Algorithm Audit

Adviesdocument

Voorspellen van ontverantwoord rijgedrag

Algoprudentie
ALGO:AA:2025:01:A

Juli 2025

Hoofdpunten normatieve adviescommissie

> **Modelvaliditeit is essentieel**
Het algoritme moet worden aangepast om specifiek rijgedrag te voorspellen dat tot schade leidt en niet algemeen misbruik van het platform. Zoals bij elk risicovoorspellingsmodel is het van cruciaal belang dat de trainingsdata aansluiten bij het beoogde doel.

> **Balans tussen monitoring en gebruikersautonomie**
Het monitoren van onverantwoord rijgedrag om schade te beperken is een legitiem bedrijfsbelang, maar mag niet ontaarden in buitensporig toezicht of betuttelend advies over rijstijl.

> **Betekenisvolle transparantie vereist**
Gebruikers hebben behoefte aan specifieke uitleg over welk rijgedrag tot een waarschuwing heeft geleid en aan duidelijke handvatten voor verbetering van rijgedrag, niet aan algemene waarschuwingen of verwarrende technische termen die voor de doorsnee platformgebruiker geen betekenis hebben.

> **Zorgvuldige variabeleselectie**
Een te hoge snelheid heeft duidelijke veiligheidsimplicaties, maar variabelen zoals acceleratie zijn lastiger te beoordelen. Dergelijke kenmerken zijn contextafhankelijk en kunnen persoonlijke rijvoorkeuren weerspiegelen. Voor opname van deze variabelen in het model moet overtuigend bewijs zijn dat ze daadwerkelijk bijdragen aan het inschatten van schaderisico en niet slechts verschillende rijstijlen of rij-omgevingen representeren.

> **Menselijke controle blijft essentieel**
Menselijke beslismedewerkers wijken momenteel voor 50–60% af van de aanbevelingen van het risicotaxatie-algoritme, wat wijst op betekenisvolle menselijke tussenkomst in plaats van conformisme. Deze betekenisvolle menselijke toetsing moet behouden bestaan.

Samenvatting advies

De commissie oordeelt dat algoritmische risicovoorspelling voor het identificeren van onverantwoord rijgedrag alleen onder strikte voorwaarden mag plaatsvinden en zorgvuldig moet worden afgewogen tegenover alternatieve methoden om schade te beperken. De validiteit van het voorspellingsmodel is cruciaal; de huidige mismatch tussen het beoogde doel (het voorspellen van onverantwoord rijgedrag) en de gebruikte doelvariabele gedurende training (geblokkeerde gebruikers op het platform vanwege uiteenlopende vormen van misbruik) moet daarom eerst worden opgelost. De commissie benadrukt dat monitoring om schade te beperken een legitiem bedrijfsbelang kan zijn, maar dat dit niet mag ontaarden in buitensporige surveillance of gebruikt moet worden voor betuttelende feedback over persoonlijke rijstijl. Gebruikers dienen concrete en begrijpelijke uitleg te krijgen over welk rijgedrag tot een waarschuwing heeft geleid. Algemene meldingen of lijsten van technische variabelen die voor de meeste gebruikers onduidelijk zijn volstaan niet aan uitlegbaarheidsvereisten. Daarnaast moeten gebruikte variabelen in het model beter onderbouwd worden, waarbij snelheid de meest legitieme variabele is; kenmerken zoals ‘snel optrekken’ of ‘hard remmen’ moet zorgvulidg worden gewogen binnen de (stedelijke) omgeving waarbinnen gereden wordt en er moet overtuigend bewijs worden verzameld dat deze kenmerken daadwerkelijk bijdragen aan schaderisico. Verder adviseert de commissie om betekenisvolle menselijke beoordeling van algoritmische aanbevelingen te behouden, zodat het risico op onterechte waarschuwingen wordt verkleind en gebruikers in bezwaar kunnen gaan tegen een beslissing.

Inhoudsopgave

Over Algorithm Audit

Algorithm Audit is een Europees kennisplatform voor AI bias testing en normatieve AI-standaarden.
De doelen van de stichting zijn vierledig:

Kennisplatform

Samenbrengen van kennis en experts om het collectieve leerproces over de verantwoorde inzet van algoritmes aan te jagen, zie onze white papers en publieke standaarden.

Normatieve adviescommissies

Adviseren over ethische vraagstukken die zich voordoen in de algoritmische praktijk door het faciliteren van deliberatieve en inclusieve adviescommissies, met algoprudentie als resultaat

Technische hulpmiddelen

Implementeren en testen van open source software voor de verantwoorde inzet van algoritmes en AI, o.a. socio-technische evaluatie van generatieve AI, unsupervised bias detectie en synthetische data generatie.

Projectwerk

Ondersteuning bij specifieke vragen vanuit de publieke en private sector over de verantwoorde inzet van algoritmes, zonder winstoogmerk.

