilt van Portugees uit Portugal met betrekking tot haatberichten op Twitter, dan is een dataset die representatief is voor Portugees uit Portugal niet representatief voor Braziliaans Portugees.
- Een AI-model dat getraind is op Portugees uit Portugal zal waarschijnlijk niet goed werken voor Braziliaans Portugees.

Wat is representativiteit?

- Of een representatie passend is, kan vanuit verschillende perspectieven beoordeeld worden:
 - accuraatheid – in hoeverre is de huidige situatie hetzelfde als gepresenteerd in de data
 - aspiratie – reflecteert de data hoe wij te werk willen gaan in de toekomst? (Chasalow & Levy, 2021)
- Verder zijn er verschillende categorieën van representativiteit (Henriksen et al., 2020), waar we bij de casechecker expliciet kijken naar:
 - Temporele representativiteit: dit draait erom of de data verkregen in het verleden representatief is voor het heden of voor de beoogde toekomst.
 - Geografische representativiteit: dit draait erom of de data representatief is voor verschillende regio's.
- Daarnaast moet ook de grootte en de opbouw van de data worden meegenomen in de bepaling of de dataset representatief is.
 - De steekproef van de totale populatie moet voldoende grootte hebben om de populatie voldoende te representeren
 - Voor de opbouw van de beschikbare data zijn verschillende vormen, waaronder:
 - 1) in de dataset zijn is de gehele populatie bekend, maar alleen in een deel van de gevallen is kennis over het label (bijv. we hebben een lijst van alle bedrijven, maar weten maar van een klein aantal of ze een overtreding begaan of niet)
 - 2) de dataset bevat een steekproef van de totale populatie, waarbij voor elk sample het label bekend is (bijv. in een alcoholfuik worden willekeurig automobilisten gestopt waarvan we voor elk exact weten of ze wel of niet teveel alcohol op hebben)

Hoe creëer je een representatieve dataset?*

- Wanneer er nagedacht wordt over representatieve sampling vanuit het accuraatheidsperspectief, dan wordt er vaak gekeken of de samenstelling van de dataset overeenkomt met de overeenkomstige populatie op basis van een aantal meetbare kenmerken of variabelen. Een techniek die dit ondersteunt is **random sampling** (Bergman et al., 2023)
 - In het voorbeeld over haatdetectie in het Portugees, houdt dat in dat de verdeling van taalvariaties (Braziliaans-Portugees, Portugees uit Portugal, etc.) evenredig is met de populatie van de gemeenschappen waarin de taalvariaties gesproken worden. Als er 19x zo veel sprekers zijn van Portugees in Brazilië dan in Portugal, dan zou er ook 19x zo veel Portugees uit Brazilië in de dataset moeten zitten.¹
- Random sampling is echter niet perfect. Op deze manier worden de metingen gedomineerd door meerderheidsgroepen. Het is mogelijk dat de metingen de prestaties van het model niet goed de minderheidsgroepen weergeven. Een slechte performance voor deze groepen is dan ook moeilijk te meten (Bergman et al., 2023) .
- Er zijn een aantal technieken die gebruikt kunnen worden als alternatief voor random sampling, waaronder **stratified sampling**. Hierbij neem je random samples per subgroep in plaats van uit de populatie. Met deze techniek kun je ervoor zorgen dat iedere groep evenredige representatie heeft.
 - Voor de casechecker kun je voor een groep denken aan een parket. Als je de casechecker landelijk wilt gebruiken, dan is het van belang dat ieder parket gerepresenteerd is in de dataset.
- Het aspiratie vraagstuk vraagt om een bredere verkenning vanuit bestuurlijk- en beleidsperspectief.



Stratified sampling (bron: [Stratified sampling - Wikipedia](#))

* We behandelen de discussie over dit onderwerp op hoofdlijnen, aangenomen dat de historische data een goede afspiegeling is van de gewenste toekomstige situatie (aspiratie).

¹ Dit is een simplificatie. Hoeveel er precies nodig zijn van iedere groep heeft ook te maken met de variatie onder de sprekers.

Hoe hangen representativiteit en bias samen?

- In het vakgebied van Trustworthy AI zijn fairness, bias en harm vaakgenoemde begrippen. Onvoldoende representativiteit wordt vaak gezien als een belangrijke reden voor het optreden van bias, harm en unfairness.
- Representatieve schade (**representational harm**) kan onder andere optreden wanneer in de trainingsdata een deel van de bevolking ondervertegenwoordigd is of verkeerd wordt gerepresenteerd, en daarmee geen goede generalisatie is voor een subgroep van de gebruikspopulatie (Bergman et al., 2023). In het voorbeeld over haatdetectie in het Portugees kan representatieve schade optreden als de detectie minder goed werkt voor sprekers van bijvoorbeeld Braziliaans Portugees.
- **Representatieve bias** ontstaat wanneer een systeem (bijvoorbeeld een zoekmachine of een generatieve AI) bepaalde sociale groepen in een minder gunstig daglicht stelt dan andere, of hun bestaan helemaal niet erkent (Bergman et al., 2023). Dit komt vaak door het ontbreken van deze groepen in de dataset.
- Er zijn meerdere typen bias die gelinkt kunnen worden aan de oorsprong, verzameling en verwerking van data, zoals bijvoorbeeld measurement bias, omitted variable bias en aggregation bias. Voor een uitgebreid overzicht van verschillende typen bias die een rol spelen bij machine learning verwijzen we door naar Mehrabi, N. et al (2022).

Uitdagingen bij representativiteit (casechecker)

Binnen de Casechecker casus zijn er verschillende redenen dat data mogelijk minder representatief is met betrekking tot het doel van de casechecker:

- **Verzamelde data is gestuurd door (regionaal) beleid**
 - Veranderingen in het beleid kan er voor zorgen dat de data niet meer representatief is in temporele zin, waar verschillen in beleid tussen de regio's juist zorgen voor een vermindering van geografische representativiteit (accuraatheid). Bijvoorbeeld bij parket X komen door het beleid meer drugszaken binnen dan bij parket Y waar juist gefocust wordt op geweldszaken of bij parket X wordt sneller voor een andere afdoening gekozen bij geweldszaken dan bij parket Y.
 - Daarnaast is de invloed van beleid ook erg belangrijk voor het aspiratievraagstuk. Wanneer er door beleid wordt gestuurd op (tijdelijk) zwaardere eisen in type X zaken of juist vaker de keuze voor een andere afdoening is de vraag of dat ook gewenst is in toekomstige scenario's?
 - Verder kan ook beleid in andere delen van de keten de verzamelde data beïnvloeden, zoals bijvoorbeeld wanneer de politie meer controleert in gebied Y of meer bij groepen A en B waardoor daar meer zaken van beschikbaar zijn in de dataset. Dit kan in de toekomst bij een wijziging van het beleid bij de ketenpartner zorgen dat de casechecker tool niet meer een goede afweging kan maken van de prioriteit van de nieuwe zaken.
- **Benodigde data is niet beschikbaar of niet correct**
 - Het ontbreken van data kan op verschillende niveaus een rol spelen, zo kunnen er bepaalde groepen uit de populatie ontbreken (bijv. de , er kunnen bepaalde variabelen ontbreken (bijv. is de verdachte een first-offender of een recidivist) of er kunnen bepaalde uitkomsten (labels) ontbreken (bijv. andere manieren van afdoening worden niet meegenomen).
 - Ook foutieve of incomplete data heeft effect op de representativiteit en daardoor de kwaliteit van het model (bijv. het verkeerd invullen van de maatschappelijke klasse) .
- **Tijdelijke historische gebeurtenissen hebben de data beïnvloed**
 - Tijdelijke gebeurtenissen, zoals COVID, kunnen ervoor zorgen dat de verzamelde data in niet representatief is voor de huidige situatie, bijvoorbeeld doordat er veel minder inbraken waren door COVID-maatregelen. Bij de casechecker is data gebruikt uit de periode 2019-2021, wat twee covid-jaren bevat.

Hoofdstuk 4: Verschil in parketsbeleid en prioritering: voorbeeld van de casechecker

Verschil in parketsbeleid en prioritering: voorbeeld van de casechecker

Verschillen in beleid van parketten kunnen van invloed zijn op de verantwoorde ontwikkeling, de inzet en het vertrouwen in AI:

- Op pagina 17 zien we dat verschillend beleid de data kan beïnvloeden zodanig dat er lagere representativiteit is. Dit kan leiden tot lagere kwaliteit en toename van risico's omtrent eerlijkheid van de AI-toepassing. Een risico is bijvoorbeeld dat bepaalde groepen benadeeld worden. Dit speelt een grote rol wanneer er op landelijk niveau geanalyseerd wordt of parketten met elkaar vergeleken worden.
- Verschillen in het beleid van parketten leiden ook tot verschillende werkwijzen. Dit kan de keuze voor een mogelijke inzet als centraal landelijk model bemoeilijken, omdat de werkwijze moet aansluiten bij het model. Het hanteren van een eenduidige werkwijze en procedures, waarin afwijkingsruimte voor de parketten is vastgelegd, zijn van belang om risico's te mitigeren en transparant te kunnen zijn.

In deze sectie gaan we in op:

- Verschillen in beleid van parketten in de context van de Casechecker
- Mogelijke ontwerpen van de inzet van een AI-toepassing (landelijk versus parketgericht)

Context OM: verschil in beleid van AP's

- **Verschil in type zaken per arrondissementsparket (AP)**
 - Elk AP van het OM heeft te maken met andere context specifieke aspecten.
 - Bijvoorbeeld: het AP in Oost-Nederland heeft vaker te maken met voetbalrellen, waar het AP in Brabant vaker tegen drugszaken aanloopt.
 - Hierdoor ontstaan ook verschillen in prioritering, en verschillen in type zittingen.
- **Verschil in werkwijze, prioriteit en beleid per AP**
 - Elk AP plant en prioriteert welke casussen als eerste op een politierechterzitting worden ingepland.
 - Elk AP heeft zijn eigen manier voor het inplannen van deze zittingen (na de beoordeling van de OvJ) en verschillende werkwijzen.
 - Zo is bij de één de werkwijze om een lijst geordende query's uit te voeren op de planningslijst in Excel om tot een prioritering te komen, waar bij de andere een automatisch toegekende score per zaak wordt bepaald op basis van vaste richtlijnen in Excel. De hoogste scores krijgen hierbij de hoogste prioriteit.
- **Organisatorische gevolgen voor het OM**
 - AP specifieke context is van invloed op de ontwikkeling, inzet en monitoring van verantwoorde AI binnen het OM. Het is reëel dat er verschillen tussen de AP's nodig zijn, hier ligt de uitdaging om dit inzichtelijk te maken en hiermee om te kunnen gaan.
 - Uit interviews blijkt dat niet overal duidelijk is wat landelijk en wat AP-specifiek beleid is (op het gebied van prioritering)
 - Uit interviews blijkt dat niet overal duidelijk is of er landelijk beleid is en hoe dat wordt vertaald / vastgelegd op regionaal niveau.

