Automatische voertuigen – Hoe blijft de mens de techniek de baas?

De rol van Meaningful human control bij het ontwerpen en reguleren van automatische voertuigen

Zelfrijdende voertuigen die ons soepel, snel en geheel automatisch van A naar B brengen – technisch gezien zijn we er al bijna. Maar voordat we deze slimme auto's op de weg laten, moeten we er wel voor zorgen dat het voldoende *veilig, beheersbaar en verantwoord* blijft. Hoe kunnen autofabrikanten, wegbeheerders en beleidsmakers daar samen aan werken? In deze whitepaper presenteren we hiervoor het raamwerk Passend menselijk toezicht, oftewel *Meaningful human control*.

Auteurs:
Simeon C. Calvert, Stig Johnsen en Ashwin George

TU Delft DiTT Lab

Deze whitepaper is een bewerking van het wetenschappelijke artikel *Designing Automated Vehicle and Traffic Systems towards Meaningful Human Control*. Deze kan geciteerd worden als:
Calvert, S.C., S. Johnsen, A. George, 2023. Designing Automated Vehicle and Traffic Systems towards Meaningful Human Control. In: Research Handbook on Meaningful Human Control of Artificial Intelligence Systems. Edward Elgar Publishing.

Introductie

De discussie over de *veiligheid* van automatische voertuigen is geen theoretische. Automatische voertuigen die een vrachtwagen uit de andere richting niet opmerken, een tunnel niet herkennen of die midden op de snelweg stil gaan staan – ze halen met enige regelmaat het nieuws. Dat die incidenten de zaak van de zelfrijdende auto geen goed doen, is nog tot daaraantoe. Het echte probleem is natuurlijk dat dit soort missers mensenlevens kunnen kosten.

Nu wordt wel aangevoerd dat er altijd iets mis kan gaan, met of zonder automatisering. Het voordeel van het automatische voertuig zou bovendien zijn, dat de techniek steeds intelligenter wordt en dat automatisch rijden dus alleen maar veiliger wordt. Gelet op de vooruitgang die de afgelopen jaren is geboekt, is dat geen onlogische gedachte. Maar we zouden met dat vertrouwen in de techniek als oplossing wel voorbijgaan aan twee andere problemen van verregaande automatisering, die van *beheersbaarheid* (controle) en *verantwoording*. De complexe, zelflerende algoritmes waar een automatisch voertuig op draait, kunnen met een te sterke focus op techniek namelijk makkelijk in een 'black box' veranderen. Voor menselijke gebruikers is dan steeds lastiger te vatten hoe het systeem tot z'n keuzes komt. Dat zou in bijzondere omstandigheden tot onverwacht en onvoorzien rijgedrag kunnen leiden. En als het voertuig hierdoor een ongeval veroorzaakt, is nauwelijks vast te stellen wie of wat daar schuld voor draagt. Dat is politiek en maatschappelijk gezien onaanvaardbaar.

Wie is de baas?

De kwestie van veiligheid is dan ook alleen goed op te pakken, als we de meer fundamentele kwestie daaronder beschouwen: hoe blijft de mens de machine de baas?

Dit punt is voor de huidige generatie automatische voertuigen nog wat rigoureus geregeld: de bestuurder van een automatisch voertuig is te allen tijde verantwoordelijk. Hij kan gebruikmaken van de (uitgebreide) rijtaakondersteuning, maar wordt wel geacht direct in te grijpen als dat nodig mocht zijn. Hoewel deze verplichting op dit moment nuttig is, is het geen oplossing voor de langere termijn. Los van de vraag of het redelijk is te verwachten dat een 'meerijder' op ieder willekeurig moment in staat is het stuur over te nemen, past de oplossing simpelweg niet in het eindplaatje van automatische voertuigen. Uiteindelijk willen we toch een voertuig waarin we rustig de krant kunnen lezen of een mail tikken, zonder dat we voortdurend met een schuin oog naar het verkeer hoeven kijken.