1. Voorwoord

Dit advies is het resultaat van de beraadslagingen door een onafhankelijke normatieve adviescommissie. Algorithm Audit heeft dit advies opgesteld op basis van een deliberatieve discussie die tijdens een fysieke bijeenkomst van de adviescommissie heeft plaatsgevonden. Tijdens deze bijeenkomst kwamen verschillende ethische vragen aan bod over het gebruik van algoritmische risicoprofilering om onverantwoord rijgedrag te voorspellen.

Het specifieke geval waarop dit advies is gebaseerd betreft een risicoprofileringmodel dat wordt ingezet door een business-to-consumer autodeelplatform. Met behulp van machine learning wordt een risicomodel (*balanced random forest*) getraind om rijgedrag dat samenhangt met schaderisico aan voertuigen te identificeren. Na iedere rit berekent het model een risicoscore voor de gebruiker. Wanneer een risicoscore een drempel overschrijdt, ontvangt de gebruiker een waarschuwing. Indien het rijgedrag bij volgende ritten niet verbetert, kan het platform de gebruiker na menselijke beoordeling blokkeren voor verdere dienstverlening.

Deze casus is door Algorithm Audit gekozen omdat het een duidelijk voorbeeld is van hoe op machine learning gebaseerde risicoprofilering wordt toegepast in de praktijk, zoals in e-commerce, het bankwezen en human resources – sectoren waar details over deze methoden zelden door organisaties worden gedeeld. Voor dit geval heeft het autodeelplatform gedetailleerde informatie verstrekt over trainingsdata, gebruikte hyperparameters in het balanced random forest-algoritme en de wisselwerking tussen aanbevelingen van het algoritme en menselijke toetsing bij het blokkeren van gebruikers. Alle specificaties zijn terug te vinden in de probleemstelling 'Voorspellen van onverantwoord rijgedrag' (ALGO:AA:2025:01:P).

Met deze casus bouwt Algorithm Audit voort op eerdere onderzoeken naar algoritmische risicoprofilering in de private sector. Een eerdere casestudy (ALGO:AA:2022:01:A) onderzocht het risicomodel van een e-commerceplatform, waarbij zorgen ontstonden over proxy-discriminatie door onderscheid op ogenschijnlijk neutrale variabelen zoals het type simkaart van klanten. Deze casus is interessant omdat het fundamenteel andere ethische overwegingen met zich meebrengt. In de ALGO:AA:2022:01 stond vooral de bescherming van bedrijfsbelangen tegen betalingsfraude centraal. Daarentegen raakt risicoprofilering van onverantwoord rijgedrag niet alleen bedrijfskosten, maar ook de verkeersveiligheid. Bovendien gaan de ethische vraagstukken in dit geval verder dan (proxy-)discriminatie op grond van beschermde wettelijke gronden. Een belangrijk aandachtspunt is de monitoring van rijgedrag en welke variabelen verantwoord kunnen worden ingezet om schaderisico te voorspellen. De combinatie van diepgaande technische informatie en de unieke toepassingscontext maakt deze casus waardevol voor onafhankelijke en deliberatieve evaluatie.

Na uitgebreid onderzoek heeft Algorithm Audit diverse ethische kwesties geïdentificeerd die als meest urgent en relevant worden beschouwd. Als onderdeel van het onderzoek zijn academici, domeinexperts en diverse belanghebbenden geraadpleegd. De resultaten van dit onderzoek vormen de basis voor het beraad van de adviescommissie en zijn te vinden in de probleemstelling 'Voorspellen van onverantwoord rijgedrag' (ALGO:AA:2025:01:P).

Daarnaast heeft Algorithm Audit een focusgroep georganiseerd met gebruikers van deelmobiliteitsplatforms.

Box 1

Algoprudentie: Case-based normatief advies voor ethische algoritmen

Algorithm Audit heeft geen mandaat om juridisch bindende uitspraken of officiële oordelen te geven. In onze casestudies verstrekken we niet-bindend ethisch advies. Ethisch advies gaat vaak verder dan advies over wat wettelijk vereist is. Maar bij het ontbreken van juridische uitspraken of duidelijke standaarden van een toezichthouder, fungeert ons onafhankelijke ethische advies ook als voorlopige wegwijzer voor organisaties. Ons casusadvies kan ook helpen bij het uitwerken van officiële standaarden of toekomstige beslissingen van juridische instanties ondersteunen. In die zin heeft ons ethisch advies ook relevantie voor het juridische domein.

Het is de eerste keer dat Algorithm Audit voor casuïstiek een focusgroep bijeen heeft gebracht om te verkennen hoe gebruikersperspectieven het beste kunnen worden meegenomen in de evaluatie van het risicotaxatie-algoritme. De resultaten van de focusgroep zijn vooraf als algemene input maar niet als bindende richtlijn aan de adviescommissie voorgelegd. Alle bevindingen van de focusgroep zijn te vinden in een apart document, inclusief bijbehorende praktisch aspecten hoe de focusgroep is samengesteld en is samengekomen.

