De ethiek van zelfrijdende auto's

Sven Nyholm

Inleiding

Toen filosofen rond 2014 begonnen met het onderzoeken van ethische kwesties met betrekking tot zelfrijdende auto's, ging de discussie over verschillende gedachte-experimenten: wie moet er verantwoordelijk worden gehouden als een zelfrijdende auto een ongeval zou veroorzaken?; hoe zou een zelfrijdende auto moeten reageren als een ongeval niet te voorkomen is? Juridische wetenschappers hadden al iets eerder dan filosofen over deze vragen, maar ook deze artikelen in de rechtstheorie gingen meestal over hypothetische scenario's.

In 2015 begonnen echter wat eerder hypothetische scenario's waren in het echte leven te gebeuren. Dat jaar waren er zo'n 20 kleine ongelukken met experimentele zelfrijdende auto's zonder persoonlijk letsel. Het ging met name om kleine blikschade. Deze ongelukken werden doorgaans veroorzaakt door menselijke bestuurders – in 'conventionele' auto's – die tegen de experimentele zelfrijdende auto's reden. Dit gebeurde omdat deze zelfrijdende auto's zich niet gedroegen zoals de menselijke bestuurders dat verwachtten. De zelfrijdende auto's accelereerden bijvoorbeeld langzamer dan de meeste door mensen aangedreven auto's.

In 2016 gebeurde het echter voor de eerste keer dat een zelfrijdende auto een ongeval veroorzaakte. Op Valentijnsdag van dat jaar botste een experimentele zelfrijdende auto van Google lichtjes tegen een bus, en Google moest toegeven dat hun auto het ongeluk had veroorzaakt. Later dat jaar kwam de eerste persoon om het leven tijdens het rijden in een zelfrijdende auto. De man had de 'autopilot'-functie ingeschakeld van de Tesla en de auto botste frontaal tegen een vrachtwagen. In 2018 gebeurde het voor het eerst dat een mens fataal werd aangereden door een zelfrijdende auto. De kunstmatige intelligentie in een experimentele zelfrijdende auto van Uber herkende niet op tijd dat een mens de weg overstak. De vrouw werd eerst geclassificeerd als een teken, vervolgens als een fiets en uiteindelijk als een mens, maar toen was het al te laat. De auto raakte de vrouw en ze stierf op weg naar het ziekenhuis.

Wat er de afgelopen vijf jaar ook is gebeurd, is dat er een enorme toename is geweest van filosofische artikelen over ethische kwesties met betrekking tot zelfrijdende auto's. Zo'n twee of drie jaar geleden was het nog mogelijk om een overzicht te geven van al deze artikelen. Nu is dat schier onmogelijk, omdat er inmiddels tal van filosofen hebben gepubliceerd over zelfrijdende auto's. Ik zal dus niet proberen om hieronder alles samen te vatten wat er in de filosofie met betrekking tot zelfrijdende auto's is besproken. Ik zal in plaats daarvan een aantal van de onderwerpen bespreken die ik het meest interessant vind en waaraan ik de afgelopen tijd zelf heb gewerkt (Nyholm, 2018).

Het trolley-probleem en wat als een ongeluk niet te voorkomen is

Het eerste dat moet worden opgemerkt, is dat zelfrijdende auto's de belofte inhouden dat ze veel veiliger zijn dan conventionele auto's en dus de verkeersveiligheid zullen bevorderen. Dit is de belangrijkste reden dat mensen enthousiast zijn over het vooruitzicht van zelfrijdende auto's. Toch zullen ze niet honderd procent veilig zijn. Hoewel verkeersongelukken in ruim negentig procent van alle gevallen worden veroorzaakt door menselijke fouten (onoplettendheid, net te laat remmen, onvoldoende afstand bewaren, te hard rijden, enzovoort), zullen ook de veiligste zelfrijdende auto's een ongeluk niet altijd kunnen voorkomen. We moeten dus nadenken over hoe we met scenario's waarbij verkeersongelukken niet zijn te voorkomen omgaan. Een mogelijkheid is dat mensen het altijd moeten overnemen in zo'n onvermijdelijk ongevalscenario. De gemiddelde reactietijd van de mens is echter te traag. Mensen zul-