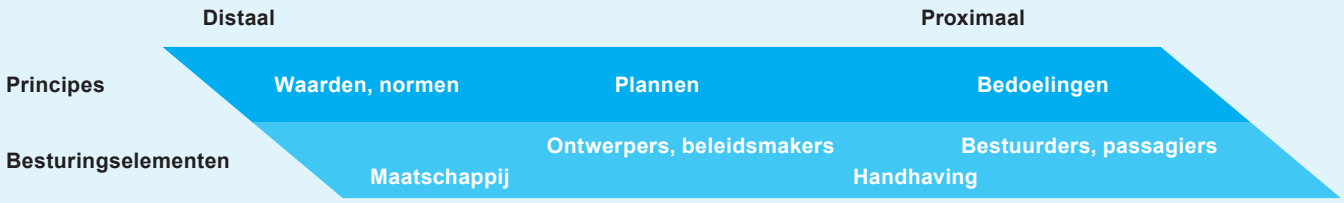
Aan de TU Delft zijn we daarom op zoek gegaan naar een bredere en robuustere aanpak. We hebben het concept *Meaningful human control* als vertrekpunt genomen en doorontwikkeld naar een 'denkraam' dat geschikt is voor automatische voertuigen. In het onderstaande lichten we het raamwerk toe en beschrijven we met enkele voorbeelden hoe deze benadering autofabrikanten, wegbeheerders en beleidsmakers kan helpen om de basis te leggen voor een veilige én beheersbare en verantwoorde inzet van automatische voertuigen.

Meaningful human control

Het begrip *Meaningful human control* werd voor het eerst gebruikt in 2015, in de context van autonome (automatische) aanvalswapens. Vooraanstaande wetenschappers, ondernemers en beleidsmakers riepen toen op tot een verbod op "autonome aanvalswapens die niet onder *passend menselijk toezicht* staan". Hiermee bedoelden ze dat geautomatiseerde systemen zo opgezet en ingericht moeten zijn dat altijd mensen, en niet computers en hun algoritmen, de controle houden over de beslissingen. Zo blijven ook altijd mensen moreel verantwoordelijk voor het handelen van de systemen.

Tracking en tracing

Dit uitgangspunt is later op veel meer (impactvolle) automatische systemen toegepast – onder meer op automatische voertuigen. In latere studies naar de mogelijkheden van *Meaningful human control* voor de verkeer- en vervoersector zijn twee voorwaarden beschreven om dit menselijke toezicht op orde te krijgen, *tracking* en *tracing*. Met het eerste wordt bedoeld dat menselijke principes en intenties altijd leidend moeten zijn. Algoritmes en systemen in automatische voertuigen mogen dus geen 'eigen' afwegingen maken, maar moeten zo zijn ingericht dat ze menselijke overwegingen en maatstaven volgen. Die kunnen heel fundamenteel zijn, van het type 'mensen geen schade toebrengen', of specifieker en subjectiever, zoals 'comfortabel rijden'.



Figuur 1 *Raamwerk voor Meaningful human control.*

De tweede voorwaarde, *tracing*, gaat over toezicht en controle: er moet altijd iemand (een persoon) zijn die direct of indirect toeziet en daarmee ook verantwoordelijk is voor het gedrag van het automatische systeem. Bij een automatisch voertuig kan dat de reiziger zijn, een medewerker in een controlecentrum of, in meer indirecte zin, de programmeur/ontwerper van het voertuig.

Figuur 1 biedt een schematische weergave van de twee voorwaarden. Het deel *Principes* betreft *tracking* en het deel *Besturingselementen* is waar de *tracing* plaatsvindt.[1] Merk op dat de figuur onderscheid maakt naar 'afstand': van meer fundamentele tot persoonlijke principes, en van toezicht op (letterlijk) afstand tot controle in het voertuig. We spreken in dit verband ook wel van distaal en proximaal.

