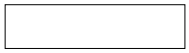
Een particularistische zelfrijdende auto

Max Velner

3983110

Begeleiding: Sander Werkhoven



Inhoudsopgave:

Samenvatting

In dit paper verdedig ik een manier om zelfrijdende auto's te programmeren met een particularistische theorie. De theorie die ik hierbij gebruik is die van Ross, zoals beschreven in het artikel *What Makes Right Acts Right?* Het paper is opgedeeld in vier hoofdstukken. In de eerste twee hoofdstukken zal ik laten zien wat de problemen zijn van het programmeren van deontologische en consequentialistische theorieën in een zelfrijdende auto en wat de voordelen zijn van het particularisme ten opzichte van deze theorieën. In het derde en vierde hoofdstuk zal ik een manier voorstellen en verdedigen om het particularisme van Ross zo goed mogelijk te implementeren in zelfrijdende auto's. Deze manier maakt gebruik van een artificiële intelligentie die ethisch handelen van mensen kan immiteren.

Inleiding

Zelfrijdende auto's zijn geen sciencefiction meer. Google is al meer dan zes jaar bezig met het testen van zelfrijdende auto's en de techniek is al ver genoeg gevorderd dat een zelfrijdende auto kan deelnemen aan het verkeer. Er wordt verwacht dat wanneer iedereen in zelfrijdende auto's rijdt dit het aantal slachtoffers in het verkeer drastisch zal verminderen.[1] Toch zullen zelfs de beste zelfrijdende auto's niet in staat zijn om ongelukssituaties te voorkomen.[2] Ongelukssituaties met zelfrijdende auto's brengen grote ethische problemen met zich mee. Om te zien wat voor problemen dit zijn zal ik hieronder een casus schetsen.

Casus 1:

Stel een zelfrijdende auto met één inzittende rijdt op een snelweg. Het is druk op de snelweg en de zelfrijdende auto is aan de voor-, rechter- en linkerkant ingesloten door andere voertuigen. Het voertuig dat voor de zelfrijdende auto rijdt is een vrachtwagen. Plotseling valt er een groot object van de vrachtwagen die voor de zelfrijdende auto rijdt. Het is te laat om te remmen om een botsing met het object te voorkomen. Links van de auto rijdt een personenauto met één bestuurder en aan de rechterkant rijdt een persoon op een motorfiets. De zelfrijdende auto is gedwongen een keuze te maken tussen: A, Het aan aanrijden van de motorfiets. Dit betekent dat de passagiers in de zelfrijdende auto het hoogstwaarschijnlijk met weinig verwondingen overleven, maar de motorrijder zal hoogstwaarschijnlijk sterven. B, de zelfrijdende auto kiest voor de botsing met het object dat van de vrachtwagen gevallen is. De passagiers van de zelfrijdende auto zullen dit hoogstwaarschijnlijk niet overleven, maar de kans dat andere voertuigen op de snelweg in het ongeluk betrokken raken is minimaal. En C, de zelfrijdende auto kan naar links uitwijken en tegen de personenauto botsen. In dit geval is de kans dat de passagiers van de zelfrijdende auto het overleven kleiner dan wanneer ze de motorfiets aanrijden, maar groter dan wanneer ze tegen het object van de vrachtwagen botsen. Hiermee worden echter wel de passagiers van de personenauto ook in levensgevaar gebracht, ook al is de kans dat ze het overleven een veel groter dan dat de persoon op de motorfiets het ongeluk overleeft. Wat moet de zelfrijdende auto doen? Kies je voor het beschermen van de passagier en offer je iemand anders op? Of juist voor het beschermen van andere weggebruikers door de passagier op te offeren? Of is de middenweg, die wel de meeste risico's met zich meebrengt het beste?[3]

---

1 'Google Self-Driving Car Project' (Laatst geraadpleegd 21 juni 2016) https://www.google.com/selfdrivingcar/.

2 Patrick Lin,'Why Ethics Matters for Autonomous Cars', in: Markus Maurer, J. Christian Gerdes, Barbara Lenz, Hermann Winner (red.), *Autonomes Fahren* (Berlin 2015) 71-72.

3 Lin, 'Why Ethics Matters for Autonomous Cars' 70-71.

Het verschil tussen een zelfrijdende auto en een menselijke bestuurder in deze situatie is dat een computerprogramma veel minder tijd nodig heeft om een afweging te maken wat het moet doen in een situatie als hierboven beschreven. Een mens kan niet op tijd een afweging maken over wat het beste is om te doen in een situatie als deze. Een mens reageert zonder na te denken op de situatie. Van mensen die accuut moeten reageren op onverwachte situaties kunnen we geen ethische overwegingen verwachten. Een computer kan daarentegen altijd een gefundeerde beslissing nemen in ongelukssituaties. De vraag is dus: Hoe programmeren we de zelfrijdende auto om de beste ethische beslissing te nemen?

Het antwoord op deze vraag is voor de hand liggend: Door de zelfrijdende auto te programmeren met een normatieve theorie. Dit zorgt ervoor dat de beslissingen die de zelfrijdende auto neemt ethisch te verantwoorden zijn. Dit roept wel onmiddelijk de volgende vraag op: Welke normatieve theorie? Op dit punt zou je kunnen denken dat de discussie dezelfde discussie wordt als in de normatieve ethiek. De vraag is dan: Welke normatieve theorie is de beste theorie? Toch is er een verschil tussen de discussie in de normatieve ethiek en de discussie over welke theorie het beste werkt bij zelfrijdende auto's. Naast de vraag welke ethische theorie het beste is, speelt de vraag - Welke theorie kan het beste in een zelfrijdende auto geïmplementeerd worden? - een even belangrijke rol. Het verschil tussen deze vragen is dat je bij de tweede vraag gebonden bent aan de praktische beperkingen van de techniek. We zijn dus op zoek naar een normatieve theorie die zich staande kan houden in het normatief-ethische debat en goed implementeerbaar is in een zelfrijdende auto. In de huidige literatuur, die nog niet erg veel is, over de ethiek van zelfrijdende auto's zijn consequentialistische theorieën het meest populair. Ik denk echter dat er nog een andere mogelijkheid is: Het particularisme van Ross.[4] Het doel van dit paper is om te laten zien dat het particularisme van Ross een plausibel alternatief is om zelfrijdende auto's mee te programmeren. Ik zal een voorstel doen hoe ik denk dat deze theorie in een zelfrijdende auto zo goed mogelijk geïmplementeerd kan worden.