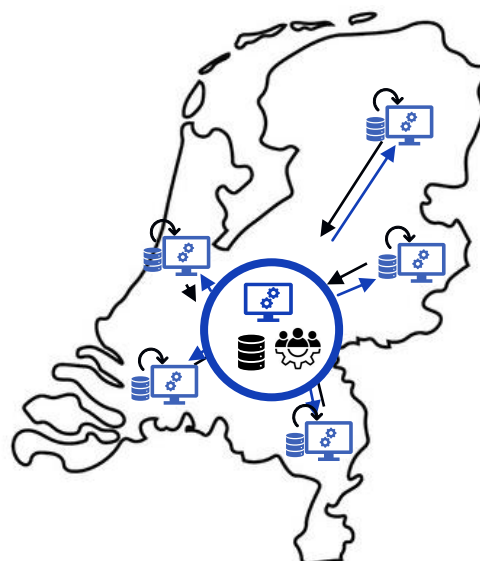
Vier mogelijke manieren van ontwerp

Er is sprake van een zekere mate van autonomie bij de parketten waardoor er verschillend lokaal beleid is, het is belangrijk om hier bij het ontwerp van het AI-toepassing en de werkwijze rekening mee te houden. De werkwijze houdt in de manier waarop het model wordt ingezet in het werkproces.

Er zijn verschillende ontwerpen mogelijk, met elk hun eigen overwegingen en consequenties. We laten hieronder vier mogelijke vormen zien en lichten ze verder toe in de volgende slides.



Centraal model en uniforme werkwijze



Centraal basismodel met fine-tuning bij de AP's, uniforme werkwijze



Model per AP, maar met een uniforme werkwijze



AP-specifieke modellen en werkwijze¹



AI-toepassing



Oorsprong Data



Werkwijze



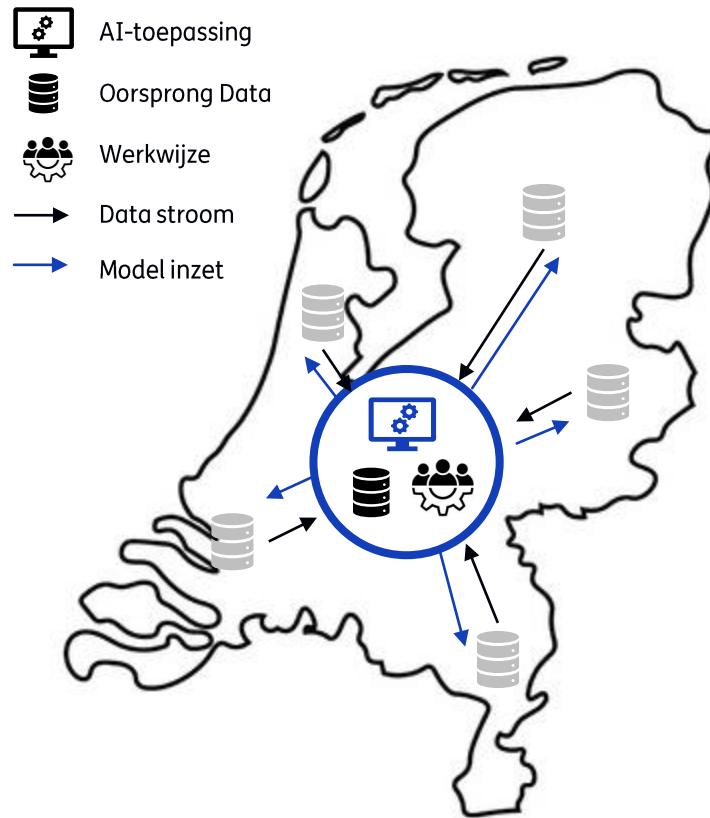
Data stroom



Model inzet

1. Dit is in het kader van uniformingslagen binnen het OM een onwaarschijnlijk scenario

1. Centraal model met uniforme werkwijze



Centraal model en uniforme werkwijze

In dit scenario wordt de data uit de verschillende AP's bij elkaar gebracht en gebruikt voor de training van een centraal model. Dit model wordt op een uniforme manier ingezet bij de AP's.

Beleid: Landelijk gestuurd beleid wat geldt voor alle AP's.

Werkwijze: Uniforme werkwijze die gelijk is in alle AP's, d.w.z. alle AP's hebben dezelfde processen en interacteren op dezelfde manier met het model.

Model & Data: De data wordt vanuit de verschillende AP's verzameld en gecombineerd. Het model wordt getraind op deze gecombineerde data, echter moet hier wel rekening gehouden worden met mogelijke regionale bias door verschillen in de AP's en mogelijk ander beleid dat in het verleden is gevoerd in de AP's. Het model moet centraal en decentraal geëvalueerd en herijkt worden.

Voordelen:

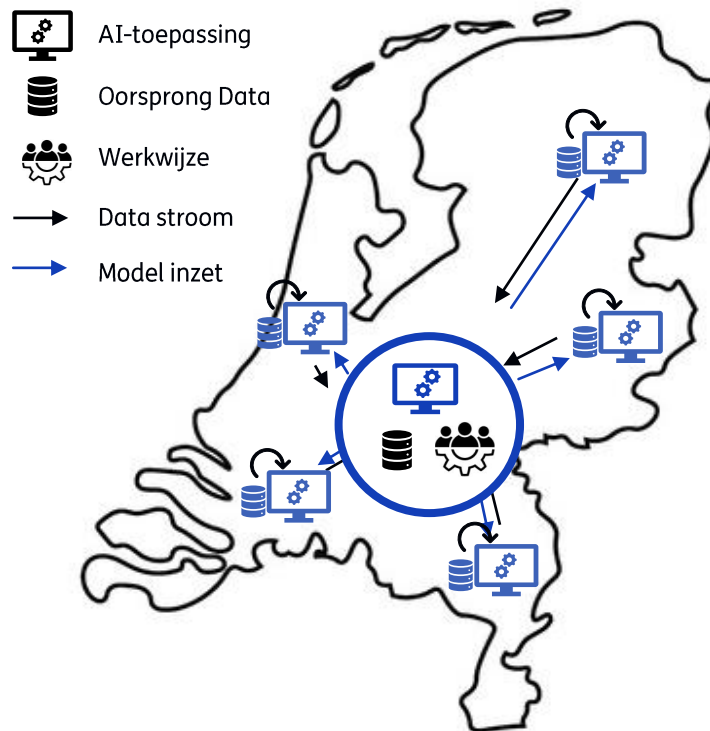
- er is een uniforme en eenduidige werkwijze in het gehele land;
- er is een centraal en eenduidig databeleid;
- het centrale model bevat landelijke ontwikkelingen;
- het model en de werkwijze sluiten aan bij landelijk beleid en doelstellingen;
- overzichtelijk en transparant wat kan zorgen voor verhoogd vertrouwen en gevoel van rechtvaardigheid;
- veranderingen in de prioritering en werkwijzen kunnen snel worden aangepast;
- AP's kunnen profiteren van elkaars data.

Aandachtspunten:

- minder mogelijkheden om aan te kunnen passen naar verschillen tussen AP's;
- verschillende data strategieën in het verleden kunnen het lastiger maken om data te combineren;
- verschillen in de AP's en verschillend AP beleid in het verleden kan zorgen voor verminderde representativiteit op landelijk niveau;
- grondige analyse nodig in hoeverre het centrale model aansluit bij de verschillende AP's.

2. Centraal basismodel met finetuning bij de AP's

In dit scenario wordt er een centraal basismodel getraind op de data vanuit verschillende AP's. De AP's krijgen de mogelijkheden om op bepaalde punten het model te finetunen naar de eigen situatie, bijvoorbeeld door het aanpassen van enkele regels of gewichten in het model. De AP's hebben vervolgens dezelfde werkwijze met het model.



Centraal basismodel met finetuning bij de AP's, uniforme werkwijze

Beleid: Landelijk beleid met afgebakende ruimte voor aanpassingen binnen de AP's.

Werkwijze: Uniforme werkwijze als leidraad met mogelijkheden voor afgebakende aanpassingen binnen de AP's.

Model & Data: Een basisset aan data vanuit de AP's wordt gecombineerd en het model wordt getraind op deze centrale data. AP's kunnen op verschillende manieren finetunen, bijvoorbeeld door het bijtrainen op basis van additionele eigen data, door het invoeren van extra expliciete regels voor hun of door het nemen van extra stappen in hun werkwijze. De flexibiliteit die AP's hebben wordt expliciet afgebakend om een centrale aanpak mogelijk te maken. Het model moet zowel centraal als per AP geëvalueerd en herijkt worden.