Integraal raamwerk voor Meaningful human control

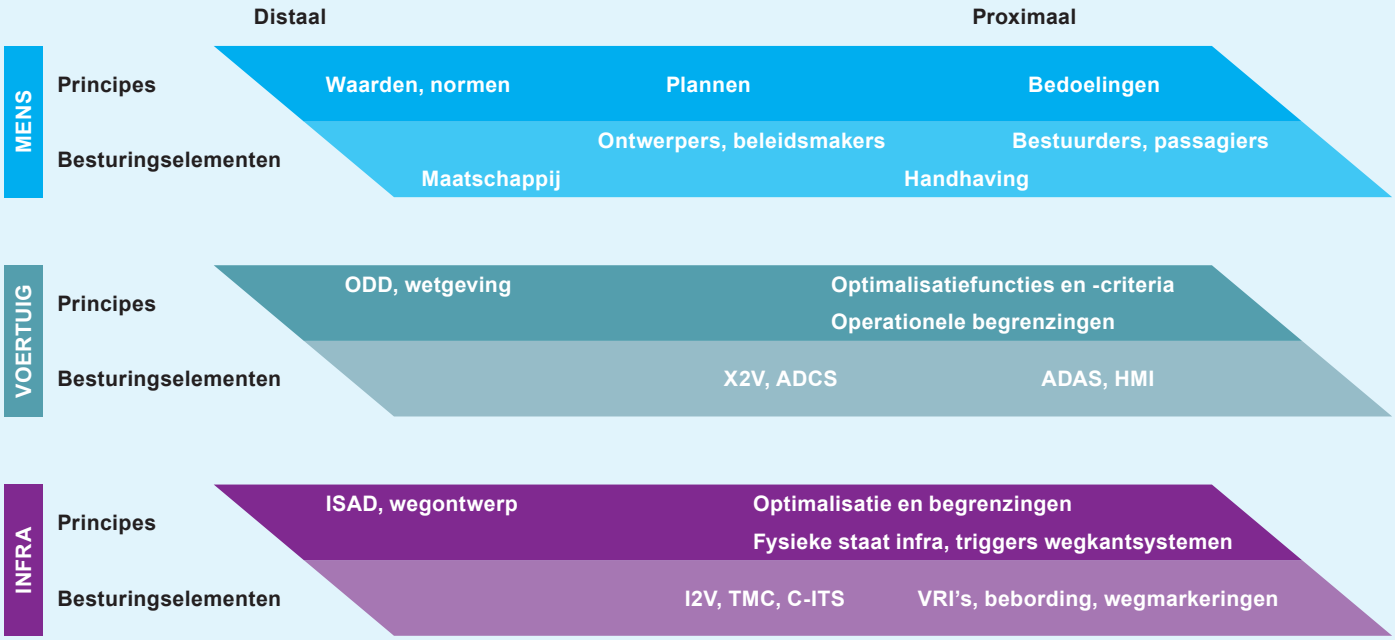
Nu beperken we ons met Figuur 1 wel erg tot de factor mens. Om het systeem van automatisch rijden meer in z'n totaliteit te kunnen beschouwen, hebben we daarom aan de 'menselijke laag' twee lagen toegevoegd, die van het voertuig en van de infrastructuur – zie Figuur 2.

Voertuig-laag

Met de twee extra lagen kunnen we expliciet maken waar de menselijke principes moeten 'landen' in het voertuig en de infra. In de Voertuig-laag staan bij *Principes* de voorbeelden *ODD*[2] en *wetgeving*. Dit zijn inderdaad belangrijke 'omgevingen' waarin menselijke principes kunnen (moeten) worden geïntegreerd om de goede werking van automatische voertuigen te garanderen. Stel bijvoorbeeld dat de sensoren van een zeker voertuigmodel moeite hebben met mindere lichtomstandigheden. Rijden in het schemerdonker zou dan tot een verhoogde kans op ongevallen leiden en dus botsen met het menselijke principe 'mensen geen schade toebrengen'. Dat principe kun je laten landen in het *ODD* van het betreffende voertuigmodel door alleen 'rijden bij voldoende daglicht' op te nemen; de *wetgeving* kan het betreffende principe ondersteunen met een expliciet verbod om een automatisch voertuig buiten z'n ODD te gebruiken.