Sven Nyholm is universitair docent aan het Ethisch Instituut van de Universiteit Utrecht. Hij schrijft over onderwerpen in de ethische theorie en toegepaste ethiek (met een focus op de ethiek van de technologie). Zijn meest recente boek is *Humans and Robots: Ethics, Agency, and Anthropomorphism* (Rowman & Littlefield Publishers, 2020).

len in veel gevallen niet op tijd kunnen reageren. Zelfrijdende auto's moeten dus worden geprogrammeerd om te reageren op dit soort scenario's.

Veronderstel het volgende ongevalscenario: een zelfrijdende auto met vijf inzittenden detecteert ineens een groot obstakel op de weg, omdat het bijvoorbeeld van een vrachtwagen afvalt. Tenzij de auto uitwijkt, zullen de vijf passagiers waarschijnlijk sterven. De enige manier om uit te wijken is via de stoep waar een voetganger loopt. De enige manier waarop de auto de vijf inzittenden kan redden, is door die ene voetganger op te offeren. Wat zou een zelfrijdende auto in dit scenario moeten beslissen? En stel nu hetzelfde scenario voor maar dan met slechts één inzittende in de zelfrijdende auto en vijf voetgangers op de stoep. Hoe moet de auto nu reageren? Moet hij de inzittende opofferen voor de vijf voetgangers?

Deze scenario's zijn zo ontworpen dat ze vergelijkbaar zijn met het zogenaamde 'trolley-probleem'. Het trolley-probleem is één van de bekendste gedachte-experimenten in de filosofie en geïntroduceerd in 1967 door de filosofe Philippa Foot. In dit gedachte-experiment rijdt een op hol geslagen tram over de spoorrails die recht op vijf mensen afgaat. U staat verderop naast een hendel. Als u aan deze hendel trekt, schakelt de tram over naar een zijspoor waar één persoon staat. Dus om de eerdere vijf mensen te redden, zal er één persoon moeten worden opgeofferd. Er zijn veel alternatieve scenario's van dit experiment, en veel filosofen en psychologen hebben hierover geschreven.

In veel artikelen – zowel in de populaire media als in de academische literatuur – wordt de ethiek van de zelfrijdende auto vergeleken met het trolley-probleem. Ik denk echter dat veel van wat filosofen en psychologen hebben geschreven over het trolleyprobleem hier niet helpt voor een beter inzicht in de ethiek van de zelfrijdende auto. Daar zijn drie redenen voor.

Ten eerste wordt ons in academische discussies over het trolleyprobleem gevraagd om ons slechts te concentreren op een zeer beperkt aantal omstandigheden, terwijl ik de praktijk van de zelfrijdende auto's we met veel meer omstandigheden rekening moeten houden. Ten tweede worden in het trolleyprobleem vragen over morele en juridische verantwoordelijkheid terzijde geschoven. In de ethiek van zelfrijdende auto's kunnen we deze vragen niet negeren omdat ze juist tot de meest centrale vragen behoren. Ik kom hier later op terug. Ten derde gaan we er bij het trolley-probleem van uit dat we exact weten wat de uitkomsten zijn van de verschillende mogelijke acties. Het is in feite een volledig deterministisch scenario. In tegenstelling tot de praktijk van de zelfrijdende auto, waar de meeste beslissingen genomen moeten worden op basis van onzekere factoren en waar allerlei risico-inschattingen nodig zijn.

Om deze redenen is de literatuur over het trolleyprobleem minder nuttig dan veel mensen denken als het gaat om de ethiek van de zelfrijdende auto. Ik ga daarom nu een andere weg verkennen, de zogenaamde empirische ethiek.