Dit adviesdocument vat de beraadslagingen samen van een groep experts en belanghebbenden die samen de adviescommissie vormen. De commissie is in haar samenstelling divers, waarbij verschillende relevante disciplines en belanghebbenden zijn vertegenwoordigd. De exacte samenstelling van de commissie wordt beschreven in 7. Samenstelling van normatieve adviescommissie. Zowel de adviescommissie als Algorithm Audit hebben dit onderzoek onafhankelijk van het autodeelplatform uitgevoerd. Het onderzoek en het advies zijn niet in opdracht van of gefinancierd door het platform. Het advies van de commissie is niet bindend, maar dient als normatieve leidraad voor alle partijen die zich verdiepen in verantwoord gebruik van algoritmische risicoprofilering in de context van (auto)deelplatforms.

2. Scope van advies

Met betrekking tot dit specifieke geval hebben Algorithm Audit en de adviescommissie een aantal belangrijke ethische kwesties geïdentificeerd die normatief beoordeling behoeven, omdat bestaande regels en richtlijnen niet direct uitsluitsel geven. In dit advies staan de volgende kernthema's centraal:

- **Doel en geldigheid van voorspellingen:** Het bepalen van het legitieme doel van het algoritme, de geschiktheid van het huidige model om dat doel te bereiken en de vraag wanneer monitoring omslaat in ongepaste surveillance en/of betutteling.
- **Transparantie en uitlegbaarheid:** Beoordelen wat gebruikers als een zinvolle uitleg mogen verwachten over de monitoring van hun rijgedrag.
- **Selectie van variabelen:** Welke rijgedragkenmerken kunnen worden gebruikt als input voor het voorspellen van onverantwoord rijgedrag en onder welke voorwaarden?
- **Modelkalibratie en balans tussen fout-positieven en fout-negatieven:** Hoe moet de balans tussen fout-positieve en fout-negatieve voorspellingen worden gewogen, rekening houdend met de impact van beide?

Dat de commissie zich richt op deze punten, betekent niet dat hiermee alle aspecten van verantwoord gebruik van risicoprofileringsalgoritmen door het autodeelplatform zijn afgedekt. Aspecten zoals datakwaliteit, governance, gegevensverwerking, documentatie, besluitvormingsprocessen, rollen en verantwoordelijkheden vallen buiten de reikwijdte van dit advies, niet omdat deze aspecten minder belangrijk zijn, maar omdat bestaande kaders hier al voldoende richting geven. Een uitzondering hierop is modelvaliditeit: dit is minder een normatieve vraag en meer een operationele vereiste en krijgt in dit advies extra aandacht.

De reikwijdte van dit adviesdocument sluit niet geheel aan bij de vragen uit de probleemstelling. In het bijzonder is vraag 3, die gaat over gevoeligheidsanalyses en de invloed van hyperparameterkeuze op modelprestaties, buiten beschouwing gelaten. De commissie beschouwt dit namelijk als een kwestie van vakmanschap die het beste bij de data scientists zelf kan worden gelaten en niet als een urgente normatieve kwestie.

3. Algemene overwegingen en doel van voorspelling

Bij het bepalen hoe algoritmische risicotaxatie op een verantwoorde manier toegepast kan worden om onverantwoord rijgedrag voor het autodeelplatform te identificeren, is het belangrijk om stil te staan bij het doel van dergelijke profilering. Voor deze casus ziet de commissie vooral een belangrijk punt in wat het model daadwerkelijk voorspelt. Het model is oorspronkelijk getraind op gebruikers die om uiteenlopende redenen door het platform zijn geblokeerd, zoals rijgedrag maar ook betalingsachterstanden, veel te laat terugbrengen, of het vuil achterlaten van auto's. Volgens de commissie is er daardoor een discrepantie tussen de doelvariabele die is gebruikt tijdens het trainen van het algoritme (geblokkeerde gebruikers voor uiteenlopende redenen) en het eigenlijke doel van het model (het detecteren van onverantwoord rijgedrag).

De commissie stelt dat deze mismatch de validiteit van het model ondermijnt en opgelost moet worden. Als het doel is om schade door onveilig rijgedrag te voorspellen en te beperken, dan moet het model specifiek getraind worden op gevallen waar schade direct samenhangt met rijgedrag, niet op de bredere groep geblokkeerde gebruikers. Dit is een voorwaarde om het model geschikt te laten zijn voor het beoogde doel. De commissie adviseert daarom het volgende:

> Een heldere definitie van wat onder "onverantwoord rijgedrag" valt, specifiek gericht op rijgedrag dat tot schade leidt;
> Trainingsdata die expliciet het verband leggen tussen rijgedrag en daadwerkelijke schadegevallen;
> Kenmerken die aantoonbaar gerelateerd zijn aan het gedefinieerde risico.

De commissie voert haar beoordeling uit op basis van een toekomstige situatie waarin het risicoprofilerings-algoritme is verbeterd en niet op het huidige model. Alle verdere adviezen in dit rapport zijn dan ook voorwaardelijk: het model moet eerst specifiek kunnen voorspellen welk rijgedrag tot schade leidt, voordat het verantwoord kan worden ingezet. Dat vraagt om zorgvuldige dataselectie, extra aanpassing van kenmerken en mogelijk aanvullende dataverzameling om statistische verbanden tussen rijgedrag en schade te onderbouwen.