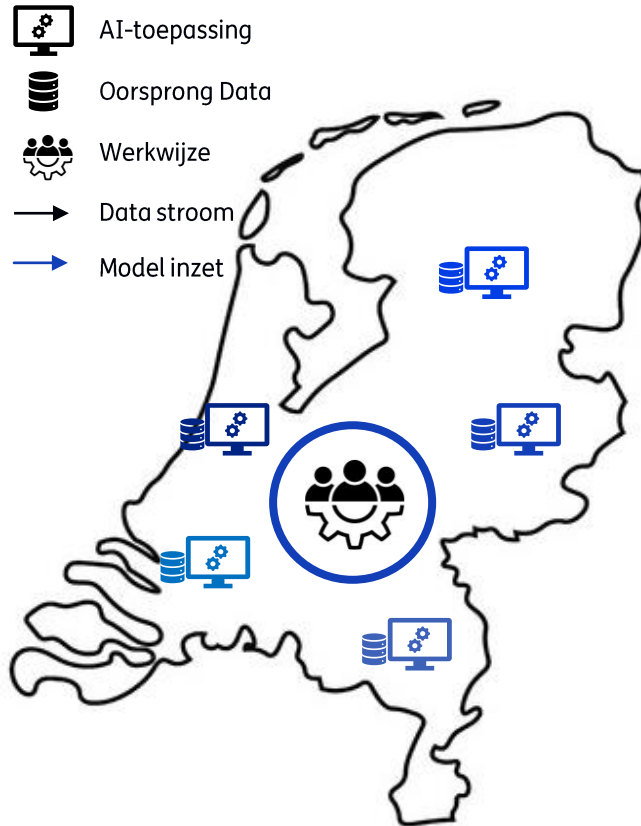
Voordelen:

- er is een centraal en eenduidig databeleid;
- het centrale model bevat landelijke ontwikkelingen;
- het model en de werkwijze sluiten aan bij landelijk beleid en doelstellingen;
- Afgebakende flexibiliteit geeft AP's de mogelijkheid in te spelen op regionale verschillen;
- representativiteit voor verschillende AP's kan beter worden afgedekt.

Aandachtspunten:

- door de aanpassingen is de manier van werken minder overzichtelijk en eenduidig en daardoor vragen transparantie en uitlegbaarheid (zowel OM-intern, als richting ketenpartners als naar burgers) meer aandacht;
- er moeten duidelijke afspraken nodig over wat wel en niet lokaal mag worden aangepast;
- verschillende data strategieën in het verleden kunnen het lastiger maken om data te combineren;
- verschillen in de AP's en verschillend AP beleid in het verleden kan zorgen voor verminderde representativiteit op landelijk niveau;
- moet zowel landelijk als per AP worden geëvalueerd en gemonitord.

3. Model per AP met uniforme werkwijze



Model per AP, maar met een uniforme werkwijze

In dit scenario blijft de data bij de verschillende AP's, de AP's trainen het model op hun eigen data. De werkwijze is uniform bepaald en het doel van de modellen is dus gelijk.

Beleid: Landelijk beleid op de werkwijze en AP specifiek beleid, mogelijk gevoed door landelijk beleid, op de prioritering.

Werkwijze: Uniforme werkwijze, alle AP's werken op dezelfde manier met een gelijksoortig model (d.w.z. het model heeft hetzelfde doel en een gelijk type uitkomsten)

Model & Data: Alle AP's gebruiken hetzelfde type model, maar trainen het met hun eigen data. Dat wil zeggen dat de modellen hetzelfde doel hebben en dezelfde type uitkomsten aanleveren, maar dat de gewichten, regels en ontwerp van het algoritme kunnen verschillen. Het model moet per AP worden geëvalueerd en herijkt. Op landelijk niveau is herijking van de werkwijze, doel en uitkomsten van het model wenselijk. Ook is een evaluatie van onderlinge verschillen wenselijk.

Voordelen:

- dit zorgt voor een flexibeler model dat beter kan inspelen op verschillen tussen AP's;
- de geografische representativiteit wordt beter afgedekt;
- uniforme werkwijze zorgt voor eenduidigheid en transparantie over de werkwijze;
- centrale data-uitwisseling is in dit geval niet noodzakelijk, hoewel het nog steeds wenselijk is om verschillen tussen de AP's in kaart te brengen en te evalueren.

Aandachtspunten:

- meer werk en complexer om te evalueren, monitoren en onderhouden omdat alle AP's dit individueel moeten uitvoeren;
- duidelijke transparantie bieden over de uitkomsten van de modellen wordt uitdagender;
- kan door de verschillen tussen AP's ervaren worden als minder rechtvaardig;
- door te concentreren op data per AP ontstaat het risico op tunnelvisie en het vergroten van verschillen tussen de AP's.

4. AP-specifieke modellen en werkwijze



AP-specifieke modellen en werkwijze

In dit scenario wordt het gehele proces bij de AP's neergelegd. Dat betekent dat de AP's zelf bepalen op wat voor manier ze de prioritering vormgeven, welk beleid ze voeren, of en op welke manier ze daar data en/of modellen voor gebruiken en hoe ze de processen vormgeven.

Beleid: AP specifiek beleid, mogelijk gevoed door landelijk beleid.

Werkwijze: Er wordt per AP bepaald of er met data en modellen wordt gewerkt, zo ja op welke manier en met wat voor type model er wordt gewerkt.

Model & Data: In het geval dat er data wordt gebruikt, richt een AP zich op haar eigen data om een model mee te trainen. Het model moet daardoor ook per AP worden geëvalueerd en herijkt.

Voordelen:

- zeer flexibele werkwijze die veel ruimte laat om in te spelen op verschillen tussen AP's;
- de geografische representativiteit wordt hierdoor beter afgedekt;
- een centrale data-uitwisseling is niet noodzakelijk.

Aandachtspunten:

- op landelijk niveau kan het uitdagender worden om transparant en uitlegbaar te zijn over de manier van werken;
- het kan door de verschillen tussen AP's en minder eenduidige manier van werken ervaren worden als minder rechtvaardig door de burgers;
- het is meer werk en complexer om te evalueren en te onderhouden, omdat elk AP dit zelf moet organiseren;
- er kunnen mogelijk grotere verschillen ontstaan tussen AP's in doelmatigheid;
- door te concentreren per AP ontstaat er een risico op tunnelvisie en het vergroten van verschillen tussen de AP's.

We merken hierbij op dat door uniformeringsslagen bij het OM dit scenario onwaarschijnlijk zal zijn.

Generieke inzichten

- Herken en erken dat er verschillen zijn tussen de verschillende AP's waar expliciet mee moet worden omgegaan.
- Analyseer de dataset die je wilt gebruiken:
 - Kan de data uit verschillende regio's/parketten worden gecombineerd: is hetzelfde gedocumenteerd, hebben de verschillende bronnen dezelfde aannames (dezelfde manier waarop de data verzameld is)
 - Onderzoek de representativiteit van de dataset: zitten hier bijvoorbeeld regionale of historische keuzes in die kunnen verschillen voor de verschillende oorsprongen van de data.
 - Analyseer wat je ziet in de data, maar ga ook het gesprek aan over wat er ontbreekt en wat je juist niet ziet in de data. Om dit effectief te doen zullen verschillende perspectieven betrokken moeten worden.
- Analyseer hoe het doel van het model aansluit bij de organisatorische doelstellingen, zowel op landelijk als op parketniveau.
- Bedenk waar uniformering wenselijk is en waar ruimte moet zijn voor flexibiliteit en leg deze afwegingen expliciet vast
- Analyseer verschillende manieren van ontwerp voor de inzet van het model waarmee invulling kan worden gegeven aan de verschillen tussen parketten en/of regio's, bijvoorbeeld een central model of een model per parket.

Hoofdstuk 5: Invloed van maatschappelijke ontwikkelingen op AI ontwikkeling

Impact maatschappelijke ontwikkelingen

Korte termijn of disruptieve maatschappelijke ontwikkelingen zijn van invloed op de verantwoorde ontwikkeling en inzet van AI.

- Werkwijzen kunnen niet meer geldig/houdbaar zijn voor de situatie
- Data kan niet meer representatief zijn voor de situatie

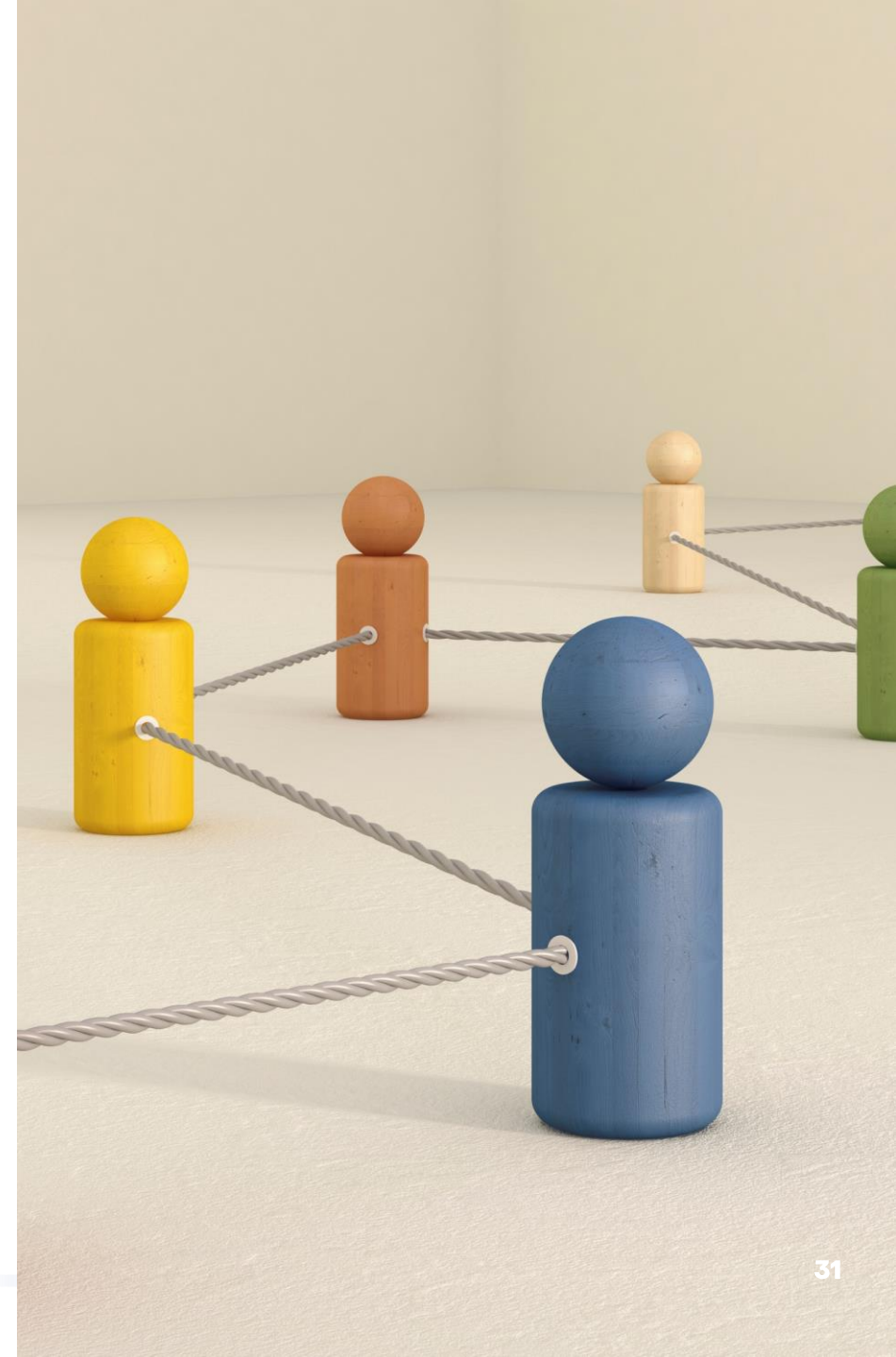
In deze sectie gaan we in op:

- Invloed van maatschappelijke ontwikkelingen op werkwijze, beleid en bestuur van organisaties
- Overwegingen in ontwikkeling, inzet en monitoring
- Impact in een casus: Scheidsrechtersgeweld
- Impact in een casus: COVID

Invloed van maatschappelijke ontwikkelingen vergt adaptive governance

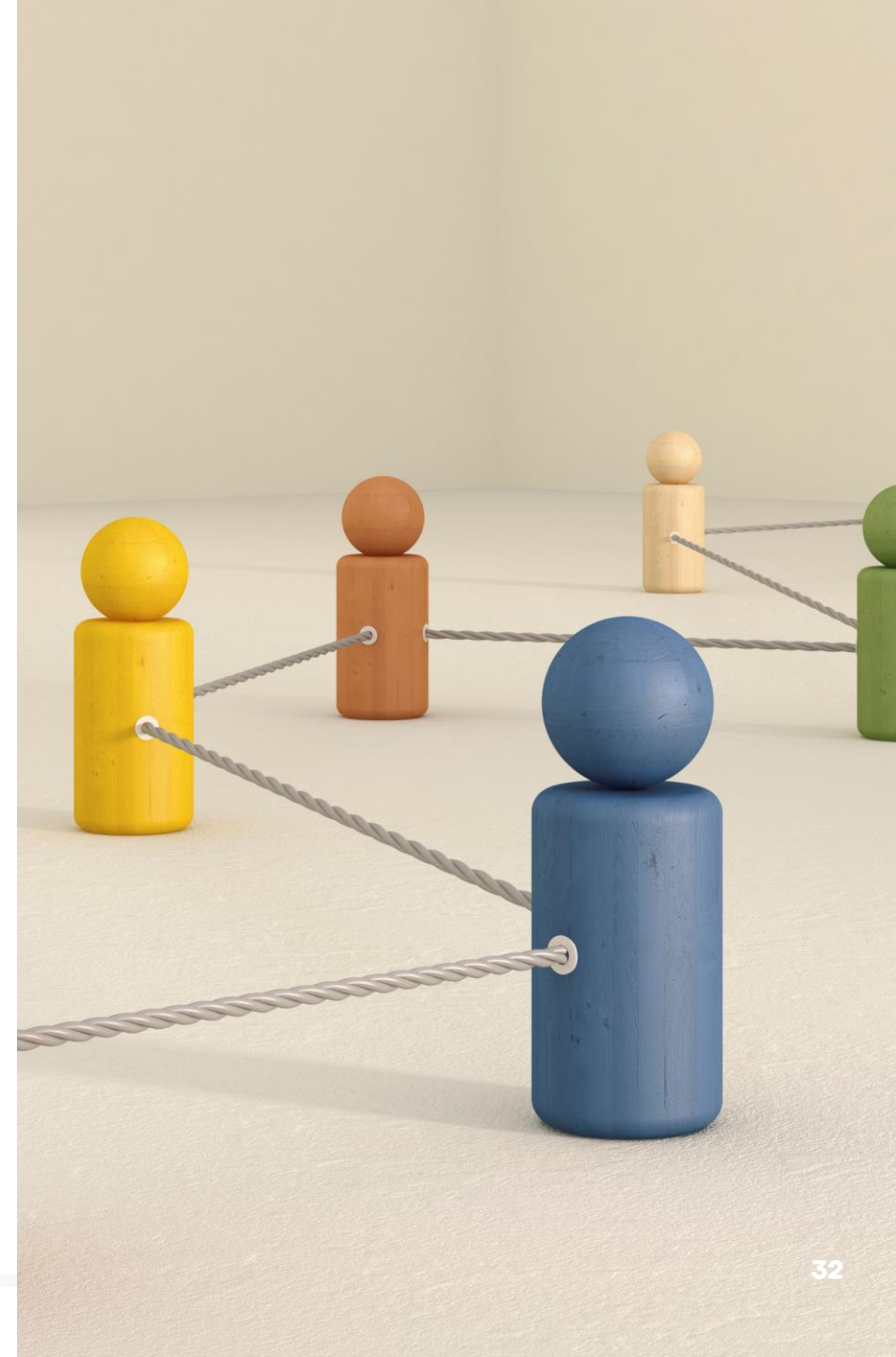
Invloed van maatschappelijke ontwikkelingen op werkwijze, beleid en bestuur van organisaties

- Door de wettelijke taak en normerende functie van het OM is het werk van het OM in sommige aspecten 'gevoelig' voor ontwikkelingen die op korte termijn kunnen plaatsvinden in de maatschappij. Als gevolg hiervan kan de werkwijze, beleid en prioriteiten van het OM wijzigen. Dit heeft ook invloed op de ontwikkeling, inzet en monitoring van AI. Data waarop de AI-toepassing is ontwikkeld kan namelijk niet meer representatief zijn voor de situatie.
- Inspelen op maatschappelijke ontwikkelingen vergt adaptief bestuur, ook wel '**adaptive governance**' (TNO, 2022) genoemd. Om als organisatie adaptief te kunnen werken en goed in te kunnen spelen op maatschappelijke ontwikkelingen is een aantal aspecten van belang (TNO, 2022). Deze aspecten worden op de volgende pagina's kort toegelicht.
- Ter illustratie: De COVID pandemie laat de invloed van disruptieve maatschappelijke ontwikkelingen op organisaties zien. Ook op het OM heeft de COVID pandemie veel invloed gehad, dit wordt op pagina's 37-40 verder geïllustreerd.
- Om als organisatie adaptief te kunnen werken zijn een aantal randvoorwaarden van belang (Janssen & van der Voort, TNO, 2022). Deze randvoorwaarden worden op de volgende pagina's toegelicht.