Dit paper is opgedeeld in vier hoofdstukken. In het eerste hoofdstuk zal ik een aantal problemen bespreken van het implementeren van deontologische en consequentialistische theorieën in een zelfrijdende auto. In het tweede hoofdstuk zal ik kort laten zien welke voordelen het particularisme heeft ten opzichte van consequentialistische en deontologische theorieën. In hoofdstuk drie laat ik zien hoe het particularisme van Ross zo goed mogelijk geprogrammeerd kan worden in een zelfrijdende auto door middel van artificiële intelligentie. In het vierde en laatste hoofdstuk bespreek

4 Met particularisme van Ross bedoel ik zijn theorie in het artikel *What Makes Right Acts Right?*

ik hoe we moeten bepalen met welke gegevens de artificiële intelligentie geprogrammeerd moet worden.

Problemen met deontologische en consequentialistische theorieën

Met deontologische normatieve theorieën bedoel ik theorieën die vanuit principes bepalen wat goed en slecht is los van de consequenties die uit het handelen naar de principes volgen.[5] Ik zal eerst de problemen van deontologie behandelen en vervolgens de problemen van consequentialisme.

Ten eerste geven deontologische principes die sturend moeten zijn voor ons handelen ons vaak geen duidelijke instructie over hoe er gehandeld moet worden in een specifieke situatie. Stel we nemen casus 1 als voorbeeld en we nemen een zelfrijdende auto die geprogrammeerd is met het principe 'geen schade aan mensen aanrichten.' In het geval van casus 1 is met dit principe onmogelijk te bepalen welke keuze de zelfrijdende auto moet maken. Ten eerste is het op deze manier al lastig om te kiezen tussen tegen de motorrijder en tegen het object aanrijden. Het wordt nog lastiger als ook de botsing met de personenauto meegenomen moet worden in de overweging. Het principe geeft geen duidelijkheid over hoe omgegaan moet worden met een situatie waar de risico's groter zijn maar de kans op succes ook groter is. Daarnaast werken deontologische principes niet aggregratief. Dat wil zeggen dat een deontoloog in dit geval geen onderscheid mag maken tussen het aanrijden van één persoon of twee personen. Het is beide even erg aangezien er niet aan het principe voldaan wordt.

Een principe zoals 'geen schade aan mensen aanrichten' biedt dus geen hulp met het kiezen van een handeling. Je zou kunnen zeggen dat in dit geval het niet uitmaakt wat de zelfrijdende auto doet. Elke optie is even slecht dus kunnen we de beslissing net zo goed overlaten aan toeval. Ik denk dat deze optie niet wenselijk is. Ik zie niet hoe het toeval een goede verantwoording kan zijn voor ethische keuzes.

Een manier om het probleem te omzeilen is door gebruik te maken van meerdere principes en te veronderstellen dat een aantal principes samen wel tot een keuze voor een bepaalde handeling kunnen komen. Ook dit is geen goede oplossing. Door gebruik te maken van meerdere principes kan er alsnog besluiteloosheid optreden wanneer de principes elkaar tegenspreken. Stel we nemen als tweede principe 'dodelijke ongelukken voorkomen.' In dit geval is het duidelijk dat de zelfrijdende auto de personenauto aanrijdt aangezien dit de enige kans is om dit principe te vervullen. Maar stel we passen casus 1 aan door te zeggen dat er in plaats van één persoon er twee volwassenen met 3

5 Larry Alexander en Michael Moore, "Deontological Ethics", *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2015/entries/ethics-deontological/>.

kinderen in de personenauto zitten. Het verschil met casus 1 is dat het eerste principe bij elke optie gelijk was en daarom kon het tweede principe uitsluitsel geven. In de aangepaste casus is er sprake van twee principes die tegenstrijdig zijn met elkaar. De computer kan nog steeds geen keuze maken. Je zou kunnen veronderstellen dat er een bepaalde hiërarchie is tussen de principes, die de zelfrijdende auto in staat stelt om te beslissen. Maar hoe bepalen we deze hiërarchie? Uit het aangepaste voorbeeld van hierboven blijkt dat het intuïtief niet duidelijk is wat de hiërarchie is tussen verschillende principes, als die er al is.

Je zou ook kunnen zeggen dat de deontologische principes specifieker geformuleerd moeten worden zodat de computer altijd weet wat hij moet doen. Dit brengt ons bij het tweede probleem. Het is onmogelijk om deontologische principes op zo'n specifieke manier te formuleren dat ze goed geprogrammeerd kunnen worden en tegelijkertijd recht doen aan het algemenere principe waaruit ze zijn afgeleid. Om principes specifieker te maken moet je weten in welke mogelijke situaties de zelfrijdende auto terecht kan komen. Ik denk dat het onmogleijk is om vantevoren elke mogelijke situatie in te schatten en bijbehorende specifieke principes te formuleren.[6]

Zelfs als de bovenstaande argumenten niet problematisch zijn blijft er nog het probleem met welke principes de zelfrijdende auto geprogrammeerd zou moeten worden. Zolang er geen consensus is over welke principes de juiste zijn is het geen goed idee om de zelfrijdende auto met een set aan principes te programmeren.[7]

De consequentialistische manier om zelfrijdende auto's te programmeren werkt in eerste instantie beter dan de deontologische manier. Met consequentialisme bedoel ik de normatieve theorieën die bekijken of een handeling goed of slecht is aan de hand van de consequenties die de handeling met zich meebrengt.[8] Met betrekking tot zelfrijdende auto's zijn consequentialistische theorieën in het huidige debat populairder dan deontologische theorieën. Hibbard, Goodall en andere auteurs over ethiek van zelfrijdende auto's verdedigen een consequentialistische zelfrijdende auto.[9] [10] [11] De reden dat consequentialistische theorieën populairder zijn dan deontologische theorieën is dat het voor

6 N. J. Goodall, 'Ethical Decision Making During Automated Vehicle Crashes' in: *Transportation Research Record: Journal of the Transportation Research Board*, volume 2424 (2014) 62-63.