ODD en wetgeving zijn nog domeinen 'op afstand'. Dichterbij, op het operationele niveau, komen menselijke principes terug in de manier waarop het voertuig z'n besturingsfuncties optimaliseert en/of inperkt om in een omgeving te interageren. Een voertuig kan bijvoorbeeld extra voorzichtig zijn afgesteld, met een lage maximumsnelheid en een ruime afstand tot andere voertuigen.

1 De figuur is gebaseerd op het *Fundamental diagram of meaningful human control proximity* van Santoni de Sio and Mecacci (2021). 'Ontwerpers, beleidsmakers' is toegevoegd.
2 Het ODD, *Operational Design Domain*, van een automatisch voertuig beschrijft onder welke omstandigheden het automatisch kan rijden (waar is het voor geschikt?).



Figuur 2 *Integraal raamwerk voor Meaningful human control.*

Deze menselijke principes op voertuigniveau zijn weer leidend voor de besturingselementen van het voertuig. De ontwerpers en programmeurs van het *Automated Driving Control System*, het hart van een zelfrijdend voertuig, en van *Advanced Driving Assistance Systems* zullen zich bijvoorbeeld naar het ODD, de geldende wetgeving en de kaders van de besturingsfuncties moeten voegen. Zo dalen de principes vanzelf in het voertuiggedrag in.

Infra-laag

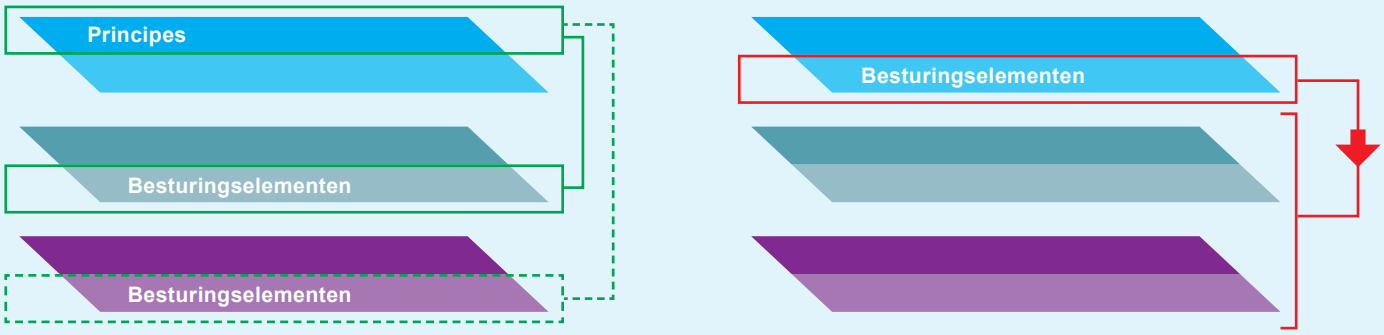
Iets soortgelijks geldt voor de laag Infra, waarmee we zowel de fysieke als de digitale infrastructuur bedoelen. De relaties met de menselijke laag zijn wat minder sterk, maar ze zijn er wel. *ISAD-niveaus*[3] maken bijvoorbeeld expliciet waar welke rijtaak-ondersteuning mogelijk en dus veilig is. En een goede belijning (*wegmarkeringen* in de figuur) zorgt ervoor dat een automatisch voertuig alleen daar rijdt waar het veilig is. Beide gevallen betreffen het principe van 'mensen geen schade toebrengen'.

In het deel *Besturingselementen* van de Infra-laag vinden we onder meer de *verkeerscentrale* (TMC) en de *verkeersregelinstallatie*, die op hun wijze bijdragen aan het juist functioneren van een automatisch voertuig. Zo kan de verkeerscentrale een voertuig begeleiden dat 'vastloopt'. Stel bijvoorbeeld dat dat door dubbel geparkeerde auto's een automatisch voertuig niet kan passeren zonder een doorgetrokken streep over te gaan. Op eigen initiatief zal het voertuig die streep niet passeren, vanwege z'n veiligheids-instellingen. De centrale kan dan een akkoord geven om in die specifieke situatie wél, uiteraard met oog voor de veiligheid, de streep te passeren.