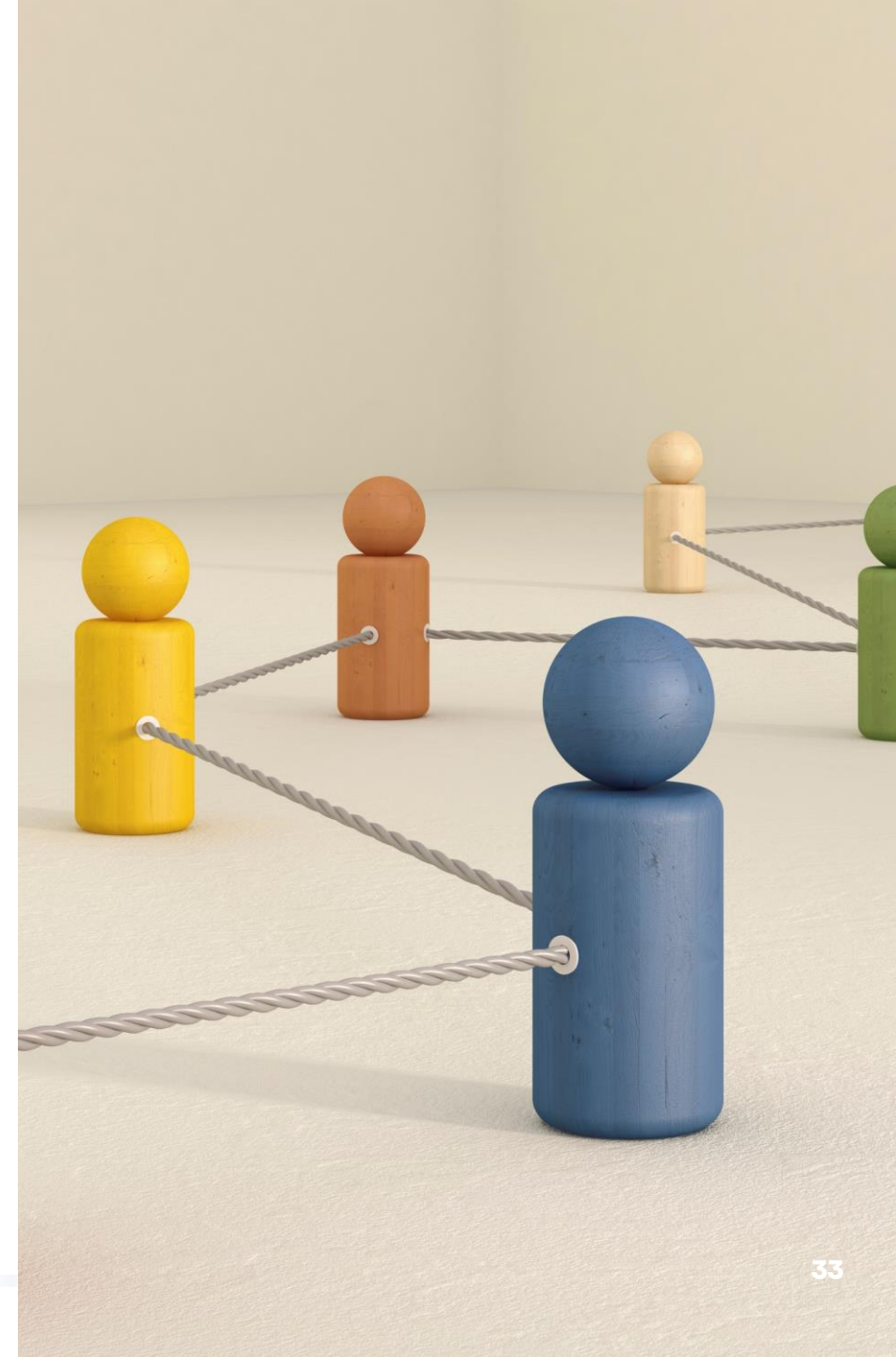
Belang van het netwerk

- Het is van belang om inzichtelijk te hebben welke **stakeholders binnen het netwerk** belangrijk zijn en deze kunnen mobiliseren indien nodig (Geuijen, 2018, TNO, 2022).
 - **Ketenpartners van het OM**, zoals de politie, NFI en de rechtbank zijn hierbij belangrijke stakeholders.
- Besluitvorming over beleid vindt plaats op meerdere niveaus. Dit heet ook wel '**Multi-level governance**'. Adaptief bestuur vereist een balans tussen centrale en decentrale besluitvorming (Janssen & van der Voort, 2020; TNO, 2022).
 - Het OM kent zowel regionale AP's als landelijke onderdelen. **Afstemming** tussen deze verschillende onderdelen is cruciaal.



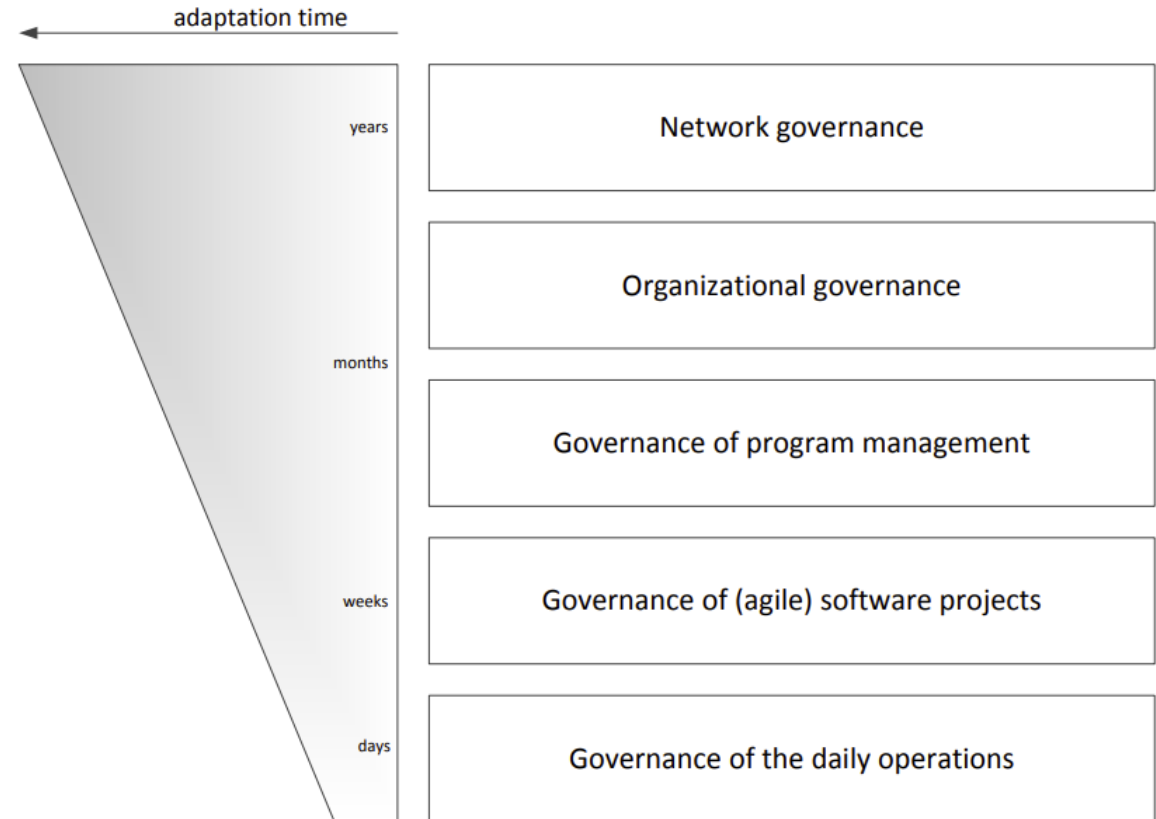
Capaciteit, leren en kennisdeling

- **Ontwikkel adaptieve capaciteit om onzekerheden en onverwachte ontwikkelingen te kunnen adresseren** (Gupta et. al. 2010; Van der Steen, 2018; TNO, 2022).
- Maak **ruimte voor experimenteren en leren** binnen een organisatie én met andere organisaties (Geuijen, 2018, TNO, 2022).
 - Het Testlab speelt hierin een centrale rol. Het is belangrijk om voldoende capaciteit binnen de organisatie te creëren zodat beoogde gebruikers samen met het Testlab kunnen experimenteren en leren.
 - Bijvoorbeeld door samen met relevante ketenpartners zoals de politie te experimenteren en te leren.
- Maak **ruimte voor kennisdeling en disseminatie**, ook met andere organisaties zoals ketenpartners (Geuijen, 2018, TNO, 2022). Een aantal manieren waarop het OM dit kan doen is via:
 - **Interne kennisdeling en disseminatie:** Bijvoorbeeld via opleidingen en trainingen via het opleidingsinstituut van het OM (SSR) of door kennis te delen via een intranet pagina.
 - **Externe kennisdeling en disseminatie:** bijvoorbeeld via kennisdelingssessies tussen het Testlab OM en andere labs binnen de keten.



Flexibiliteit versus robuustheid van een organisatie

- **Bestuursstructuren moeten zowel robuust als flexibel zijn.** Dit levert een spanning op tussen het vermogen om snel te kunnen inspelen op onverwachte ontwikkelingen versus het maken van goed doordachte, transparante en verantwoorde besluiten (Janssen & van der Voort, 2016; TNO, 2022).
- **Daarnaast kost het tijd om werkwijzen en organisatiestructuren aan te passen.** Hoeveel tijd dit kost is afhankelijk van de type en complexiteit van de governance structuur (zie figuur 2). (Janssen & van der Voort, 2016).
- Om op deze spanning in te kunnen spelen is het **belangrijk om beleid en strategieën te ontwikkelen** waarin duidelijk is op welke onderdelen het robuust moet zijn, en waar ruimte is voor flexibiliteit (Janssen & van der Voort, 2016).



Figuur 2. Bron: uit Janssen & van der Voort, 2016

Overwegingen in AI ontwikkeling, inzet en monitoring

Maatschappelijke ontwikkelingen vragen om een bepaalde mate van adaptiviteit en dit moet ook worden meegenomen in de ontwikkeling, inzet en monitoring van AI. We geven hieronder een aantal belangrijke overwegingen mee in deze fases.

Invloed op ontwikkeling

- Analyseer de data die je wilt gebruiken op representativiteit;
 - Uit welk tijdsframe komt de data?
 - Zijn er verschillen in de maatschappij tussen nu en moment waarop de data is verzameld?
- Maak een keuze over het type model dat je gaat gebruiken;
 - In hoeverre ondersteunt het type model adaptiviteit naar maatschappelijke ontwikkelingen? Bijv. kan het model met kennis en/of regels worden bijgestuurd of kan het op basis van data zelf ontwikkelingen detecteren en daarop adapteren?
- Evaluer de gevoeligheid en limitatie van het model.
- **Invloed op inzet**
 - Ontwikkel een duidelijke werkwijze om de manier waarop de AI-toepassing wordt ingezet door medewerkers eenduidig te houden.
 - Ontwikkel een strategie en procedure(s) om te kunnen omgaan met abrupte veranderingen die impact hebben op de inzet van het model: bijv. hoe wordt er gewerkt wanneer het model (plotseling) niet meer voldoet?
- **Invloed op monitoring**
 - Monitor of het model aansluit bij het beoogde doel en de huidige situatie in de maatschappij en de beoogde kwaliteit levert.
 - Maak afspraken over hoe vaak er wordt gekeken naar veranderingen in de maatschappij en beleid?
 - Maak procedures voor het updaten en bijtrainen van het model: wanneer, hoe vaak en op wat voor manier ga je bijtrainen?

Twee casussen ter illustratie

Om deze inzichten meer concreet te maken voor de context van het OM, kijken we naar twee verschillende casussen in de context van de casechecker. Dat wil zeggen in de context van prioritering van politierechter (PR) zaken. Deze casussen zijn gebaseerd op historische ontwikkelingen, maar zijn puur ter illustratie. Op sommige punten worden aannames gedaan die wellicht geen goede representatie zijn van de werkelijkheid.

We behandelen twee casussen:

- Toename van scheidsrechtersgeweld
- COVID-19 pandemie

Bij elke casus behandelen we:

- Beschrijving van de casus
- Gevolg van de casus
- Impact op de casechecker
- Aanbevelingen

Casus 1: Scheidsrechtersgeweld

Omschrijving

In 2012 werd de 41-jarige Richard Nieuwenhuizen dodelijk verwond na het optreden als scheidsrechter in een voetbalwedstrijd. Dit geval van zinloos geweld tegen scheidsrechters wakkerde een maatschappelijke verontwaardiging aan. Op basis hiervan werd er strenger opgetreden tegen scheidsrechtersgeweld en kregen deze zaken een hogere prioriteit.