7 Goodall, 'Ethical Decision making,' 63.

8 Walter Sinnott-Armstrong, "Consequentialism", The Stanford Encyclopedia of Philosophy (Winter 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2015/entries/consequentialism/>.

9 Goodall, 'Ethical Decision' 58-65.

10 Bill Hibbard, *Ethical Artificial Intelligence* (Berkeley: Space Science and Engineering Center University of Wisconsin - Madison and Machine Intelligence Research Institute, 2015), 78-93, https://arxiv.org/ftp/arxiv/papers/1411/1411.1373.pdf (laatst geraadpleegd 21 juni 2016).

11 Lin, 'Why Ethics Matters for Autonomous Cars' 69-85.

sommige vormen van consequentialisme veel makkelijker is om een een zelfrijdende auto met een consequentialistische theorie te programmeren. Dit komt doordat deze consequentialistische theorieën uit te drukken zijn in het maximaliseren van een functie. Via een waardesysteem kunnen consequenties van acties gekwantificeerd worden zodat de computer kan uitrekenen wat de beste actie is.[12] Niet elke vorm van consequentialisme is echter geschikt om op deze manier in een zelfrijdende auto geïmplementeerd te worden. Voorkeursutilisme is bijvoorbeeld problematisch. De zelfrijdende auto kan natuurlijk niet zien welke voorkeuren andere weggebruikers op het moment van een ongelukssituatie hebben.

Hoewel een consequentialistische benadering de makkelijkste manier lijkt om een zelfrijdende auto te programmeren en uit onderzoek blijkt dat relatief veel mensen comfortabel zijn met een auto die geprogrammeerd is met een handelsutilistische ethiek, zijn er toch een aantal problemen met een consequentialistische zelfrijdende auto.[13]

Eén probleem is dat het niet duidelijk is hoe een consequentialistische auto om moet gaan met kansberekening. Stel we nemen casus 1 in combinatie met het handelsutilistische principe 'zo veel mogelijk welzijn voor zoveel mogelijk mensen.' In het geval van de zelfrijdende auto is welzijn gedefiniëerd in zo min mogelijk fysieke schade aan mensen. We nemen aan dat er één persoon in de zelfrijdende en één persoon in de personenauto rijdt. Stel dat de kansen om te overleven in casus 1 zijn als weergegeven wordt in de tabel:

Botsingen met:	Overlevingskans passagier zefrijdende auto	Overlevingskans aangereden persoon	Kans op geen dodelijke slachtoffers	Kans twee dodelijke slachtoffers	Gemiddeld aantal slachtoffers
Motorrijder	95%	5%	4,75%	4,75%	1
Personenauto	50%	50%	25%	25%	1
Object	0%	n. v. t.	0%	0%	1

In dit geval kan een handelsutilistische zelfrijdende auto geen keuze maken tussen de opties aangezien het gemiddelde aantal slachtoffers per optie overal even groot is. De opties leveren kanstechnisch gezien allemaal evenveel fysieke schade aan. Een utilistische calculus uitvoeren is dus niet altijd genoeg om de doorslag te geven in hoe een zelfrijdende auto moet handelen. Ook het

12 Hibbard, 'Ethical Artificial Intelligence' 9-21.

13 Jean-François Bonnefon, Azim Shariff, Iyad Rahwan, *'Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?'* (2015). http://arxiv.org/abs/1510.03346. 4-6.

regelutilisme kan niet ontsnappen aan deze objectie aangezien ook bij regelutilisme gebruik moet worden gemaakt van een utilistische calculus.

Een ander probleem van consequentialisme is dat het net als deontologie een vorm van principalisme is. Dit vormt een groot probleem bij het toepassen van de theorieën in de werkelijkheid. Het nadeel van principalisme is dat door het toepassen van absolute principes om te bepalen wat de juiste handeling is het voor kan komen dat handelingen, die zeer tegenintuïtief aanvoelen, als de juiste handeling gelden. In het volgende hoofdstuk zal ik dit probleem aan de hand van voorbeelden uitleggen en laten zien waarom het particularisme van Ross beter in staat is om recht te doen aan onze intuïties.

Voordelen particularisme van Ross

Belangrijk om op te merken is dat hoewel deontologische en consequentialistische theorieën weerlegd kunnen worden met intuïtieve tegenvoorbeelden dit niet betekent dat de theorieën in hun geheel niet bruikbaar zijn. Het betekent slechts dat in sommige gevallen de theorieën tekortschieten, maar in de meeste gevallen wel degelijk zeggen welke handeling het beste is in oveernstemming met onze intuïties. Het particularisme van Ross is in staat om de sterke kanten van deontologie en consequentialisme te combineren zonder dat het last heeft van tegenintuïtieve tegenvoorbeelden.[14] De reden dat de theorie van Ross dit kan is omdat het een anti-principalistische theorie is. Met anti-principalisme bedoel ik de visie dat ethische redenen contextafhankelijk zijn en niet in elke situatie even sterk gelden. Naast Ross staat ook Dancy kritisch tegenover het principalisme.[15] Dancy en Ross zetten zich hiermee af tegen consequentialistische en deontologische theorieën die beiden handelingsprincipes verdedigen die in elke ethische situatie zo goed mogelijk toegepast moeten worden. Een groot nadeel van principalisme is dat dit situaties kan opleveren waarin een handelingsprincipe ons tot een handeling verplicht die in strijd is met wat we intuïtief als moreel goed aanvoelen. Dit soort situaties komen in het dagelijks leven zo vaak voor dat het gemakkelijk is om een goed voorbeeld te geven zonder een extreem hypothetische situatie te schetsen.