Verder vraagt de commissie zich af wanneer monitoring van rijgedrag overgaat in surveillance, die verder reikt dan de kernfunctie van het autodeelplatform: het aanbieden van toegang tot voertuigen. Wanneer het platform zich opstelt als autoriteit op het gebied van rijgedrag en gebruikers zich voortdurend gemonitord en beoordeeld voelen, ontstaat een omgeving die als controlerend en betuttelend kan worden ervaren. Sommige commissieleden menen dat feedback op rijgedrag, bijvoorbeeld door gebruikers te stimuleren veiliger te rijden, een dienst kan zijn, maar benadrukken dat dit wel op een manier moet gebeuren die de autonomie van de gebruiker respecteert. Er is een duidelijk verschil tussen monitoring om schade te beperken (een legitiem bedrijfsbelang) en het opleggen van gedragsnormen die verder gaan dan dat doel.

De commissie onderkent dat sommig rijgedrag, zoals snelheidsovertredingen, direct met veiligheid te maken hebben, terwijl andere, zoals acceleratie, contextafhankelijk zijn en niet altijd een risico met zich meebrengen. Het platform moet daarom terughoudend zijn met het geven van sturing over rijgedrag waarvan de relatie met schade of onveiligheid onbewezen is.

Wat betreft de noodzaak van het huidige algoritmische model, kijkt de commissie eerst naar het bedrijfsbelang. Schadeposten vormen ongeveer 7% (€2M-€3M) van de jaarlijkse omzet (€25M-€45M), dus het is logisch dat het platform deze kosten wil beperken. De platformbeheerder geeft aan dat de schadelast sinds de invoering van het model enigszins is gedaald, maar concrete cijfers ontbreken. Door het uitblijven van bewezen effectiviteit en onvoldoende vergelijking met alternatieven, doet de commissie geen definitieve uitspraak of het huidige model proportioneel en effectief is. Mogelijk zijn eenvoudigere, transparantere methoden net zo effectief, zoals:

> Regelgebaseerde profilering met duidelijke, handmatig ingestelde drempelwaarden;
> Directe waarschuwingen bij concrete incidenten, in plaats van op basis van een algoritmisch risicomodel;
> Andere monitoringsmethoden, bijvoorbeeld gebruikers een voertuig laten inspecteren bij aanvang en einde van het gebruik.

De commissie adviseert het platform om deze alternatieven grondig te onderzoeken voordat wordt vastgehouden aan een complex algoritmisch risicotaxatie-systeem.

4. Transparantie en uitlegbaarheid

Wanneer een risicoprofileringsmethode wordt ingezet om mogelijk onverantwoord rijgedrag te signaleren, is het essentieel dat gebruikers weten dat hun rijgedrag wordt gemonitord en dat beslissingen op basis van de uitkomsten hen duidelijk kunnen worden uitgelegd. Dit principe is cruciaal om het vertrouwen en de legitimiteit van het platform te waarborgen.

Toegang tot de diensten van het platform moet gebaseerd zijn op instemming van de gebruiker, niet alleen voor het verzamelen van persoonsgegevens, maar ook voor het monitoren van individueel rijgedrag. De commissie benadrukt dat deze instemming geïnformeerd moet worden afgegeven. Er bestaat zorg dat het verlenen van toestemming, waarbij gebruikers akkoord gaan met een lang en ingewikkeld privacybeleid waarin het gebruik van een risicotaxatie is opgenomen, ertoe kan leiden dat men formeel toestemming geeft zonder zich daadwerkelijk bewust te zijn van monitoring. De commissie vindt het redelijk dat gebruikers

die niet akkoord gaan met het verzamelen van hun data over hun rijgedrag, zich kunnen afmelden na een melding hierover in de app te hebben ontvangen.

Wanneer een gebruiker een waarschuwing krijgt over het rijgedrag, moet de uitleg hierover concreet genoeg zijn om te begrijpen welk gedrag door het platform als problematisch wordt gezien. Huidige, algemene meldingen over "onverantwoord rijgedrag" zijn hiervoor onvoldoende. De commissie adviseert om waarschuwingen te voorzien van specifieke gedragingen die de waarschuwing hebben veroorzaakt, evenals uitleg over waarom dit gedrag als risicovol wordt gezien en praktische tips voor verbetering. Hoewel een te gedetailleerd overzicht van alle data de gebruiker kan overweldigen, moeten de belangrijkste factoren die tot de waarschuwing hebben geleid, duidelijk worden gecommuniceerd, zodat bijvoorbeeld duidelijk is dat hard remmen of te hard rijden tijdens een bepaalde rit de aanleiding was van de waarschuwing.

De commissie vindt het huidige overzicht van gebruikte, geaggregeerde kenmerken niet geschikt om op deze manier aan gebruikter inzicht te verschaffen. Zo bevatten de variabelen rondom snelheid veel overlap en zijn de gebruikte categorieën (zoals 'totaal aantal rijgebeurtenissen') technisch en niet direct inzichtelijk voor de gemiddelde gebruiker (zie Box 2). Alleen de vermelding van het kenmerk dat het meeste bijdroeg aan de risicoscore is volgens de commissie daarom geen zinvolle uitleg. Het is aan te bevelen om de data te verwerken tot bruikbare en begrijpelijke categorieën, waarmee beter richting kan worden gegeven aan gebruikers. Meer technische details kunnen altijd op verzoek verstrekt worden.