Gevolg*

- Meer van dit soort zaken bij het OM
- Zaken kregen meer prioriteit bij inplannen op zitting
- Hogere straf ge-eist en hogere straffen opgelegd bij deze zaken

Impact op trainingsdata casechecker

- In de jaren na het voorval werden hogere straffen ge-eist en opgelegd. Dit zal ervoor zorgen dat de casechecker ook hogere strafeisen voorspelt voor deze zaken, en deze zaken dus hogere prioriteit krijgen.
- Door meer te handhaven zijn er logischerwijs ook meer zaken. Als de casechecker getraind kan worden op meer zaken, zal het de strafeis beter kunnen voorspellen.

* In de casussen zijn aannames gemaakt zodat we concreet kunnen illustreren wat de impact is op de casechecker. Op deze slide zijn de gevolgen gebaseerd op aannames.

 de Volkskrant
<https://www.volkskrant.nl> › Topverhalen vandaag

[Nog steeds is er geweld op het veld](#)

2 dec 2013 — De **KNVB** treedt harder op tegen wangedrag sinds het overlijden van grensrechter Richard Nieuwenhuizen, vandaag precies een jaar geleden.

 de Volkskrant
<https://www.volkskrant.nl> › Mensen

[Hof handhaaft straffen voor doodschoppen grensrechter](#)

Harder optreden **KNVB** De **KNVB** treedt inmiddels harder op tegen wangedrag sinds het overlijden van grensrechter Richard Nieuwenhuizen. Teams die hebben ...

 de Volkskrant
<https://www.volkskrant.nl> › Topverhalen vandaag

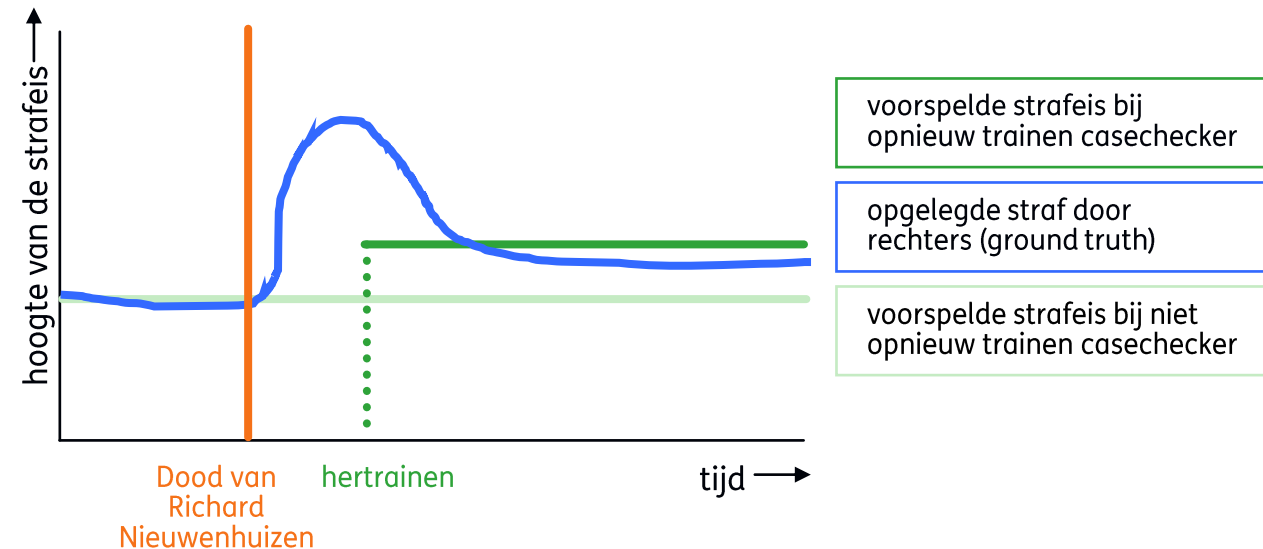
[KNVB lanceert campagne tegen voetbalgeweld](#)

26 okt 2015 — Na de gewelddadige dood van grensrechter Richard Nieuwenhuizen, maandag, heeft de bond talloze telefoontjes en e-mails ontvangen van personen ...
Bevat niet: ~~treedt wangedrag~~

Casus 1: Scheidsrechtersgeweld

Impact op casechecker

- Na het voorval (**oranje lijn**) leggen rechters hogere straffen op voor scheidsrechtersgeweld (**blauwe lijn**). Het is mogelijk dat dit daalt na verloop van tijd, wat is weergegeven in de figuur.
- De casechecker, die is getraind op historische data, beweegt niet uit zichzelf mee met deze veranderingen.
- Bij niet hertrainen zullen de **voorspelde strafeisen** van de casechecker (**lichtgroene lijn**) **lager** zijn dan de straffen die worden opgelegd door rechters.
 - Wanneer de prioritering wordt bepaald door de zaken te sorteren op hoogte van de strafeis, zullen zaken met scheidsrechtersgeweld dan ook een **lagere prioritering** krijgen dan gewenst is.
- Als het model opnieuw getraind wordt, dan is de **voorspelde strafeis** van de casechecker (**donkergroene lijn**) meer in lijn met de straf die worden opgelegd door rechters.
 - Wanneer de prioritering wordt bepaald door de zaken de sortering op hoogte van de strafeis, zullen de zaken de **gewenste prioritering** krijgen.
- Geleidelijke veranderingen kunnen worden opgevangen door te werken met een vast tijdsframe (time window) die alleen de recentste data in meeneemt in de voorspelling. Het tijdsframe schuift mee als er nieuwe data binnenkomt.



Conclusie

- Bij een veranderende situatie kan er opnieuw getraind worden, maar een machine learning model zoals de casechecker zal altijd achterlopen op een veranderende werkelijkheid.
- Hertrainen is mogelijk, maar het is belangrijk om dit mee te nemen bij de governance. Te weinig hertrainen leidt tot een verouderd model, maar te veel hertrainen vraagt mogelijk ook om meer evaluaties.

* In de casussen zijn aannames gemaakt zodat we concreet kunnen illustreren wat de impact is op de casechecker. Op deze slide wordt verder geredeneerd op basis van de aangenomen gevolgen van de vorige slide.

Casus 1: Scheidsrechtersgeweld

Aanbevelingen

- **Houd in de ontwikkeling van het model rekening met het aanpassend vermogen van het model.** Er moet in de ontwikkeling van het model een expliciete keuze gemaakt worden om het model wel of niet automatisch te laten bijleren op basis van nieuwe, recentere data. Modellen kunnen zelf adapteren door bijvoorbeeld het gebruik van een tijdsframe of het toekennen van zwaardere gewichten aan recentere data. Echter brengen zelflerende systemen ook weer andere uitdagingen met zich mee.
- **Monitor of de voorspelde strafeis consistent hoger/lager is dan de opgelegde straf.** Als de voorspelling consistent anders is dan de straffen die door een rechter worden opgelegd, is dit een indicatie van een veranderde werkelijkheid. Het is belangrijk om dit niet te beoordelen in een momentopname, maar om dit over de tijd te monitoren. Stel dat er op wordt gekeken of de voorspelde strafeis overeenkomen met de werkelijke straffen, en dat er geobserveerd wordt dat dit niet het geval, dan hoeft dat niet noodzakelijk te duiden op een veranderde werkelijkheid. Er kan dan ook sprake zijn van een model met een slechte performance. Stel nu dat er regelmatig gemonitord wordt, en het bekend is dat de voorspelde strafeis in het verleden goed overeenkwam met de opgelegde straf. Als er dan wordt geobserveerd dat de voorspelde strafeisen niet meer overeenkomen met de werkelijk opgelegde straffen, dan kan er geconcludeerd worden dat het model niet meer aansluit bij de werkelijkheid.
- **Herzie het model en/of de werkwijze bij grote veranderingen in de maatschappij en het beleid dat daaruit voortvloeit.** Een machine learning model zal altijd achterlopen op een veranderende werkelijkheid. Het kan een oplossing zijn om het model opnieuw te trainen als je observeert dat het model niet meer aansluit bij de werkelijkheid (bijvoorbeeld doordat de voorspelde strafeisen niet meer kloppen met de opgelegde straffen). Naast het updaten van het model, kan er ook voor worden gekozen om de werkwijze aan te passen. Stel dat de normale werkwijze is om de op strafeis gesorteerde lijst één-op-één over te nemen als prioritering, dan kan een update van de werkwijze inhouden dat de prioritering van zaken met scheidsrechtersgeweld handmatig verhoogd of verlaagd wordt.
- **Analyseer de dataset van het model,** en ga daarbij na wat voor kritieke gebeurtenissen en of veranderingen er hebben plaatsgevonden in het tijdsframe dat de dataset beslaat. Hiervoor moet een analysekader worden ontwikkeld.

* In de casussen zijn aannames gemaakt zodat we concreet kunnen illustreren wat de impact is op de casechecker.

Casus 2: COVID

Omschrijving

De coronapandemie is een disruptieve gebeurtenis die leidde tot veranderingen en uitdagingen voor overheidsorganisaties als handhaving en politie, en zo ook het OM. Tijdens deze periode van drie jaar veranderde het overheidsbeleid namelijk vaak én op korte termijn. Van intelligente lockdown tot avondklok, van mondkapje en anderhalve meter tot verbod op “samenscholing” van drie of meer personen: een nieuwe prioritering werd vanuit de overheid van de één op de andere week – soms dag – opgelegd. Inspelen op dergelijke ontwikkelingen vergt wendbaarheid en een vermogen om snel te schakelen, echter: een model zoals de casechecker is niet gemaakt om te anticiperen op disruptieve gebeurtenissen. Daarom is het noodzakelijk dat adaptieve capaciteit is ingebed in de organisatie zodat er ruimte is om onzekerheden en onverwachte veranderingen te adresseren (zie pagina 27-29).