Voorbeeld 1: Stel persoon A doet zijn vriendin de belofte dat hij boodschappen doet en het avondeten zal koken omdat hij vakantie heeft en de hele dag thuis is en zij van negen tot vijf moet werken. De vriendin zegt echter al een tijdje dat het haar leuk lijkt om weer eens uit eten te gaan en halverwege de dag besluit A om zijn vriendin te verrassen en reserveert hij bij een restaurant.

Ik denk dat het in dit geval intuïtief niet moeilijk is om toe te geven dat A hier een moreel juiste keuze heeft gemaakt door zijn belofte te breken. Een deontoloog zou misschien kunnen zeggen dat het niet per se slecht is om je niet houden aan een plicht als er een andere plicht is die sterker is. Ik denk dat deze verdediging niet toegepast kan worden op het voorbeeld hierboven. Welke plicht zou sterker moeten zijn dan de belofte? Is er een plicht om af en toe je vriendin te verrassen? Ik denk van niet. Als die er al is lijkt me die niet principieel sterker dan een belofte houden.

Om te laten zien waarom consequentialistische ethiek niet altijd met intuïties overeenkomt moeten we het voorbeeld iets aanpassen. De situatie blijft hetzelfde, alleen komt persoon A op weg naar de supermarkt een groep vrienden van vroeger tegen en wordt hij spontaan uitgenodigd om met hen

14 W.D. Ross, 'What Makes Right Acts Right?' in: Russ Shafer Landau (red.) '*Ethical Theory an Anthology*' (2013) 756-763.

15 Jonathan Dancy, 'An Unprincipled Morality' in: Russ Shafer Landau (red.) '*Ethical Theory an Anthology*' (2013) 772-776.

samen te eten bij één van de vrienden. In dit geval zeggen consequentialistische theorieën dat het goed is als persoon A met zijn vrienden uit eten gaat omdat zij waarschijnlijk meer geluk produceren of voorkeuren vervullen dan als persoon A zijn belofte houdt. Ook in dit geval lijkt het mij makkelijk om te zeggen dat mensen dit intuïtief niet als de juiste keuze zouden beschouwen.

Aanhangers van deontologische en consequentialistische theorieën zouden dit bezwaar aan de kant kunnen leggen door te zeggen dat onze intuïties er soms gewoon naast zitten. Hier heb ik twee reacties op. Ten eerste liggen aan de consequentialistische en deontologische theorieën aannames ten grondslag die op intuïties zijn gebaseerd. Simpel gezegd is de consequentialistische aanname dat geluk gelijk is aan goed een intuïtie en bij deontologie de aanname dat we een vrije wil hebben. Consequentialistische en deontologische theorieën kunnen intuïties dus niet zonder meer aan de kant vegen. Ze moeten een verhaal hebben waarom hun intuïties beter zijn dan andere intuïties. Ten tweede denk ik net als Dancy dat de bewijslast voor de stelling dat morele redenen wel absoluut zijn bij de consequentialisten en deontologen ligt.[16] Ik zie niet in waarom het vanzelfsprekend is dat morele redenen zich zo fundamenteel anders gedragen dan andere redenen. Zeker niet omdat ook nog eens blijkt dat onze intuïties niet aan de kant van de atomisten staan. Zolang consequentialisten en deontologen hier geen tevredenstellend antwoord op kunnen geven denk ik dat het particularisme een betere theorie over moreel handelen heeft.

Ik denk dat Dancy en Ross een beter verhaal hebben over hoe we ethisch moeten handelen in dit soort situaties. Dancy en Ross zeggen niet dat het toepassen van een absoluut principe ons tot de juiste ethische keuze lijdt, maar door de context juist te interpreteren weten we wat de goede ethische handeling is.[17] [18] Ross zegt dat er een veelheid aan ethische principes zijn, die weliswaar motiverend zijn en daarom bijdragen aan het maken van een morele afweging, maar ook contextafhankelijk zijn waardoor het verschilt per context in hoeverre de principes bijdragen. Volgens Ross wordt de juiste ethisch keuze gemaakt door het afwegen van verschillende ethische principes.[19] De goed ethische beslissing wordt gemaakt door een goed inzicht in de situatie. In voorbeeld 1 zou Ross kunnen zeggen dat er sprake is van twee ethische principes. Het deontologische principe om je belofte niet te breken en het handelsutilistische principe om voor zoveel mogelijk geluk te zorgen. Beiden principes kunnen terechte motivaties zijn voor een ethische beslissing, maar wat in deze situatie de doorslag geeft is niet dat één van de principes beter is, maar dat het inzicht van persoon A

16 Dancy, ‘Unprincipled Morality’ 772-773.

17 Ibidem. 772-774.

18 Ross, ‘Right Acts’ 756-763.

19 Ibidem. 756-763.

in de situatie hem in staat stelt om te bepalen dat het in dit geval beter is om het geluk te vergroten dan de belofte te houden.

Dancy gaat nog een stap verder door te zeggen dat ethische principes ook niet altijd bijdragen in het maken van een morele keuze. Volgens Dancy zou een ethisch principe in de ene situatie een bijdrage moeten leveren aan de motivatie om te handelen, terwijl in een andere situatie het houden aan hetzelfde ethische principe juist afbreuk zou moeten doen aan de motivatie om te handelen. Het enige wat bepaalt wat de juiste keuze is in een ethisch vraagstuk is inzicht in de situatie.[20] Dit is echter waar ik met Dancy breek. Ik denk net als Ross dat er wel degelijk morele principes zijn die altijd bijdragen aan de motivatie voor een ethische keuze. Ik denk dat het niet sterk is van Dancy om de bijdragende principes die Ross veronderstelt weg te laten omdat je dan geen verhaal meer hebt over wat ethisch goed handelen is behalve de situatie goed inschatten. Hoe weet je wanneer je de situatie goed inschat? Ross kan in antwoord op deze vraag nog verwijzen ethische principes die ten grondslag liggen aan de motivatie van een ethische keuze. Dit kan Dancy niet. Daarnaast kan Dancy geen recht doen aan het idee dat ethische casussen met elkaar te vergelijken zijn. Als elke morele reden context afhankelijk is kan er geen gebruik worden gemaakt van ervaring in eerdere ethische casussen om toe te passen op nieuwe casussen. Elke morele beslissing is zo contextafhankelijk dat het geen enkele indicatie kan geven hoe er juist moet worden gehandeld in een andere context. Bij Ross is het makkelijker te zien hoe ethische casussen met elkaar vergeleken kunnen worden. De bijdragende ethische principes van Ross kunnen vergeleken worden tussen meerdere casussen.