Empirische ethiek

Wat is 'empirische ethiek'? Het is een poging om empirisch onderzoek naar de intuïtieve attitudes en oordelen van mensen te integreren in onze ethische analyse. We kunnen bijvoorbeeld de ethische intuïtie van mensen bestuderen door ze te laten oordelen over veel verschillende reële of gesimuleerde scenario's. We kunnen dan proberen patronen te onderscheiden in hun oordelen en intuïties, en deze bevindingen te verwerken in ethische argumenten.

Psychologen en gedragseconomen zijn reeds begonnen met het onderzoeken van de intuïtieve oordelen van mensen over hoe zelfrijdende auto's moeten omgaan met onvermijdelijke ongevalscenario's. Een interessante bevinding komt van onderzoekers van MIT (Massachusetts Institute of Technology) in de Verenigde Staten. De bevinding is dat wanneer mensen gevraagd worden hoe een zelfrijdende auto moet reageren als het een ongeluk niet kan voorkomen, de meeste mensen dan zeggen dat de auto zo geprogrammeerd moet worden dat de algehele schade minimaal is. Op de vraag wat voor soort algoritme ze zelf zouden willen hebben in hun eigen zelfrijdende auto, veranderen ze echter van mening. Mensen willen niet verplicht worden om auto's te kopen die altruïstisch zijn, die zich dus opofferen om zodoende te voldoen aan het minimaliseren van de algehele schade.

Op de webpagina www.moralmachine.net die ook door onderzoekers van het MIT is gemaakt, kan men over tal van verschillende dilemma's met betrekking tot zelfrijdende auto's een intuïtief oordeel vellen. De onderzoekers gebruiken deze oordelen voor het bepalen van een breed gedeelde mening over hoe zelfrijdende auto's geprogrammeerd zouden moeten worden. Is dit een goede basis voor een ethische analyse over hoe zelfrijdende auto's moeten reageren als ze een ongeluk niet kunnen voorkomen? Ik ben hier zeer sceptisch over om de volgende drie redenen. Ten eerste hebben mensen nog geen praktijkervaring met zelfrijdende auto's. Het is waarschijnlijk dat de houding van mensen zal verande-

ren zodra ze meer ervaring opdoen met de participatie van zelfrijdende auto's in het verkeer. Dit geeft ons een reden om niet te veel belang te hechten aan de huidige houding van mensen. Ten tweede zeggen de spontane reacties van mensen op hypothetische gevallen ons niet noodzakelijkerwijs iets over wat voor soort argumenten en redenen ze aanvoeren om hun intuïtieve oordelen te rechtvaardigen. Bij ethische theorievorming willen we reflecteren en argumenten evalueren, en niet alleen intuïtieve reacties zonder ondersteunende argumenten en redenen. Ten derde lijken mensen inconsistent te zijn in hun oordelen. Zoals ik eerder al opmerkte, willen mensen dat anderen een zelfrijdende auto aanschaffen die de schade beperkt tot het minimum, maar voor hen zelf willen ze een auto die zo geprogrammeerd is dat die de inzittenden redt. Zelfrijdende auto's die geprogrammeerd zijn om de algehele schade te minimaliseren zullen soms de inzittenden redden, maar soms moeten opofferen.

Met andere woorden, de ethische intuïtie van mensen is zeker interessant als we nadenken over de ethiek van zelfrijdende auto's. Het is echter niet duidelijk hoe deze intuïtieve oordelen, die soms wel breed gedragen worden, kunnen leiden tot solide conclusies over hoe een zelfrijdende auto het beste kan reageren wanneer de auto een ongeluk niet kan voorkomen. We lijken dus opnieuw tot de conclusie te komen dat we waarschijnlijk elders moeten gaan kijken om meer solide argumenten te vinden. Ik zal daarvoor de mogelijkheid onderzoeken of de traditionele ethische theorieën uit de morele filosofie hiervoor van betekenis kan zijn.