De commissie merkt op dat de manier van communiceren sterk bepaalt hoe gebruikers het platform ervaren. Uit ervaringen van gebruikers blijkt dat de huidige communicatie als streng en controlerend wordt ervaren, wat kan leiden tot gevoelens van onrust en kan ontmoedigen om het platform verder te gebruiken. Vooral als

Box 2

Vertaal variabelen naar begrijpbare kenmerken

Geen betekenisvolle uitleg

Tijdens je laatste ritten hebben we de volgende afwijkingen geconstateerd:

TOTAL_DRIVING_EVENTS_PERKM

Het risico op schade aan onze voertuigen wordt hierdoor vergroot. Wij vragen u vriendelijk om dit te gedrag aan te passen en toekomstige schade te voorkomen.

Wel betekenisvolle uitleg

Tijdens je laatste ritten hebben wij uw gedrag geobserveerd en het volgende opgemerkt:

> Overmatig hard remmen;
> Bochten met hoge snelheid nemen.

Het risico op schade aan onze voertuigen wordt hierdoor vergroot. Wij vragen u vriendelijk om dit te gedrag aan te passen en toekomstige schade te voorkomen.

TOTAL_DRIVING_EVENTS_PERKM is een combinatie van de variabelen TOTAL_CORNERING_EVENTS_PERKM en TOTAL_BRAKING_EVENTS_PERKM

men zich niet bewust is van monitoring, kunnen plotselinge waarschuwingen als indringend en achterdochtig overkomen, in het bijzonder als opvallend rijgedrag – bijvoorbeeld hard remmen – soms legitiem verklaarbaar is. De commissie adviseert daarom dat de communicatie uitnodigend en samenwerkend is, gericht op het gezamenlijke belang van veilig rijden en goed onderhoud van voertuigen, waarbij de gebruiker wordt gestimuleerd tot verantwoord gedrag.

De commissie doet bewust geen uitspraak over de precieze vorm van uitleg die het meest geschikt is, omdat dit aan de gebruikers zelf moet worden gevraagd. Er wordt aangeraden om verschillende alternatieven te testen en hierover feedback op te halen bij de gebruikers. Ook het hiervoor benoemde risico op betutteling moet in deze feedback worden meegenomen.

Een belangrijk doel van transparante communicatie over waarschuwingen is het mogelijk maken van bezwaar. De commissie adviseert het platform om eenvoudige en toegankelijke procedures voor bezwaar en correctie te bieden. Het is positief dat menselijke beslismedewerkers nu al de algoritmische voorspellingen beoordelen voordat er waarschuwingen worden opgelegd of gebruikers worden geblokkeerd. Gebruikers moeten weten dat het algoritme niet geautomatiseerd besluiten neemt en dat er altijd sprake is van menselijke beoordeling, evenals duidelijke mogelijkheden om bezwaar te maken of om extra toelichting te vragen.

Daarnaast kan transparantie gebruikers meer inzicht geven in hun rijgedrag. De commissie waarschuwt wel dat continue feedback snel kan overslaan in betutteling, maar erkent dat sommige gebruikers het juist op prijs als zij inzicht kunnen krijgen in hun rijgedrag. Het ontwikkelen van een online dashboard waar gebruikers hun rijstatistieken kunnen bekijken, kan uitkomst bieden. Zeker voor gebruikers die eerder een waarschuwing ontvingen, kan het geruststellend zijn om te zien dat hun rijgedrag is verbeterd en dat hun risico op blokkering verminderd is. Dit bevordert het vertrouwen in het platform en maakt het monitoringsproces inzichtelijker.

5. Variabeleselectie voor profilering

Er vanuitgaande dat de aanpassingen zoals beschreven bij sectie 1 over de validiteit van het model zijn doorgevoerd, waardeert de commissie dat bij risicoprofilering uitsluitend rijgedragdata worden gebruikt en geen andere persoonsgegevens, zoals leeftijd, postcode of het aantal jaren dat iemand een rijbewijs heeft. Ook vindt de commissie het positief dat het platform werkt met geaggregeerde data (zoals het aantal snelheidsovertredingen per gereden kilometer) en niet met exacte details van individuele overtredingen, wat de privacy van gebruikers enigszins beschermt.

Toch is het belangrijk zorgvuldig te wegen welke rijgedragvariabelen geschikt zijn om op te nemen in het risicomodel. De commissie raadt aan het aantal variabelen tot een minimum te beperken en alleen variabelen te gebruiken waarbij er een direct en inhoudelijk verband is tussen rijgedrag en het risico op schade. Ook dient gekeken te worden naar mogelijke samenhang tussen rijgedrag en beschermde discriminatiegronden, zoals etniciteit en geslacht, zodat het model niet onterecht slechter presteert voor bepaalde groepen.