20 Dancy, ‘Unprincipled Morality’ 772-774.

Voorstel: Implementeer particularisme van Ross in zelfrijdende auto door middel van artificiële intelligentie.

De volgende vraag is: Hoe programmeren we het particularisme van Ross in een zelfrijdende auto? Mijn antwoord: Door gebruik te maken van een artificiële intelligentie. Om te begrijpen hoe dit kan zal ik in dit hoofdstuk eerst uitleggen hoe een artificiële intelligentie in een zelfrijdende auto kan werken.

Noah J. Goodall stelt in zijn stuk *Ethical Decision Making During Automated Vehicle Crashes* voor om in de situatie dat alle auto's op de weg vervangen zijn door zelfrijdende auto's gebruik te maken van artificiële intelligentie om ethische beslissingen te nemen. Goodall komt tot deze visie doordat hij van mening is dat er problemen zijn met het programmeren van bestaande ethische theorieën. Goodall wil uiteindelijk naar een situatie waarin iedere zelfrijdende auto is uitgerust met een artificiële intelligentie die ethisch handelen van mensen kan immiteren met als randvoorwaarde een aantal ethische waarden die niet geschonden mogen worden.[21] Ik ben het gedeeltelijk met Goodall eens dat dit de juiste manier is om zelfrijdende auto's te programmeren, maar voordat ik hier dieper op in ga wil ik eerst formuleren wat ik precies bedoel met een artificiële intelligentie die ethisch handelt.

Ten eerste is het belangrijk om te weten wat een artificiële intelligentie precies onderscheidt van een normaal computerprogramma. Over deze vraag is veel discussie omdat er verschillende definities van artificiële intelligentie in gebruik zijn. In dit paper wil ik me daarom beperken tot een specifiek soort artificiële intelligentie: *Machine Learning*.[22] Kort gezegd is het belangrijkste verschil tussen *Machine Learning* en een normaal computerprogramma het verschil tussen zelflerende algoritmes en geprogrammeerde algoritmes. Een normaal computerprogramma wordt van het begin tot het eind door de mens geprogrammeerd. Er wordt door de mens bepaald bij welke input welk algoritme gebruikt moet worden. Deze algoritmes worden door mensen geschreven en dus ook door mensen begrepen. De uitkomsten van een normaal computerprogramma zijn de uitkomsten die een programmeerder voor ogen had tijdens het programmeren van het programma. Bij *Machine Learning* werkt dit anders omdat de algoritmes die geprogrammeerd worden geen instructies zijn voor het genereren van output. De algoritmes die geprogrammeerd worden, zijn erop gericht om nieuwe algoritmes te creëren. Deze nieuwe algoritmes zijn uiteindelijk bepalend in welke handeling de artificiële intelligentie uitvoert. Een belangrijk verschil tussen een normaal computerprogramma en

21 Goodall, 'Ethical Decision making,' 58-65.

22 Monica Anderson, Chris Nicholson and Cristopher Schrader, *What are the main differences between artificial intelligence and machine learning*? (blog). 2015. https://www.quora.com/What-are-the-main-differences-between-artificial-intelligence-and-machine-learning. (laatst geraadpleegd 21 juni 2016).

de artificiële intelligentie is dat de algoritmes die uiteindelijk bepalend zijn voor de handeling van de artificiële intelligentie niet bepaald zijn door de programmeurs en mogelijk ook niet door mensen begrepen kunnen worden.

Ten tweede is het belangrijk om te weten hoe een artificiële intelligentie zichzelf nieuwe algoritmes aanleert. Kort gezegd is de manier waarop dit gaat patroonherkenning. Anders dan een normaal computerprogramma is een ethische artificiële intelligentie niet klaar wanneer mensen klaar zijn met de productie ervan. Een ethische artificiële intelligentie heeft input nodig om zijn eigen algoritmes te kunnen vormen. Deze input wordt geleverd door statistische gegevens. Op basis van de statistische gegevens maakt de artificiële intelligentie omgevingsmodellen. Met deze modellen kan de artificiële intelligentie de uitkomsten van zijn handelingen voorspellen. Deze omgevingsmodellen zijn niet statisch, maar kunnen veranderen met elke nieuwe waarneming die de artificiële intelligentie doet. Ook kunnen deze modellen zeer complex zijn, waardoor ze met grote zekerheid de uitkomsten van hun handelingen kunnen bepalen.[23]

Het idee van Hibbard is dat een artificiële intelligentie naast het voorspellen van uitkomsten van handelingen menselijke waarden kan aanleren volgens hetzelfde proces. Het is moeilijk om een machine die precies, getalmatig en consistent is de ambigue, subjectief, oneindige en inconsistente menselijke waarden aan te leren. Er zijn in het verleden pogingen gedaan om menselijke waarden uit te drukken in een verzameling regels. Deze pogingen zijn nooit erg succesvol geweest omdat mensen niet accuraat zijn in het rapporteren van hun eigen waarden. Hibbard stelt dat aan de hand van statistische gegevens menselijke waarden in kaart worden gebracht. Volgens Hibbard laten recente successen in het gebied van vertaalmachines, die door het gebruik van grote databases van menselijke taal en statistische analyse accurater zijn in hun vertalingen dan oudere machines, zien dat een complexe statistische analyse ook een goede weergave kan geven van menselijke waardes.[24] Ook Goodall neemt aan dat het mogelijk is om een artificiële intelligentie menselijke ethiek aan te leren door observatie van menselijk gedrag. Hij geeft in zijn artikel een voorbeeld van een artificiële intelligentie die het rijgedrag van een normaal mens kan imiteren na slechts 2 minuten observeren. Onderzoek moet uitwijzen of deze aanname klopt, maar voor dit paper ga ik er vanuit dat Hibbard en Goodall op dit gebied gelijk hebben.[25]

---

23 Bill Hibbard, *Ethical Artificial Intelligence*, (Berkeley 2015) PDF e-book. https://arxiv.org/ftp/arxiv/papers/1411/1411.1373.pdf. 78-93.