Ethische theorieën

Ik kijk in het bijzonder naar het utilisme, de deontologie en de deugdenethiek. Utilisme is de theorie dat als uitgangspunt heeft dat een handeling moet bijdragen aan 'het grootste geluk voor het grootste aantal mensen' oftewel het bevorderen van het algemeen welzijn. De deontologie is een ethische stroming die uitgaat van absolute gedragsregels, vaak gesteld als normen. Deugdenethiek stelt dat we het goede leven bereiken door het ontwikkelen van deugden (goede karaktereigenschappen). We kunnen deze theorieën gebruiken om de vraag te onderzoeken hoe zelfrijdende auto's moeten reageren wanneer ze een ongeluk niet kunnen voorkomen. Misschien moeten we 'utilistische auto's' nemen, of 'deontologische auto's', of 'deugdenethische auto's'.

Sommige filosofen zullen zeggen dat we hier een keuze voor één ethische theorie moeten maken. Ik zou echter willen voorstellen dat we gebruik moeten maken van alle drie de ethische theorieën, omdat we van elk van deze drie theorieën iets kunnen leren. De les vanuit het utilisme (of gevolgenethiek in het algemeen) is dat we over hoe zelfrijdende auto's zouden moeten reageren in een onvermijdelijk ongevalscenario moeten nadenken met het oog op het grotere belang van de samenleving. We moeten dus goed nadenken over wat het algemene welzijn en andere belangrijke menselijke waarden bevordert. De les vanuit de deontologie is dat wat voor regels we ook stellen voor zelfrijdende auto's, die regels universeel geldig moeten zijn. Dat wil zeggen dat omwille van rechtvaardigheid en gelijkheid zelfrijdende auto's hun keuze bijvoorbeeld niet op basis van ras en leeftijd mogen baseren bij een onvermijdelijk ongevalscenario. Het toegepaste algoritme moet dus iedereen gelijk behandelen. Ten slotte de deugdenethiek. Veel mensen in het verkeer laten zien dat ze belangrijke deugden bezitten die behoren bij een goede bestuurder. Zo gedragen de meeste mensen zich voorzichtig en redelijk verantwoord in het verkeer. Er zijn natuurlijk uitzonderingen, maar de meeste mensen voelen zich verantwoordelijk als ze omgaan met risicovolle technologieën, zoals auto's. Nu beargumenteert filosoof Mark Coeckelbergh (2016) dat hoe voorzichtig mensen zijn bij het gebruik van hun auto en hoe verantwoordelijk ze zich gedragen, beide afhangen van het ontwerp van de auto die ze gebruiken. De les die ik hier vanuit de deugdenethiek wil trekken, is dat zelfrijdende auto's zo moeten worden ontworpen dat mensen zich nog steeds verantwoordelijk gedragen voor wat er voor, tijdens en na ongelukken gebeurt. Dit is een belangrijke deugd.

Ter afsluiting van dit deel: we hebben nog geen goede ethische theorie over hoe zelfrijdende auto's met ongevalscenario's moeten omgaan. Maar we hebben wel veel materiaal voorhanden om dit verder te onderzoeken. Er zijn belangrijke lessen te trekken uit alle belangrijke ethische theorieën. Er zijn interessante resultaten vanuit de empirische ethiek waar we ook rekening mee moeten houden, en de discussie over het trolleyprobleem kan ons ook enkele interessante inzichten opleveren. Maar er is meer werk aan de winkel, zoals het verantwoordelijkheidsvraagstuk.