Scherper beeld tracking en tracing

Met dit uitgebreide, integrale raamwerk kunnen we ons beeld van *tracking* en *tracing* scherper krijgen. Als we ervoor zorgen dat menselijke principes landen in de voertuig- en infralaag, zullen de besturingssystemen op voertuig- en infraniveau de mens en z'n principes vanzelf volgen (*tracken*) – zie de groene lijnen links in Figuur 3. De menselijke 'besturingssystemen' zullen er daarbij op toe moeten zien dat de principes juist geland zijn en ook operationeel gevolgd worden – de rode lijnen rechts.

3 ISAD staat voor *Infrastructure Support for Automated Driving*. Er worden verschillende niveaus onderscheiden, die aangeven in hoeverre de weg klaar is voor rijtaakondersteuning en automatische voertuigen.



Figuur 3 *De groene lijnen geven 'tracking' aan: voertuig- en wegkantsystemen volgen menselijke principes. Dat het blok bij Infra stippellijnen heeft, geeft aan dat de relaties vanuit die laag vaak wat minder direct en sterk zijn dan vanuit het voertuig. De rode lijnen rechts staan voor 'tracing': menselijke actoren die toezien op het sterk geautomatiseerde mobiliteitssysteem.*

Procesdiagram Meaningful human control

Met het integrale raamwerk voor *Meaningful human control,* Figuur 2, hebben we al een goed beeld van hoe menselijke principes (zouden moeten) landen in het voertuig. Maar het is wel een statisch beeld van een dynamisch systeem. Om nog wat scherper te krijgen wie voor wat verantwoordelijk is – en daarmee: hoe het concept van *Meaningful human control* versterkt kan worden – hebben we het raamwerk uitgewerkt als procesdiagram. Zie Figuur 4. Dit diagram geeft duidelijker aan wat de plek en rol is van bijvoorbeeld maatschappelijke organisaties en overheden (regelgeving) en hoe zij invloed kunnen uitoefenen om *Meaningful human control* te vergroten.

De kern van het diagram is een automatisch voertuig, met twee 'besturingssystemen', namelijk de ADCS en/of een menselijke bestuurder. Beide zijn in principe lerend – zie de blauwe pijlen. Bij menselijke bestuurders kan *ervaring* de prestaties verbeteren, terwijl een ADCS dankzij (zelflerende) kunstmatige intelligentie beter wordt op basis van hetzij eigen ervaring hetzij de ervaring van andere voertuigen, via draadloze communicatie of updates. Omdat deze vorm van leren intern gebeurt, noemen we het proximaal.

Het automatische voertuig, inclusief de ADCS, wordt ontworpen en (wat de software betreft) onderhouden door wat we in de figuur *voertuigontwerpers* noemen. Daar scharen we alle partijen onder die betrokken zijn bij het ontwerp van het voertuig en onderdelen ervan.

Geredeneerd vanuit het concept *Meaningful human control* moeten deze ontwerpers ervoor zorgen dat de menselijke principes en maatstaven op de juiste wijze in de ADCS landen. In veel gevallen is het verstandig om hierbij een extra (veiligheids)buffer aan te houden. Die buffer kan van alles zijn, zoals het vragen van een toestemming aan de menselijke bestuurder/inzittende: 'Weet u zeker dat …?'.

Welke principes de voertuigontwerpers in de ADCS opnemen, bepalen ze uiteraard niet alleen. De principes zullen in belangrijke mate reflecteren wat de *maatschappij* vindt. Daar worden normen en waarden ontwikkeld, maar ook geldt dat de maatschappij staat voor hun potentiële kopers/gebruikers – en die moet het naar de zin worden gemaakt. Principes kunnen ook rechtstreeks van *overheden* komen, via bijvoorbeeld beleid of regelgeving. En dan zijn er nog de mogelijke aanscherpingen van principes als gevolg van de interactie met andere voertuigen (weggebruikers) en de infrastructuur. We noemden net al het zelflerend vermogen, de blauwe pijlen, die tot optimalisaties leiden. Maar (bijna-) ongevallen en andere incidenten kunnen tot verdergaande aanpassingen leiden met bijvoorbeeld software-updates en wijzigingen in het ontwerp. Dat kan zijn omdat de voertuigontwerpers zelf daartoe besluiten (pijl vanuit *Interacties* over links) of omdat het incident tot ophef heeft geleid en er druk vanuit maatschappij en overheid is ontstaan (pijl over rechts). Deze externe interventies beschrijven we als *distaal*, ofwel 'van buitenaf'.