Onder de rijgedragskenmerken beschouwt de commissie (geaggregeerde) snelheidsovertredingen als de meest gerechtvaardigde variabele in het model. Er is immers een directe en inhoudelijke relatie tussen te hard rijden en onveilig rijgedrag, wat kan leiden tot ongevallen en schade. Snelheidsovertredingen zijn doorgaans objectief vast te stellen en gebruikers kunnen hier zelf invloed op uitoefenen, waardoor dit een redelijke basis vormt voor risicobeoordeling.

Bij het gebruik van snelheidsdata wijst de commissie wel op mogelijke problemen met de betrouwbaarheid. GPS-fouten kunnen ertoe leiden dat een snelheidsovertreding ten onrechte wordt geregistreerd. Het platform probeert dit te ondervangen door onwaarschijnlijke snelheidsmetingen te filteren, maar dit blijft een aandachtspunt. De commissie adviseert te blijven werken aan nauwkeurigere GPS-metingen, geavanceerdere filters en andere manieren om fouten te verminderen. Transparantie, goede uitleg en eenvoudige bezwaarprocedures (zie sectie 2) zijn nodig, zodat gebruikers foute metingen eenvoudig kunnen melden.

Voor andere variabelen zoals snel optrekken, hard remmen en scherp sturen, vindt de commissie het belangrijk om terughoudend te zijn. Dergelijk rijgedrag hoeft niet altijd op onverantwoord rijdrag te wijzen en kan afhankelijk zijn van de verkeerssituatie, weersomstandigheden of onvoorziene gebeurtenissen. Hard optrekken is bijvoorbeeld niet gevaarlijk als je bij groen licht als eerste wegrijdt op een lege kruising. Hard remmen kan juist een teken zijn van oplettendheid in onverwachte situaties.

Verder merkt de commissie op dat de omgeving waarin wordt gereden van invloed is op deze variabelen. In de stad of tijdens spitsuren zal vaker stevig moeten worden geremd dan op het platteland, ongeacht hoe verantwoord iemand rijdt. Dit roept zorgen op over mogelijke onbedoelde nadelen voor mensen die vooral in stedelijke gebieden of bepaalde buurten rijden.

Hoewel de commissie niet aanbeveelt om deze variabelen volledig uit te sluiten, vindt zij dat het gebruik ervan goed onderbouwd moet worden met bewijs dat ze daadwerkelijk bijdragen aan het voorspellen van schade. Het platform zou hiervoor data verder moeten analyseren, bijvoorbeeld uitgesplitst naar groepen, en nagaan of deze variabelen niet indirect een effect hebben op gevoelige kenmerken. Ook adviseert de commissie om representatieve focusgroepen te raadplegen, waaronder mensen uit gemarginaliseerde gemeenschappen en mensen met een medische aandoening, omdat bepaald rijgedrag – zoals hard remmen – daar een andere oorzaak kan hebben. Deze diverse gebruikersgroepen moeten betrokken worden bij het ontwikkelen van het algoritme en het bepalen welke variabelen relevant zijn.

Tot slot dringt de commissie erop aan het principe van dataminimalisatie toe te passen: gebruik alleen de variabelen die noodzakelijk en proportioneel zijn voor het doel van schadevoorspelling. Het platform zou een zorgvuldige selectie van variabelen moeten onderbouwen en documenteren, waarbij inzichtelijk wordt gemaakt welke variabelen daadwerkelijk bijdragen aan de nauwkeurigheid van het model en welke zonder grote gevolgen kunnen vervallen. Dit voorkomt dat gebruikers onterecht worden benadeeld voor gedrag dat nauwelijks risico oplevert én bevordert de transparantie en uitlegbaarheid van het model.

6. Modelkalibratie: balanceren van FP's en FN's

Bij het beoordelen van een risicovoorspellingsmodel is het van groot belang om stil te staan bij hoe het model wordt gekalibreerd, met name als het gaat om de afweging tussen fout-positieven (FP's) en fout-negatieven (FN's). Deze percentages zijn doorgaans met elkaar verbonden: een verlaging van het aantal FP's (en dus een stijging van het aantal terecht-positieven) leidt vaak tot een verhoging van het aantal FN's – zoals weergegeven in Figuur 1. Om deze balans goed te kunnen beoordelen, kijkt de commissie eerst naar de gevolgen van beide fouttypes: FP's (verantwoordelijke bestuurders die onterecht als onverantwoordelijk

worden aangemerkt) en FN's (onverantwoordelijke bestuurders die onterecht als verantwoordelijk worden beschouwd). Zie ook Box 3.

De commissie merkt op dat het minimaliseren van FN's aansluit bij het primaire doel van het risicotaxatie-algoritme. Bij de weging van deze risico's moeten veiligheid en bedrijfsbelangen worden afgewogen tegen gebruikerservaring en vertrouwen. Hoewel er zorgen zijn over zowel FP's als FN's, is er brede consensus dat een menselijke beoordeling veel van de nadelen van FP's kan ondervangen. Een zorgvuldige menselijke check zorgt ervoor dat een aanzienlijk deel van de FP's tijdig wordt opgemerkt, zodat er niet onterecht waarschuwingen worden verstuurd naar verantwoordelijke bestuurders.