24 Hibbard, *Ethical Artificial Intelligence*, *78-79*.

25 Goodall, 'Ethical Decision making,'

Een belangrijk voordeel van de artificiële intelligentie ten opzichte van een normale computer is dat het in staat is om daadwerkelijk ethisch handelen van mensen te immiteren in plaats van het uitvoeren van vooraf geprogrammeerde ethische instructies. Goodall stelt dat dit voor bedrijven interessant is omdat de auto beter zal verkopen als de ethische beslissingen van de zelfrijdende auto voor de meerderheid intuïtief goed aanvoelen.[26] De redenen die Goodall voor een ethische artificiële intelligentie heeft, zijn dus niet per se filosofisch. Toch denk ik dat de ethische artificiële intelligentie die Goodall voorstelt filosofisch goed te verdedigen is. Hoewel Goodall het zelf niet expliciet maakt denk ik dat zijn idee over hoe een ethische artificiële intelligentie eruit moet zien goed aansluit bij het particularisme.

De ethische artificiële intelligentie neemt net als het particularisme de praktijk van ethisch handelen als basis. De artificiële intelligentie komt overeen met de ideeën van het particularisme omdat het niet met een theorie is geprogrammeerd over welke specifieke handeling in een bepaald geval het beste is, maar hoe tot een goede handeling gekomen wordt. Net als het particularisme heeft de artificiële intelligentie het voordeel dat elke specifieke situatie benaderd wordt als zodanig en er niet veralgemeniseerd wordt naar absolute regels die soms wel en soms niet overeenkomen met onze intuïties. De praktijk van het ethisch handelen bepaalt hoe de artificiële intelligentie ethisch handelt net zoals de praktijk van menselijk ethisch handelen het ethisch handelen van de mens bepaalt.

Een vraag die het gebruik maken van een artificiële intelligentie oproept is: Kan de ethische artificiële intelligentie op deze manier leren om beter ethisch te handelen dan mensen doen? Mijn antwoord op deze vraag is nee. Wanneer we de theorie van Ross aannemen is het onmogelijk dat een artificiële intelligentie die gebruik maakt van *machine learning* beter in staat is om ethisch te handelen dan mensen. Dit komt doordat er een essentiëel verschil is tussen hoe de artificiële intelligentie handelt en hoe mensen ethisch handelen. Zoals in het hoofdstuk hierboven is uitgelegd maakt de artificiële intelligentie gebruik van statistische gegevens om beslissingen te maken. Wat de artificiële intelligentie in wezen doet is dat hij gebruik maakt van de regelmaat in menselijk ethisch handelen om menselijk ethisch handelen te immiteren. Dit is fundamenteel anders dan hoe mensen ethisch handelen. Mensen maken in hun ethisch handelen niet uitsluitend gebruik van gegevens van eerdere ethische beslissingen. Elke ethische beslissing bestaat uit een afweging van ethische principes die voor die context relevant zijn. Het afwegen van de ethische beslissing is een proces dat niet weergegeven kan worden in de statistische gegevens. Alleen de uitkomst van het afwegen kan worden doorgegeven aan de artificiële intelligentie. De artificiële intelligentie maakt dus beslissingen op basis van een patroon in de uitkomsten van menselijk ethisch handelen en niet op basis van een

26 Ibidem. 61.

afweging van principes in een specifieke context. De artificiële intelligentie is op deze manier wel in staat om ethisch menselijk handelen te immiteren, maar niet om op dezelfde manier ethisch te handelen als mensen.

Dit maakt dat de theorie van Ross uiteindelijk niet één op één implementeerbaar is in een zelfrijdende auto. Dit kan leiden tot de volgende kritische reactie: Als de artificiële intelligentie zelf niet ethisch handelt, maar dit slechts immiteert, wat is er dan ethisch aan? Mijn antwoord: De artificiële intelligentie handelt inderdaad niet ethisch, maar dit is ook niet waar het om gaat. Waar het wel om gaat is de vraag hoe we de zelfrijdende auto het beste kunnen programmeren. Als verdediger van het particularisme van Ross denk ik dat een artificiële intelligentie die het ethisch handelen van mensen kan immiteren de beste manier is om de zelfrijdende auto te programmeren. Ethisch handelen kunnen alleen mensen. Een artificiële intelligentie die het ethisch handelen van mensen kan immiteren is de beste optie die we hebben als we in de buurt willen komen van ethisch handelen.

Met welke gegevens programmeren we de artificiële intellgentie

Een andere vraag die mijn voorstel oproept is: Wie en hoe wordt bepaalt welke statistische gegevens relevant zijn voor de artificiële intelligentie?

De manier waarop de artificiële intelligentie gaat leren om ethisch handelen van mensen te immiteren is door een groot bestand aan ethische casussen en menselijke beslissingen in die casussen als uitgangspunt te nemen voor het vormen van algoritmes. Het gevaar dat hier in schuilt is dat de artificiële intelligentie makkelijk gestuurd kan worden om een bepaald soort ethische beslissingen te nemen. Stel we geven de artificiële intelligentie uitsluitend casussen waarin een utilistische calculus de doorslag geeft in de beslissingen om te handelen. De artificiële intelligentie zal dan zelf ook volledig utilistisch worden. Hetzelfde effect treed op als we de machine alleen casussen met deontologische beslissingen geven. De machine zal dan volledig deontologisch worden. Het is dus van belang dat de casussen en bijbehorende ethische beslissingen een goede weergave zijn van de complexe manier waarop mensen ethische beslissingen nemen.