Overigens kan een bestuurder ook een ‘update’ ontvangen, bijvoorbeeld door een (al dan niet verplichte) bijscholing over hoe hij zijn rijgedrag of rol als inzittende van een automatisch voertuig kan verbeteren. Dit is ook een distale aanpassing.

Werken met Meaningful human control

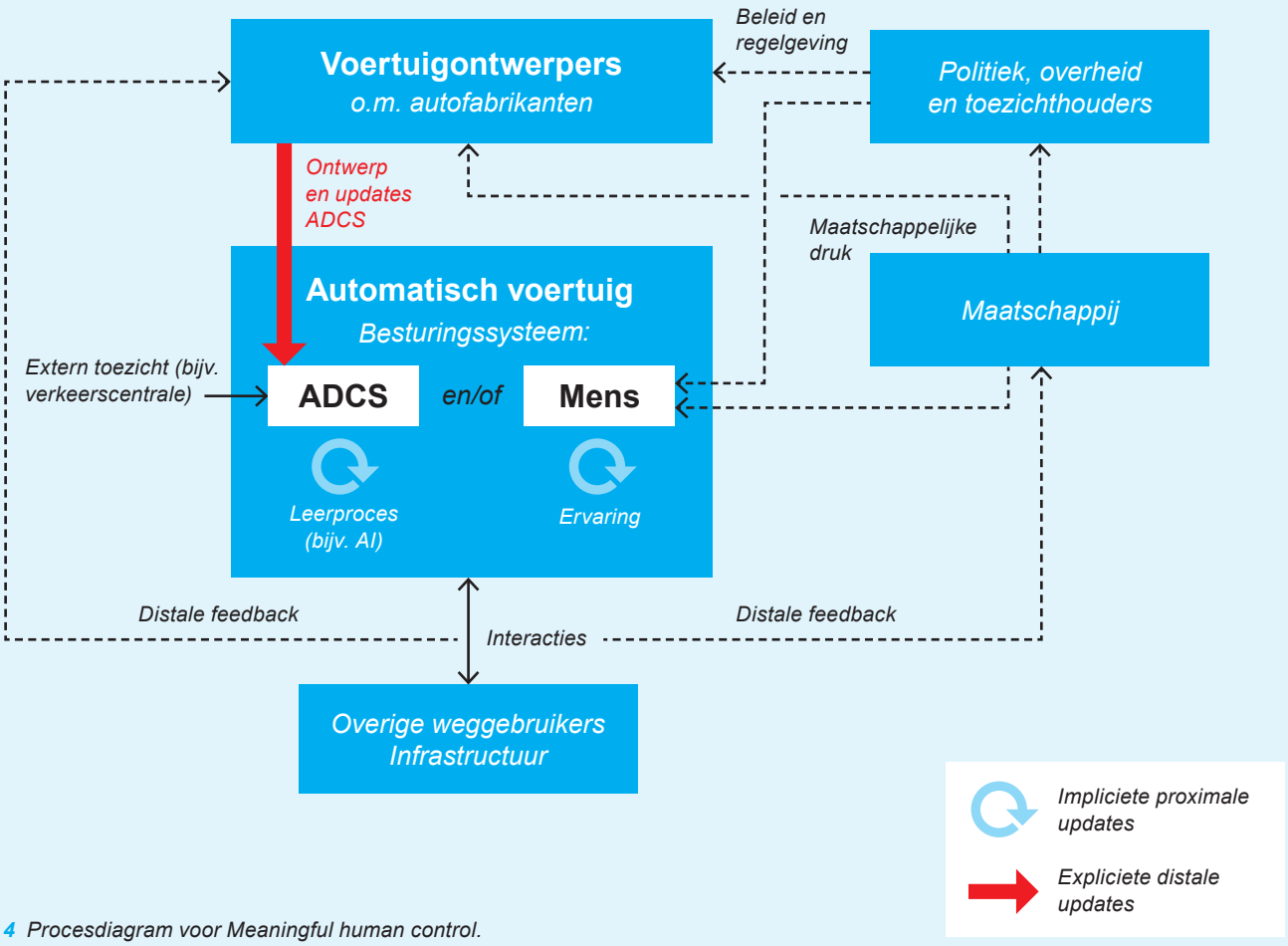
Tot zover de beschrijving van onze twee instrumenten voor *Meaningful human control*: het integrale raamwerk, Figuur 2, en het procesdiagram, Figuur 4. Zoals gezegd is het doel ervan om autofabrikanten, wegbeheerders en beleidsmakers te helpen om te komen tot een veilige, beheersbare en verantwoorde inzet van automatische voertuigen. Maar met het beschrijven en het hebben van een instrument zijn we er niet – en daarom willen we tot slot kort stilstaan bij het *toepassen* van (de instrumenten voor) *Meaningful human control*. Dat is geen onderdeel geweest van ons onderzoek, dus we schetsen slechts een mogelijke implementatie van het beschreven concept.