De commissie stelt dat het zonder verbetering van de modelvaliditeit niet mogelijk is om de scenario's uit de probleemstelling (Figuur 1) met verschillende terecht- en fout-positieven-percentages goed te beoordelen. Desondanks doet de commissie enkele suggesties. Bij het beoordelen van deze percentages is het nuttig om abstracte meetwaarden te vertalen naar concrete gebruikerservaringen: een FP-percentage van 0,0157 betekent dat ongeveer 1 op de 64 ritten onterecht als overtreding wordt geflagd; een veelrijder wordt dan ongeveer eens per twee maanden onterecht gewaarschuwd. Vanuit deze zienswijze stelt de commissie voor dat maximaal één onterechte waarschuwing per jaar voor een veelrijder een acceptabele grens vormt. Vervolgonderzoek onder gebruikers is nodig om beter te bepalen wat als acceptabel wordt ervaren. De commissie benadrukt dat deze beoordeling afhankelijk is van een effectieve menselijke beoordeling en eerlijke behandeling van alle gebruikersgroepen.

Box 3

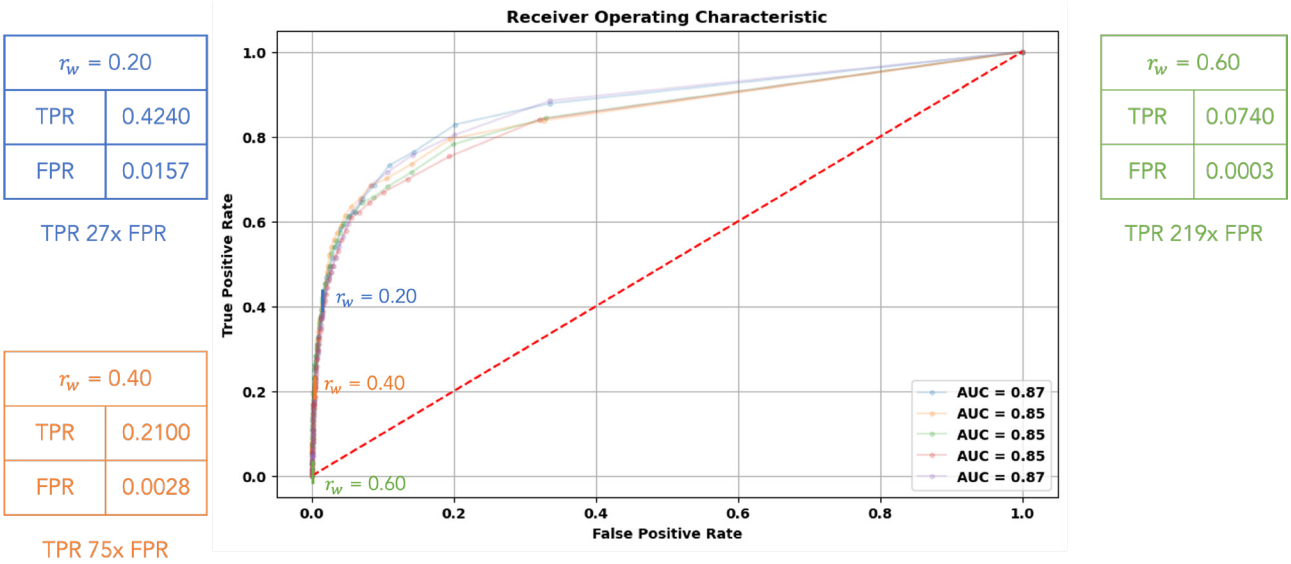
Betekenis van fout-positieven (FP's) en fout-negatieven (FN's)

Overwegingen met betrekking tot fout-negatieven (FN's):

> Materiële risico's voor het platform door het niet signaleren van gebruikers die eerder geneigd zijn om voertuigen te beschadigen, wat leidt tot hogere schadekosten;
> Verminderde verkeersveiligheid voor alle weggebruikers en het algemene publiek;
> De voertuigen van het platform kunnen worden geassocieerd met onverantwoord rijgedrag op de weg.

Overwegingen met betrekking tot fout-positieven (FP's):

> Gebruikers die onterecht worden beschuldigd van onverantwoord rijgedrag;
> Sterker gevoel gesurveilleerd te worden;
> Ervaring dat het platform oneerlijk is;
> Risico dat gebruikers overstappen naar concurrerende platforms vanwege negatieve ervaringen, incl. verlies van reputatie van het platform;
> Hogere kosten door het inzetten van meer menselijke analisten om FP's te beoordelen;
> Bepaalde groepen worden mogelijk disproportioneel vaak onterecht geclassificeerd, wat leidt tot bias en ongelijke behandeling;
> Gebruikers nemen waarschuwingen niet serieus als er te vaak onterecht wordt gewaarschuwd.



Figuur 1 - ROC-curve van het *balanced random forest* (BRF)-voorspellingsmodel voor 5-fold kruisvalidatie.