Op dit punt zou ik kunnen zeggen dat de beste manier om aan casussen en ethische beslissingen te komen door in te kijken hoe mensen in de wereld ethische beslissingen maken. Toch denk ik dat hier grote problemen mee zijn. Ten eerste is het heel vaak niet duidelijk wanneer beslissingen die genomen worden in de wereld ethisch zijn. Een persoon kan een beslissing maken die ethisch lijkt, maar eigenlijk door hele andere dingen gemotiveerd zijn. Daarnaast zijn ethische beslissingen in de echte wereld bijna nooit uitsluitend ethische beslissingen. Meestal spelen ook andere niet-ethische motivaties een rol bij het nemen van een ethische beslissing. Hierdoor kun je je afvragen of daadwerkelijke beslissingen van mensen in de echte wereld wel een goede leidraad zijn voor de artificiële intelligentie. Het ethisch handelen van mensen is verre van perfect en één van de voordelen van de artificiële intelligentie is juist dat hij niet onderhevig hoeft te zijn aan andere motivaties dan ethische.

In plaats van ethische beslissingen in de wereld stel ik voor om gebruik te maken van geconstrueerde ethische casussen. Wie moet deze casussen opstellen? Ik denk dat ethici het beste in staat zijn om de cacsussen te ontwerpen. Hier moet niet te makkelijk over gedacht worden. De uitkomsten van de ethische casussen worden in grote mate bepaald door hoe de casussen ontworpen zijn. Het is de uitdaging voor ethici om een zo groot mogelijke variëteit aan casussen te ontwerpen. De reden dat ik denk dat ethici deze casussen moeten ontwerpen is omdat zij het meest verstand hebben over welke

ethische redenen mensen kunnen hebben. In de casussen waar de juiste ethische beslissing duidelijk is kunnen ethici de desbetreffende casus en het bijbehorende antwoord meteen aan de artificiële intelligentie doorgeven. De juiste ethische beslissing is duidelijk wanneer vanuit verschillende theoretische perspectieven en de intuïtie dezelfde keuze als de juiste wordt aangeduid. Alleen als aan beide eisen wordt voldaan en er geen andere optie is die ook theoretisch verantwoord kan worden, spreken we van een duidelijke casus.

In sommige gevallen is het echter niet duidelijk wat de juiste beslissing is. Toch moeten ook deze casussen als input worden gebruikt voor de artificiële intelligentie om een zo volledig mogelijk beeld te kunnen geven van menselijk ethisch handelen. Om goed te kunnen weergeven hoe mensen handelen in de moeilijke casussen kan gebruik worden gemaakt van ethische enquêtes. Het idee is dat een grote groep mensen een groot aantal moeilijke ethische casussen krijgt voorgelegd waarin ze een beslissing moeten maken. De deelnemers van de enquête worden gevraagd om antwoord te geven op de vraag wat ethisch de beste beslissing is. Het voordeel van deze oplossing ten opzichte van ethische beslissingen in de wereld is dat mensen die geconfronteerd worden met de casussen in de enquête zich uitsluitend kunnen richten op ethische motivaties om hun antwoord te geven aangezien er voor hen geen gevolgen zijn van hun keuzes. Dit zorgt ervoor dat het patroon dat de artificiële intelligentie ziet uitsluitend uit ethische motivaties bestaat.

Een mogelijk bezwaar tegen deze manier van input geven aan de artificiële intelligentie is dat op deze manier de artificiële intelligentie simpelweg weerspiegelt wat de meerderheid moreel juist vindt. Op dit bezwaar heb ik twee dingen te zeggen. Ten eerste geldt dit niet voor de makkelijke casussen. Bij de makkelijke casussen is er maar één keuze die te verantwoorden is met een ethische theorie en aansluit bij de intuïtie. De meerderheid is in dit geval dus niet bepalend, maar de ethische kennis van de ethicus in combinatie met zijn inzicht. Het bezwaar geldt dus alleen voor de moeilijke casussen, waar de ethische enquête wordt gebruikt. Ten tweede heeft de meerderheid ook in de moeilijke casussen geen allesovertreffende invloed. Dit kan door niet slechts het antwoord dat het meest gegeven is door te geven aan de artificiële intelligentie. De antwoorden die deelnemers van de ethische enquête geven worden allemaal als input aan de artificiële intelligentie doorgegeven. De artificiële intelligentie weet dus niet alleen welk antwoord het meest gegeven is in een bepaalde casus, maar ook welke andere antwoorden er zijn gegeven en hoe vaak die zijn gegeven. Deze gegevens worden meegenomen in het opstellen van de modellen door de artificiële intelligentie. De invloed die de meerderheid heeft op de artificiële intelligentie is dus beperkt.

Wat doen we als de artificiële intelligentie, ondanks de bovenstaande methode, alsnog een keuze maakt die tegen de voorkeur van een meerderheid in de samenleving in gaat? Dit geeft eigenlijk aan

dat de voorkeur van de meerderheid inconsistent is met de input die de artificiële intelligentie tot nu toe heeft gehad. In dit geval denk ik dat we moeten reflecteren of onze inconsistentie gerechtvaardigd kan worden door middel van argumenten. Als na de reflectie blijkt dat de beslissing van de artificiële intelligentie zeer goed verdedigbaar is terwijl de voorkeur van de meerderheid niet verder komt dan een intuïtie, dan moet de voorkeur worden doorgegeven aan de artificiële intelligentie en verder niet meer dan dat. Op deze manier kan de artificïele intelligentie gebruik maken van de input dat veel mensen hier een bepaalde intuïtie over hebben. Als blijkt dat de voorkeur van de meerderheid even goed of zelfs beter verdedigbaar is dan moet aan de artificiële intelligentie doorgegeven worden dat hij fout zat. Op deze manier maakt de artificiële intelligentie een sterkere aanpassing in zijn algoritmes waardoor hij deze fout niet meer zal maken. Een fout als deze kan echter alleen maar voorkomen wanneer de input van de artificiële intelligentie tekortschiet. Bij een artificiële intelligentie die genoeg input heeft gehad zou dit niet moeten kunnen gebeuren.