Een invulexercitie voor Meaningful human control

Als we de huidige, staande praktijk bekijken, dan zien we dat veel acties en inspanningen van autofabrikanten, overheden en andere stakeholders in principe al voldoen aan de beginselen van *Meaningful human control*. Maar ‘veel’ is nog niet ‘alle’, wat ook duidelijk naar voren komt in verschillende misstanden en incidenten die al gebeurd zijn. Wil *Meaningful human control* echt z’n stempel kunnen drukken, zonder hiaten en witte vlekken, dan moet het gedachtengoed diep in de strategieën en processen komen.

Een nationale of beter nog Europese overheid zou hierbij het initiatief kunnen nemen, in nauwe samenspraak met autofabrikanten en, wellicht als klankbordgroep, de meer maatschappelijke organisaties. Een eerste stap kan zijn om de bovenste, menselijke laag van Figuur 2 in te vullen, met als tijdshorizon het ‘nu’ of juist de situatie over één, twee, vijf of meer jaar. Welke principes gelden er (nu of dan) op het vlak van verkeer en vervoer in het algemeen en automatische voertuigen in het bijzonder? Welke zijn expliciet en beschreven in bijvoorbeeld visie- en beleidsdocumenten? Welke (nog) niet? Hoe verhouden ze zich tot elkaar, welke zijn algemeen en welke afgeleid? En wie zijn op de menselijke laag de ‘besturingselementen’? Over welke partijen en organisaties hebben we het? Wie precies zijn die voertuigontwerpers, overheden en maatschappelijke organisaties?

Eenzelfde invulexercitie kan daarna plaatsvinden voor de lagen Voertuig en Infra. Welke ODD’s en ISAD’s onderscheiden we, welke wetten en regels zijn van kracht? Wat zijn de besturingselementen op deze niveaus? Welke typen ADCS zijn er zoal? Welke C-ITS-diensten? Enzovoort.



Figuur 4 *Procesdiagram voor Meaningful human control.*

Is alles eenmaal ingevuld dan kunnen de lijnen van *tracking* en *tracing* worden getrokken, aan de hand van Figuur 3. Landen alle menselijke principes in de lagen Voertuig en Infra? Welke (nog) niet expliciet? Wie heeft welke rol als het om toezicht gaat? Is dat (menselijke) toezicht operationeel of meer op afstand? En hoe zit het met de verantwoordelijkheid? Aan de hand van Figuur 4 kunnen die rollen en verantwoordelijkheden nog wat scherper worden ingevuld.