Verantwoordelijkheid

Laten we nu nogmaals kijken naar bovenstaande inleiding, waar ik behandelde hoe de ethiek rondom de zelfrijdende auto veranderde van een theoretisch gedachte-experiment naar een vraag uit de dagelijkse praktijk bij

echte ongevallen. Toen ik over die ontwikkeling schreef, benoemde ik het vraagstuk van wie er verantwoordelijk kan worden gehouden bij een ongeval met een zelfrijdende auto. Dit was steeds de centrale vraag in de dagelijkse praktijk van zelfrijdende auto's. Dus ik wil deze vraag over verantwoordelijkheid behandelen zowel in theoretisch filosofisch opzicht als ook in de dagelijkse praktijk.

Om te beginnen met casuïstiek uit de dagelijkse praktijk. Google heeft doorgaans in situaties van een ongeval met hun experimentele zelfrijdende auto's alle verantwoordelijkheid van de hand gewezen. Hoewel, zoals ik hierboven beschreef, in één specifieke situatie – de casus op Valentijnsdag 2016 – moest Google toegeven dat hun auto een aanrijding had veroorzaakt. De Google-auto was tegen een bus aan gebotst.

Google gaf toe "gedeeltelijk verantwoordelijk" te zijn en beloofde een update van de software van hun auto's, zodat de auto het gedrag van bussen in de toekomst beter zou kunnen voorspellen. Toen later dat jaar een man stierf in een Tesla-auto in de 'autopilot'-modus, ontkende Tesla daarentegen alle verantwoordelijkheid. Tesla publiceerde een blogpost waarin ze haar medeleven betuigde met de familie van de overleden man. Maar het bedrijf wees op de gebruikersovereenkomst waarin staat dat degene die gebruik maakt van de autopilot-functie verantwoordelijk is voor de eventuele gevolgen daarvan. Tegelijkertijd heeft Tesla aangekondigd om de software in haar auto's te updaten zodat die beter in staat zijn om obstakels te detecteren die mogelijk een botsing kunnen veroorzaken.

Voor sommigen was bovengenoemde reactie van Google verstandiger dan die van Tesla. Wanneer Tesla toegaf dat het een goed idee is om de software een update te geven, was dit dan niet hetzelfde als een schuldbekentenis voor de botsing? Als Tesla niet

verantwoordelijk was voor de crash, welke reden was er dan voor hen om de technologie in hun auto te updaten?

Sommige critici stellen dat het oneerlijk lijkt om de gebruikers van zelfrijdende auto's de schuld te geven van ongevallen die veroorzaakt zijn door hun auto, zelfs als die ongelukken schade aan andere mensen met zich meebrengen. Gebruikers van zelfrijdende auto's die het geluk hebben dat hun auto geen ongeval veroorzaakt handelen immers hetzelfde als gebruikers die de pech hebben dat hun auto wel een ongeval veroorzaakt.

Waarom houden we niet gewoon de bedrijven die de zelfrijdende auto's produceren verantwoordelijk? Sommige deskundigen – zowel ethici als juristen – zijn bezorgd dat dit ten koste zal gaan van de motivatie van bedrijven om deze techniek te ontwikkelen. Dit wordt als een slechte ontwikkeling beschouwd aangezien er veel voordelen worden verwacht van zelfrijdende auto's, waaronder een positief effect op de verkeersveiligheid.

Een ander argument dat wordt aangedragen om ontwikkelaars en bedrijven van zelfrijdende auto's niet verantwoordelijk te maken voor eventuele ongevallen met zelfrijdende auto's, is dat zij niet kunnen voorspellen hoe de auto's zich zullen gedragen wanneer zij daadwerkelijk aan het verkeer gaan deelnemen. Er wordt gesteld dat, wanneer de auto's zich in het verkeer begeven en autonoom gaan functioneren, de makers geen controle meer hebben over hetgeen de auto doet. Het is immers een zelfrijdende auto, die gemaakt is om zelfstandig te opereren. Dus geen mens heeft de controle over wat de auto doet. Immers, geen mens is in staat om volledig te voorspellen wat de kunstmatige intelligentie in de auto besluit als beste respons op bepaalde verkeerssituaties.