Ik ben me ervan bewust dat ik door het bovenstaande voor te stellen me niet helemaal los kan maken van de kritiek dat de meerderheid bepalend is voor hoe de artificiële intelligentie zich gedraagd. Hiermee verwijder ik me ook engiszins van Ross, die niet verdedigt dat de morele intuïties van de meerderheid hetzelfde zijn als de juiste ethische keuze. Dit ben ik met Ross eens, maar wat ik hier verdedig is een poging om zo dicht mogelijk in de buurt te komen van een particularistische artificiële intelligentie. Ik denk dat het gebruikmaken van de morele intuïties van de meerderheid ons het meest dicht in de buurt brengt bij een particularistisch geprogrammeerde zelfrijdende auto. De reden dat ik denk dat een meerderheid ons in de buurt brengt is omdat ik de aanname doe dat de intuïtie van de meerderheid in gevallen waarin alle relevante informatie gegeven is en er slechts om een ethische overweging wordt gevraagd meestal een goede indicatie geeft wat de goede ethische beslissing is. De casussen die aan de artificiële intelligentie gegeven worden zijn voorbeelden van een geval waarin alle relevante informatie gegeven is en er slechts een ethische overweging hoeft te worden gemaakt. Deze aanname lijkt me niet problematisch omdat geen enkele normatieve theorie zich uiteindelijk helemaal los kan maken van de morele intuïties van een meerderheid. Zoals ik ook in het eerste hoofdstuk van dit paper heb betoogd zijn deontologische en consequentialistische theorieën uiteindelijk gebaseerd op een intuïtie. Deze intuïtie moet op zijn minst gedeeld door een grote groep mensen om het tot een plausibele theorie te maken.

Conclusie

In dit paper heb ik betoogd dat het particularisme van Ross een plausibel alternatief is om zelfrijdende auto's mee te programmeren. Eerst heb ik laten zien wat de problemen zijn van zelfrijdende auto's. Deontologische theorieën hebben vooral moeite met het implementeren van deontologische principes in een zelfrijdende auto. Consequentialistische theorieën zijn makkelijker te implementeren in zelfrijdende auto's, maar ook daar is een probleem mee. Daarnaast hebben deontologische en consequentialistische theorieën ook problemen met intuïtieve bezwaren. Het particularisme van Ross kan beter omgaan met deze intuïtieve bezwaren. Vervolgens begin ik met mijn voorstel. Ik heb laten zien op welke manier ik denk dat het particularisme van Ross het best geïmplementeerd kan worden in een zelfrijdende auto. Dit kan worden gedaan door een artificiële intelligentie die gebruik maakt van *machine learning*. Volgens het particularisme van Ross zal de artificiële intelligentie niet in staat zijn om ethisch te handelen op dezelfde manier als mensen dat doen. Het beste wat de artificiële intelligentie kan doen is menselijk ethisch handelen immiteren. Dit doet de artificiële intelligentie door middel van grote hoeveelheden statistische gegevens. Als laatst verdedig ik waarom de statistische gegevens moeten bestaan uit ethische casussen en bijbehorende ethische beslissingen. Deze casussen en beslissingen worden vormgegeven door ethici. Hierbij maak ik een onderscheid tussen duidelijke en onduidelijke casussen. In het geval van duidelijke casussen is het inzicht en theoretische verantwoording van de ethicus genoeg om de casus met beslissing direct door te geven aan de artificiële intelligentie. In het geval van onduidelijke casussen doe ik een beroep op de ethische intuïties van een groep mensen om beslissingen door te kunnen geven.

Ik hoop dat ik met dit paper heb kunnen laten zien dat consequentialisme niet onze enige mogelijkheid is om zelfrijdende auto's te programmeren in de toekomst. Ik vind het zorgelijk dat de techniek van zelfrijdende auto's al zo ver gevorderd is terwijl het debat over ethiek van zelfrijdende auto's in mijn optiek nog in de kinderschoenen staat. Er is ongetwijfeld veel aan te merken op het voorstel dat ik hier verdedigd heb, maar wat het op zijn minst laat zien is dat het debat niet breed genoeg is en er hoogstwaarschijnlijk mogelijkheden zijn die nog niet verkend zijn.

Bibliografie

1. Alexander, Larry and Moore, Michael, "Deontological Ethics", The Stanford Encyclopedia of Philosophy (Spring 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2015/entries/ethics-deontological/>.
2. Anderson, Monica, Nicolson, Chris, en Schrader, Christopher. *What are the main differences between artificial intelligence and machine learning?* (blog). 2015. https://www.quora.com/What-are-the-main-differences-between-artificial-intelligence-and-machine-learning. (laatst geraadpleegd 21 juni 2016).
3. Bonnefon, Jean-François, Shariff, Azim en Rahwan, Iyad, *'Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?'* (2015). 1-15. http://arxiv.org/abs/1510.03346.
4. Dancy, Jonathan, 'An Unprincipled Morality' in: Russ Shafer Landau (red.) 'Ethical Theory an Anthology' (2013) 772-776.
5. Goodall, N. J., 'Ethical Decision Making During Automated Vehicle Crashes' in: *Transportation Research Record: Journal of the Transportation Research Board*, volume 2424 (2014) 58-65.
6. 'Google Self-Driving Car Project' (Laatst geraadpleegd 21 juni 2016)https://www.google.com/selfdrivingcar/
7. Hibbard, Bill. *Ethical Artificial Intelligence.* Berkeley: Space Science and Engineering Center University of Wisconsin - Madison and Machine Intelligence Research Institute, 2015. http://arxiv.org/pdf/1510.03346v1.pdf. (laatst geraadpleegd 21 juni 2016).
8. Lin, Patrick, 'Why Ethics Matters for Autonomous Cars', in: Markus Maurer, J. Christian Gerdes, Barbara Lenz, Hermann Winner (red.), *Autonomes Fahren* (Berlin 2015) 69-85.
9. Ross, W.D., 'What Makes Right Acts Right?' in: Russ Shafer Landau (red.) 'Ethical Theory an Anthology' (2013) 756-763.
10. Sinnott-Armstrong, Walter, "Consequentialism", *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2015/entries/consequentialism/>.