Ten tweede, als de hierboven genoemde scenario's als uitgangspunt worden genomen (ook al veranderen deze als de modelvaliditeit wordt verbeterd), merkt de commissie op dat het FP-percentage momenteel laag is. Op dit moment beoordelen menselijke analisten alle door het model gemarkeerde gevallen en in circa 50-60% van de gevallen wijken beoordelaars af van de aanbevelingen van het model. Dit wordt gezien als een positief teken: er wordt daadwerkelijk kritisch gekeken naar de modeluitkomst in plaats van deze automatisch over te nemen. Hierdoor lijken de risico's van FP's relatief goed te worden gemitigeerd, waardoor er meer ruimte is om de focus te leggen op het terugdringen van FN's. Wel raadt de commissie aan om extra analyses te doen naar groepen die mogelijk relatief vaker een FP krijgen. Naast het voorkomen van schade is het immers ook voor de algemene verkeersveiligheid van belang om écht onverantwoord rijgedrag te signaleren.

Deze aanbevelingen gelden onder de voorwaarde dat het menselijke element in het reviewproces behouden blijft. Beoordelaars moeten voldoende informatie en training krijgen om tot een eerlijke beoordeling te komen en het is raadzaam motiveringen voor beslissingen goed te documenteren voor verantwoording, consistentie en verbetering. Daarnaast zijn de aanbevelingen uit sectie 2 van belang om de risico's van FP's te beperken, bijvoorbeeld door waarschuwingen constructief, behulpzaam en minder wantrouwend te formuleren richting gebruikers. Zo wordt het ontvangen van een waarschuwing minder belastend, zelfs voor gebruikers die in werkelijkheid verantwoord rijden.

7. Samenstelling van normatieve adviescommissie

Dit advies is het resultaat van een gezamenlijk deliberatief proces. Specifieke standpunten in dit document hoeven daarom niet volledig overeen te komen met de persoonlijke mening van individuele leden van de advoescommissie. Commissieleden kunnen niet individueel verantwoordelijk worden gehouden voor dit advies.

Datum

De auditcommissie is op 13 januari 2025 fysiek bijeen gekomen in Den Haag. Dit adviesdocument is door alle leden van de adviescommissie goedgekeurd op 26 juni 2025.

Samenstelling van adviescommissie

De normatieve adviescommissie voor deze casus bestaat uit:

> Cynthia Liem, Associate Professor Multimedia Computing Group, TU Delft
> Hilde Weerts, Assistant Professor Fair and Explainable Machine Learning, TU Eindhoven
> Joris Krijger, AI & Ethics Officer, De Volksbank
> Maaike Habers, Professor of Applied Sciences (lector) Artificial Intelligence & Society, Hogeschool van Rotterdam
> Monique Steijns, oprichter van The People's AI agency
> Anne Rijlaarsdam, gebruiker van autodeelplatform.

Een data scientist van het deelautoplatform was eveneens aanwezig tijdens de commissiebijeenkomst om feitelijke vragen te beantwoorden, maar maakt geen deel uit van de adviescommissie.

Dankwoord

Naast de vele mensen met wie we hebben gesproken of die ons werk aandachtig hebben gelezen, willen we bijzondere dank uitspreken aan de volgende personen en organisaties voor hun waardevolle bijdragen aan dit project:

> Vardâyani Djwalapersad
> Tom Driessen
> Joel Persson

Over Algorithm Audit

Algorithm Audit is een Europees kennisplatform voor AI bias testing en normatieve AI-standaarden.
De doelen van de stichting zijn vierledig:

Kennisplatform

Samenbrengen van kennis en experts om het collectieve leerproces over de verantwoorde inzet van algoritmes aan te jagen, zie onze white papers en publieke standaarden.

Normatieve adviescommissies

Adviseren over ethische vraagstukken die zich voordoen in de algoritmische praktijk door het faciliteren van deliberatieve en inclusieve adviescommissies, met algoprudentie als resultaat

Technische hulpmiddelen

Implementeren en testen van open source software voor de verantwoorde inzet van algoritmes en AI, o.a. socio-technische evaluatie van generatieve AI, unsupervised bias detectie en synthetische data generatie.

Projectwerk

Ondersteuning bij specifieke vragen vanuit de publieke en private sector over de verantwoorde inzet van algoritmes, zonder winstoogmerk.

Structurele partners van Algorithm Audit

SIDN Fonds

Het SIDN Fonds staat voor een sterk internet voor iedereen. Het Fonds investeert in projecten met lef en maatschappelijke meerwaarde, met als doel het borgen van publieke waarden online en in de digitale democratie.

European AI&Society Fund

Het European AI&Society Fund ondersteunt organisaties uit heel Europa die AI beleid vormgeven waarin mens en maatschappij centraal staan. Het fonds is een samenwerkingsverband van 14 Europese en Amerikaase filantropische organisaties.

Opbouwen van *publieke kennis*

over verantwoorde AI *zonder winstoogmerk*

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Parkstraat 22, 2514 JK Den Haag

Algorithm Audit

Stichting Algorithm Audit is geregistreerd bij de
Kamer van Koophandel onder nummer 83979212