Corona-gerelateerde misdrijfzaken en overtredingen (Bron: OM.nl)

1. Mensen die hoesten/spugen richting agenten (209 zaken) en anderen met een publieke taak (81 zaken mbt BOA's, zorgpersoneel en OV-personeel), waarbij de daders beweerden besmet te zijn met het coronavirus.
2. Demonstranten die zich keren tegen het coronabeleid van de regering (werden met name vervolgd voor het niet opvolgen van politiebevel, bedreigingen, en (openlijk) geweld). Samen 623 zaken.
3. Avondklokrellen, 407 zaken
4. Overtreden van de avondklok (80.419 zaken)
5. Onvoldoende afstand houden (+/- 43.937)
6. Overtreden van mondkapjesplicht in het OV en publieke ruimte (8.391)

* In de casussen zijn aannames gemaakt zodat we concreet kunnen illustreren wat de impact is op de casechecker.

Casus 2: COVID

Gevolg

Door de pandemie zag de samenleving er drastisch anders uit dan normaal. Hierdoor vonden bepaalde delicten juist in mindere mate of andere vorm plaats. Zo is het aantal huisinbraken flink gedaald omdat iedereen thuis moest blijven maar heeft diefstal zich juist verplaatst naar het cybercrimedomein omdat mensen zich meer begaven in het online domein: voor werk, om sociale contacten te onderhouden, en om te winkelen en boodschappen te doen. Ook was het mogelijk dat overtredingen tijdens corona van een ander kaliber waren dan ervoor. Neem bijvoorbeeld het samenscholingsverbod. Van samenscholing is in de regel sprake wanneer een groep een bepaalde dreiging vormt voor de openbare orde, wat samenhangt met wanorde of een vechtpartij. Echter, tijdens de Coronapandemie vormde een groep van drie mensen uit verschillende huishoudens een dreiging voor de maatschappelijke gezondheid, zonder dat er sprake was van geweld of fysieke dreiging tot geweld, wat een reden kon zijn tot aanhouding op heterdaad.

Het snel veranderende coronabeleid, maakt daarnaast dat ook de prioritering snel veranderde. Zo kon het bijvoorbeeld zijn dat een geweldsdelict als gevolg van een avondklokkel eerder in behandeling werd genomen dan andere geweldsdelicten, omdat het OM zo de ernst van de avondklok wilde onderstrepen.

* In de casussen zijn aannames gemaakt zodat we concreet kunnen illustreren wat de impact is op de casechecker. Op deze slide berusten de gevolgen en het veranderen van de aard van een overtreding niet op de werkelijkheid.

Casus 2: COVID

Impact op de case checker

Zou je dus, ten tijde van Corona, gebruik maken van een AI-model dat is getraind op historische data van vijf jaar ervoor, dan loop je tegen een aantal problemen aan (als we uitgaan van de zojuist genoemde voorbeelden).

Doordat de samenleving er zo anders uit ziet, is je data niet **representatief**. Het model zou immers getraind zijn op een dataset waar méér huisinbraken en minder cybercrime in voorkomt. Ook waren de overtreding van de avondklok en de mondkapjesplicht, als noodverordening, een nieuw soort overtreding. Het is aannemelijk dat dergelijke noodverordeningen niet voorkomen in de dataset, laat staan op deze schaal, en dat – door een **herdefiniëring** van strafbare feiten – het overtreden van de mondkapjesplicht met een nieuwe strafcode werd geregistreerd. Dit heeft consequenties voor het model:

Het model zal dus **geen (accurate) voorspelling kunnen maken** voor het overtreden van de nationale avondklok, want die komt niet voor in je data. Hetzelfde geldt voor het overtreden van de mondkapjesplicht. Ook kan het model geen onderscheid maken tussen rellen als gevolg van de avondklok en andere rellen, hoewel het volgens het prioriteringsbeleid wel wenselijk is om avondklokrellen eerder op te pakken.

Hetzelfde kunnen we terugzien bij de veranderde aard van samenscholingszaken. Het is aannemelijk dat het model een patroon herkent waarbij straffen voor samenscholing hoger uitvielen naarmate de groep groter was, bijvoorbeeld omdat het samenhangt met geweld en bedreiging van de openbare orde. Echter is de groepsgrootte in het geval van de coronapandemie niet betekenisvol voor de strafbepaling: een groep van drie personen uit verschillende huishoudens is namelijk een grotere bedreiging voor de maatschappelijke gezondheid dan een groep van twaalf personen uit een huishouden, zoals een studentenhuus. **Het model kan hier niet op anticiperen.**

In het kort: **De casechecker kan geen accurate strafeis voorspellen voor de strafzaken die in hoge mate – of zelfs het meeste – voorkomen.**

* In de casussen zijn aannames gemaakt zodat we concreet kunnen illustreren wat de impact is op de casechecker. Op deze slide is de herdefiniëring van overtredingen fictief, er is geen fact-check gedaan op de strafcode voor het overtreden van de mondkapjesplicht.

Casus 2: COVID

Aanbevelingen

Het voorbeeld van de coronapandemie laat zien dat **bij disruptieve gebeurtenissen in de samenleving je niet op je model kan vertrouwen**, en dat je je ervan bewust moet zijn dat de data waarop het model is getraind niet overeenkomt met hoe de samenleving er werkelijk uit ziet.

Disruptieve gebeurtenissen veranderen dus hoe en 'of' je het model (de casechecker) kan *gebruiken*, maar zorgen ook voor vertekeningen in de data als je een model wil *trainen*. Er moet dan worden gekeken naar wat er in de maatschappij gebeurde in het tijdsframe van de dataverzameling. Een laag aantal huisinbraken in de dataset zegt in dit geval namelijk niet dat het niet veel voorkomt op het moment dat je het model wilt gaan toepassen. Het lage aantal illustreert de gevolgen van de quarantaine maatregelen, al kan het aantal huisinbraken na de coronapandemie weer stijgen naar het niveau vóór de coronapandemie.

Om te voorkomen dat je wordt verrast, is het dus van belang dat je:

- 1) op de hoogte bent van kritieke gebeurtenissen in het tijdsframe van de dataset, en
- 2) op de hoogte bent van kritieke gebeurtenissen die spelen in de periode waarin je het model wil toepassen.

Zo kan je identificeren wat de kloof is tussen het model en de werkelijkheid, en hierop anticiperen. Dit onderstreept ook het belang van **adaptive governance**, zoals beschreven op pagina's 27-29. Het oog hebben voor maatschappelijke gebeurtenissen zowel nu als ten tijde van de historische dataset, vraagt om adaptief vermogen in de organisatiestructuur. Er moet ruimte en tijd zijn om hier kritisch naar te kijken, en de kennisdeling hierover (bijvoorbeeld tussen verschillende AP's) is zeer waardevol.

* In de casussen zijn aannames gemaakt zodat we concreet kunnen illustreren wat de impact is op de casechecker.

Hoofdstuk 6: Conclusies en aanbevelingen

Conclusies en aanbevelingen

Er is in dit onderzoek gekeken naar de volgende drie onderzoeksvragen.

- Welke **organisatie** en **context specifieke eigenschappen** van het OM vragen extra aandacht in de verantwoorde ontwikkeling, inzet en monitoring van AI?
- Hoe krijg je inzicht in de **context** (maatschappelijke gebeurtenissen, beleid, historie) waarin de data zijn verzameld en de **representativiteit** daarvan voor de toepassing?
- Wat is de **invloed van verschil in beleid van parketten** bij de ontwikkeling, inzet en monitoring van AI binnen het OM?

Aan de hand hiervan zijn er verschillende aanbevelingen opgesteld, deze zijn onderverdeeld naar verschillende onderwerpen en naar verschillende doelgroepen binnen het OM (bestuur, beleid, ontwikkelaars, gebruikers en opleiding)

Organisatie en context specifieke eigenschappen

De ontwikkeling en inzet van AI-toepassingen is geen statisch proces, maar moet continu worden getoetst aan de huidige omstandigheden. Bij deze ontwikkeling en inzet is de context van de organisatie waarin de AI-toepassing wordt gebruikt en de manier waarop zij interacteert met de maatschappij van groot belang. In dit onderzoek zijn de volgende eigenschappen naar boven gekomen:

- **Verschil in werkwijzen en uitvoering van AP's** doordat AP's deels zelf hun beleid kunnen vormgeven;
- **Impact van maatschappelijke ontwikkelingen**, waar het werk van het OM wordt beïnvloed door korte termijn veranderingen in de maatschappij;
- De directe en sterke **samenwerking** met de partners **in de keten** die elkaar beïnvloeden;
- De **normerende rol van het OM** in de maatschappij;
- **Rechtvaardigheid, transparantie en veiligheid** als kern publieke waarden voor het OM die belangrijk zijn bij verschillende afwegingen over publieke waarden.

Deze eigenschappen vragen om **flexibiliteit en wendbaarheid** in de organisatie. Tegelijkertijd is de rechterlijke macht een essentiële bouwsteen van onze maatschappij en is rechtvaardigheid een kernwaarde van onze rechtstaat. Er moet daarom zorgvuldig met flexibiliteit in beleid en de inzet van AI-toepassingen worden omgegaan.

Om te waarborgen dat de inzet van AI-toepassingen de uitvoering van het OM ondersteunt en versterkt is het van belang dat deze **AI-toepassingen betrouwbaar zijn en publieke waarden versterken**.

Daarom bevelen wij aan om **voldoende ruimte te bieden** (middelen, capaciteit en initiatief) **voor experimenteren en leren met AI** hoe dit verantwoord kan en om **kennisdeling in gang te zetten** binnen de organisatie, met organisaties in de keten en met andere overheidspartijen.

Inzicht in representativiteit

Begrip, analyses en kennis van de data die gebruikt wordt zijn essentieel om tot een betrouwbaar en kwalitatief goed model te komen.

Geografische representativiteit is met name van belang wanneer data uit verschillende parketten gecombineerd wordt om een landelijk model op te trainen of er op basis van de data vergelijkingen tussen parketten of tussen landelijk en parketniveau worden gemaakt. Daarom bevelen wij aan om per usecase **de achterliggende aannames en verschillen in kaart te brengen**.

Representativiteit moet vanuit een **accuraatheidsperspectief** (hoe goed past de data bij de huidige situatie) en een **aspiratieperspectief** (in hoeverre willen we op deze manier verder gaan) worden geanalyseerd.