Sommige ethici die zich met dit soort onderwerpen bezig houden vrezen voor een, wat zij noemen, een 'verantwoordelijkheidskloof'. Dit zou betekenen dat er niemand verantwoordelijk kan worden gehouden of schuldig bevonden kan worden in het geval van een aanrijding met letsel. Wanneer er geen menselijke bestuurder is en de auto is volledig zelfrijdend, hoe kan dan een mens schuldig of verantwoordelijk zijn voor de eventuele schade die de auto aanricht?

Zijn deze zorgen om de verantwoordelijkheidskloof reëel? Het zou kunnen dat de oude manier van denken over verantwoordelijkheid en schuld van een bestuurder bij een ongeval niet zomaar kan worden gekopieerd naar een situatie van een ongeval met een zelfrijdende auto. Het zou volgens juristen ook kunnen zijn dat de bestaande wettelijke kaders voor het toewijzen van de verantwoordelijkheid bij auto-ongelukken niet voldoen bij zelfrijdende auto's. Onze huidige verkeerswet gaat er immers van uit dat er altijd een mens achter het stuur zit.

Er zijn echter manieren om de bestaande morele en juridische verantwoordelijkheid toe te passen op het verantwoordelijkheidsvraagstuk rondom de zelfrijdende auto. Men kan bijvoorbeeld kijken naar wie het meeste voordeel haalt uit de aanwezigheid van zelfrijdende auto's in het verkeer. Wanneer een vervoersbedrijf bijvoorbeeld zelfrijdende auto's verhuurt en dat bedrijf verdient daar veel geld mee, dan zou het gerechtvaardigd zijn om het vervoersbedrijf verantwoordelijk te maken voor eventuele ongevallen die deze lucratieve zelfrijdende auto's veroorzaken.

Bovendien, zelfs als mensen geen directe controle hebben over het gedrag van de zelfrijdende auto (want

deze rijdt zelf), dan hebben ze nog wel indirecte controle over het gedrag van de zelfrijdende auto. Zelfrijdende auto's zullen moeten worden geüpdatet en onderhouden en de frequentie en wijze van dit noodzakelijke onderhoud zal worden bepaald door de gebruikers. Dit kan helpen om tenminste een gedeeltelijke verantwoordelijkheid voor het gedrag van zelfrijdende auto's in het verkeer toe te kennen aan iedereen die verantwoordelijk is voor dit onderhoud. Dit geeft de gebruiker indirecte controle over het gedrag van de auto in het verkeer. Dat zou genoeg kunnen zijn om hen verantwoordelijk te maken hiervoor.

Hoe we ook besluiten om de verantwoordelijkheid te organiseren wanneer zelfrijdende auto's schade veroorzaken, we zullen ons moeten realiseren dat we onze kijk op verkeersveiligheid en verantwoordelijkheid in het verkeer moeten aanpassen. Onze kijk op auto's en hun bestuurders zullen we moeten aanpassen op de nieuwe situatie waarin auto's zelfstandig deelnemen aan het verkeer en waarbij de mensen in de auto slechts passagiers zijn. Er is nog veel onontgonnen terrein op dit gebied voor ethici die geïnteresseerd zijn in de wereld van de toekomst waarin autonome technologie steeds vaker taken gaat overnemen die voorheen door mensen werden uitgevoerd, zoals rondrijden in een auto.

Literatuur

Marc Coeckelberg, Responsibility and the moral phenomenology of using self-driving cars, in: *Applied Artificial Intelligence*, 2016, 30(8).

Sven Nyholm, The ethics of crashes with self-driving cars, a roadmap (I en II), in: *Philosophy Compass*, 2018, 13(7).

MIT, *Moral Machine*, http://moralmachine.mit.edu/.

Waymo, de zelfrijdende auto van Google zonder stuur en zonder pedalen
Bron: Wikimedia Commons, CC BY-SA 4.0