Hiaten blootleggen en aanpakken

Het samen verkennen van het speelveld legt mogelijk hiaten bloot, zoals menselijke principes die nog te weinig expliciet zijn of die wel expliciet zijn maar nog niet zijn geland in wetgeving of ODD, of misschien wel in wetgeving maar niet in een besturingssysteem. Om hierbij nog een stukje dieper te graven, kunnen de partijen aan de hand van het ingevulde raamwerk en procesdiagram ook *wat als*-scenario's doorlopen. Als dit gebeurt of dat misgaat, wie is dan waar (moreel) verantwoordelijk voor? Zijn er mogelijkheden om de risico's op het incident te verkleinen?

Het zal niet eenvoudig of zelfs onmogelijk zijn om op deze theoretische wijze alle hiaten meteen op te sporen en op te lossen. Maar het gestructureerd verkennen aan de hand van een denkraam, is wel een eerste essentiële stap. Na zo'n gezamenlijke verkenning ontstaat er wellicht ook voldoende draagvlak voor regelgeving over *Meaningful human control* in de werkprocessen. Zo zouden autofabrikanten, maar bijvoorbeeld ook wegbeheerders, verplicht kunnen worden om *Meaningful human control* op te nemen in hun ontwerp- en beheerprocessen en om daarvan verslag te doen in bijvoorbeeld een actieplan.

Mocht zich toch een incident voordoen, dan kan aan de hand van de gezamenlijke verkenning en de eventuele actieplannen teruggeredeneerd worden wat er waar is misgegaan en hoe *Meaningful human control* verder in de processen verwerkt moet worden.

Tot slot

In deze bijdrage hebben we besproken hoe we automatische voertuigen voldoende veilig maar ook beheersbaar en verantwoord kunnen houden. Een focus op alleen techniek en innovatie is dan niet voldoende: het draait er vooral om dat de mens de machine de baas blijft. Om aan dat doel te werken is het concept *Meaningful human control* interessant. De insteek is dat geautomatiseerde systemen zo opgezet en ingericht moeten zijn dat altijd mensen, en niet computers en hun algoritmen, de controle houden over de beslissingen. Zo blijven ook altijd mensen moreel verantwoordelijk voor het handelen van de systemen.

In dit onderzoek hebben we het concept uitgewerkt in respectievelijk een integraal raamwerk en een procesdiagram voor *Meaningful human control*.

Deze twee instrumenten helpen om de lijnen van het *tracking* en *tracing* in beeld te brengen: hoe landen menselijke principes in de voertuig- en infrasystemen en hoe is het toezicht geregeld? Overheden, auto-fabrikanten en andere stakeholders kunnen met die instrumenten aan de slag om hun processen, rollen en verantwoordelijkheden inzichtelijk te maken en waar nodig aan te scherpen.

Door op deze wijze de menselijke factor gestructureerd mee te nemen in het vormgeven van een mobiliteits-systeem met automatische voertuigen, kunnen we ongevallen en/of ongewenste effecten zoveel mogelijk voor zijn. De mens zal zo de steeds slimmere maar soms ook onvoorspelbare voertuigen voldoende de baas blijven.

Over de auteurs

Dr. ir. Simeon C. Calvert is universitair docent op de afdeling Transport & Planning van TU Delft. Ook is hij codirecteur van het Delft Data Analytics and Traffic Simulation Lab (DiTTlab) en van het CiTy-AI-lab. Zijn onderzoek richt zich op de invloed van technologie op het wegverkeer. E-mail: **s.c.calvert@tudelft.nl**

Dr. Stig O. Johnsen is senior onderzoeker bij SINTEF (Safety and Reliability Group) en docent bij NTNU en NORD. Hij is verantwoordelijk voor het Human Factors Network (HFC) in Noorwegen en is voorzitter van Onderzoek naar ongevallen en incidenten bij ESRA. Zijn onderzoek richt zich op *Meaningful human control* van autonomie in het vervoer en de olie- en gasindustrie.

Ashwin George MSc. is promovendus bij de Human-Robot Interaction Group van de afdeling Cognitive Robotics van de TU Delft en maakt deel uit van het HERALD Lab. Hij onderzoekt hoe de introductie van technologie in het verkeer kan leiden tot gedragsaanpassingen en ethische uitdagingen.