- Voor het **accuraatheidsperspectief** is het noodzakelijk een **analyse uit te voeren van herkomst en de opbouw van de beschikbare data**. Hiervoor adviseren wij om voor elke usecase een analyse uit te voeren met verschillende perspectieven (ontwikkelaar, beleid, gebruiker, etc.) over onder welke aannames de data verzameld is en wat de data beïnvloed kan hebben, maar ook juist wat er mogelijk ontbreekt en niet zichtbaar is in de dataset.
- Het **aspiratieperspectief** vraagt om een **breder verkenning vanuit bestuurlijk- en beleidsperspectief** wat de gewenste aanpak is. Hiervoor is het noodzakelijk dat bestuur en beleid een duidelijke visie en kernwaarden beschrijven wat de gewenste manier van werken en de gewenste uitkomsten zijn van het werkproces.

Het evalueren van representativiteit vraagt om verschillende afwegingen en het is van belang deze **te beoordelen in het licht van het beoogde doel** van de toepassing. Een **vast analysekader** met hulpvragen kan daarbij helpen. De stappen en vragen op de volgende pagina kunnen als basis worden gebruikt door het ontwikkelteam. Het is daarbij van belang om in deze discussies en afwegingen **verschillende perspectieven te betrekken**.

In kaart brengen representativiteit

Deze vragen helpen met de eerste stap om representativiteit in kaart te brengen, maar zijn geen uitputtend overzicht. Het is van belang om verschillende perspectieven (data scientists, gebruikers, beleid etc.) te betrekken in de discussies om een zo'n volledig mogelijk beeld te krijgen.

1) Zorg dat doel, scope, limitaties en werkwijze van het bedoelde model in kaart zijn gebracht

- Sluiten de data en het model aan bij wat we willen bereiken?
- De manier van inzetten (landelijk of regionaal) heeft invloed op de evaluatie van representativiteit van de data. Andersom heeft ook representativiteit invloed op de keuze voor de manier van inzetten.

Dit kan bijvoorbeeld gaan om sensitieve groepen (zoals bepaalde nationaliteiten), maar kan ook gaan om verschillende type zaken

- Welke variabelen zijn aanwezig en welke ontbreken? Hoe beïnvloedt dit de representativiteit?
- Hoe kan de representativiteit beïnvloed worden door foutieve collectie/invoer van data (bv invoeren van 'overig' bij maatschappelijke classificatie)?

2) Analyseer in hoeverre de data representatief is voor het doel van het model en de huidige/gewenste situatie

- Op welke manier speelt representativiteit een rol in deze toepassing, bijvoorbeeld geografisch, temporeel?
- Wat is een robuuste representatie voor deze toepassing? Wanneer is de representativiteit voldoende?
- Uit welke tijdsperiode komt de data? Wat speelde er in deze tijd en is dit representatief voor de huidige situatie en doel? Spelen er in het heden kritieke gebeurtenissen?
- Voor welke groepen moet het representatief zijn? Welke groepen zitten er in de data? Hoe zijn deze groepen vertegenwoordigd? Is er een groep ondervertegenwoordigd of zijn er groepen helemaal niet aanwezig?

3) Definieer vervolgacties

- Bepaal of de representativiteit voldoende is om door te gaan met de ontwikkeling
- Indien nodig, onderneem actie om de representativiteit te verhogen, bijvoorbeeld door een relevante vorm van sampling om extra data te vergaren.
- Bepaal hoe je de representativiteit over tijd wilt monitoren en waarborgen

Generieke conclusies en aanbevelingen

De verschillen in beleid en uitvoering tussen AP's en korte termijn maatschappelijke ontwikkelingen zijn twee aspecten die de representativiteit van de data voor het beoogde doel kunnen beïnvloeden. Hier moet zorgvuldig mee om worden gegaan om te komen tot betrouwbare inzet van AI-toepassingen.

- **Landelijk en AP specifiek beleid:** Zowel voor het inzicht in de data als voor het bepalen van de manier waarop AI wordt ingezet is het van belang dat er afwegingen worden gemaakt over de balans en interactie tussen landelijk en regionaal beleid. Onze aanbeveling is om vanuit het bestuur en beleid duidelijke afspraken te maken over landelijk en AP specifiek beleid, deze te communiceren naar de AP's en het ontwikkelteam en deze expliciet mee te nemen in de ontwikkelfase van nieuwe AI-toepassingen.
- **Adaptieve capaciteit:** De organisatie moet goed om kunnen gaan met onzekerheden en moet onverwachte veranderingen snel kunnen adresseren. Ons advies is om te investeren in de ontwikkeling van adaptieve capaciteit binnen de organisatie.
- **Ontwikkelkeuzes:** Het is van belang om duidelijke afwegingen te maken in de mate waarin je als organisatie wilt kunnen inspelen op maatschappelijke ontwikkelingen en hoe dat moet worden ingeregeld in processen en AI-toepassingen. Hier moet bijvoorbeeld ook in de ontwikkeling van AI-systemen over na worden gedacht om de keuzes tussen verschillende type modellen (bijvoorbeeld zelflerend of statisch) te faciliteren. Een goede afstemming tussen beleid en ontwikkelteam is daarbij essentieel.
- **Monitoring en evaluatie:** AI-toepassingen en de werkwijzen moeten regelmatig worden gemonitord en geëvalueerd om afwijkingen en ontwikkelingen te detecteren. Onze aanbeveling is om de verschillende mogelijkheden voor monitoring en evaluatie te onderzoeken en procedures te ontwikkelen die deze monitoring inregelen.
- **Feedback loops:** Het is van belang om te blijven communiceren tussen technische staf, beleidsmedewerkers en gebruikers om te zorgen dat ontwikkelingen snel worden gedetecteerd en kunnen worden verwerkt in nieuw beleid en dat nieuw beleid of aangepaste doelen snel kunnen worden opgenomen in modellen en de manier van werken. Wij bevelen aan om feedbackloops te organiseren tussen techniek, beleid en gebruikers voor belangrijke beleidsaanpassingen en de herkenning van belangrijke ontwikkelingen.

De Casechecker

In dit onderzoek is de casechecker als leercasus genomen. De casechecker casus is een experiment waarin een algoritme is ontwikkeld door het Testlab OM om de kansen van AI bij prioritering van politierechterzaken (PR-zaken) te kunnen onderzoeken. Het algoritme wordt dus niet ingezet in de praktijk. We gebruiken de casechecker als leercasus in dit onderzoek concreet en praktisch te maken en lessen te trekken over de verantwoorde ontwikkeling en inzet van AI-toepassingen binnen het OM. We hebben daarom inzichten opgedaan die specifiek gelden voor de casechecker. We noemen hier de conclusies en aanbevelingen indien er wordt besloten een vervolgstap te zetten met de casechecker.

Er zijn verschillende uitdagingen voor representativiteit bij de casechecker:

- 1) De data is beïnvloed door verschillend beleid bij de parketten;
onze aanbeveling is om expliciet de verschillen tussen de werkwijzen en het beleid van alle parketten in kaart te brengen.
- 2) Er is ontbrekende of foutieve data in de dataset aanwezig;
onze aanbeveling is om een gedegen analyse te doen van de data kwaliteit en te onderzoeken hoe het vastleggen van de data verbeterd kan worden.
- 3) Maatschappelijk gebeurtenissen zoals Covid-pandemie vertekenen mogelijk de data;
de exacte invloed van de Covid-pandemie op de data is nog niet in kaart gebracht. Een eerste stap zou zijn om te onderzoeken in hoeverre de instroom van zaken en de afhandeling van de prioritering is veranderd in deze periode.
- 4) De gewenste aanpak van prioritering bij PR-zaken wordt veelal bij parketten neergelegd.
We raden aan om duidelijkheid te creëren over wat de gewenste aanpak is om PR-zaken te prioriteren.

Daarnaast is er gebleken dat de werkwijze rondom prioritering van PR-zaken verschilt tussen parketten en dat veel betrokkenen geen zicht hebben op wat landelijk en wat AP specifiek beleid is in deze werkwijze. De gewenste aanpak voor prioritering van PR-zaken lijkt dus veelal bij de parketten te worden neergelegd. Voor een betrouwbare en effectieve ontwikkeling en inzet van de casechecker is het van belang duidelijkheid te creëren over landelijk en AP beleid.

Referenties

Bergman et al. (2023) “Representation in AI Evaluations”. Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency

Chasalow & Levy (2021) “Representativeness in statistics, policy and machine learning.”, Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency, pp. 77-89.

Henriksen et al. (2020) Data representation in LCA: A framework for the systematic assessment of data quality relative to technology characteristics, Journal of Industrial Ecology, 25(8), doi: 10.1111/jiec.13048

Geuijen, K. (2018). “Noodzaak en mogelijkheden voor adaptiviteit in netwerken,” in Adaptief bestuur. Essays over adaptiviteit en openbaar bestuur. pp. 54--62.

Gupta, J. et al. (2010). “The Adaptive Capacity Wheel: A method to assess the inherent characteristics of institutions to enable the adaptive capacity of society,” Environ. Sci. Policy, vol. 13, no. 6, pp. 459–471, 2010, doi: 10.1016/j.envsci.2010.05.006.

Mehrabi, N. et al (2022), “A Survey on Bias and Fairness in Machine Learning, ACM Computing Surveys, vol. 54, no. 6, pp. 1-35

Janssen, M. and van der Voort, H. (2016). “Adaptive governance: Towards a stable, accountable and responsive government,” Gov. Inf. Q., vol. 33, no. 1, pp. 1–5, 2016, doi: 10.1016/j.giq.2016.02.003.

Janssen, M and van der Voort, H. (2020) “Agile and adaptive governance in crisis response: Lessons from the COVID-19 pandemic,” Int. J. Inf. Manage., vol. 55, doi: 10.1016/j.ijinfomgt.2020.102180.

van der Steen, M.A. (2018). “Adaptief bestuur: organiseren voor een voorspelbaar verrassende toekomst,” in Adaptief bestuur. Essays over adaptiviteit en openbaar bestuur.,, pp. 73--84.

TNO. (2022) Adaptive Governance for Transitions and Transformations: Definition, Characteristics and Methods. TNO paper.