Calculo Ergo Sum (2)

Over het recht en de ethiek van autonome systemen in computernetwerken[1]

dr. mr. C.N.J. de Vey Mestdagh[2]

1. Autonome systemen in computernetwerken kunnen twijfelen

Van computersystemen wordt steeds aangenomen dat zij niet twijfelen. Computersystemen zijn weliswaar net als mensen wat beperkt in hun kennis, maar zij zijn volgens velen wel rationeel en objectief. Ons beeld van een computersysteem wordt nog steeds bepaald door het model van dedicated machines, zoals een rekenmachine, met volledig beschrijfbare doelen, functies en invoer- en uitvoerverzamelingen. Het probleem dat hiermee wordt geassocieerd is dat wij als subjecten van computerbeslissingen over een kam worden geschoren, met andere woorden dat er geen rekening wordt gehouden met onze individuele omstandigheden. Ongelijke gevallen worden gelijk behandeld. En zelfs de voorspelbaarheid van die behandeling komt in gevaar omdat wij geen inzicht hebben in de gegevens en kennis waarop deze ongenuanceerde beslissingen zijn gebaseerd. Indien de gebruikte algoritmen en de omgeving (*context*) waar ze in werken een bepaalde (niet eens zo hoge) mate van complexiteit bereiken geldt deze *onvoorspelbaarheid* niet alleen voor de subjecten van de beslissingen, maar ook voor de *principalen*[3] van de computersystemen die beslissingen nemen. Het laatste maakt de beslissingen van deze systemen *autonoom*.[4] Ons hoogste juridische adviescollege formuleert het als volgt: *Bij digitalisering van de besluitvorming […] dreigt de burger in toenemende mate te worden geconfronteerd met besluiten die volautomatisch zijn genomen, zonder menselijke tussenkomst. Die burger kan niet meer nagaan welke regels zijn toegepast en het is niet meer vast te stellen of de regels ook werkelijk doen waarvoor ze bedoeld zijn. Ook dreigt de burger slachtoffer te worden van een robotachtige gelijkheid, waarbij geen oog meer bestaat voor de eigenheid van zijn situatie.*[5] De Raad adviseert op grond hiervan de transparantie van dergelijke besluitvorming aanzienlijk te verbeteren.

Dit beeld van een computersysteem als volledig gedetermineerd, rationeel, objectief en derhalve ongenuanceerd systeem[6] is gebaseerd op de gebruikelijke wijze van automatisering van de (juridische) besluitvorming. Van zowel de gebruikte gegevens als de gebruikte kennis wordt voor het gemak aan-

1. Dit artikel is een vervolg op C.N.J. de Vey Mestdagh, 'Calculo Ergo Sum (1)', *Recht en Elektronische media* 1999, afl. 1, p. 2-3, http://law-and-ict.org/wp-content/uploads/2019/10/Calculo-ergo-sum.pdf en een uitwerking van C.N.J. de Vey Mestdagh, 'Verdienen autonome systemen die onrechtmatig handelen de doodstraf?', Lezing op het Seminar kunstmatige intelligentie en ethiek, NVvIR Jong, 7 maart 2019, Deloitte Amsterdam.
2. Kees de Vey Mestdagh is directeur R&D bij het Softwareborg Instituut en verricht wetenschappelijk onderzoek voor de Foundation for Law&ICT (e-mail: c.n.j.de.vey.mestdagh@law-and-ict.org). Daarnaast is hij hoofdredacteur van het Tijdschrift voor Internetrecht.
3. Hier wordt met opzet het begrip *principaal* (opdrachtgever) gebruikt en bijvoorbeeld niet *verantwoordelijke* of *vertegenwoordigde* omdat het nu net de vraag is of de principaal verantwoordelijk en/of vertegenwoordigd is.
4. Zie bijvoorbeeld C.N.J. de Vey Mestdagh, *Juridische Kennissystemen, rekentuig of rekenmeester?* (diss. Groningen), Serie Informatica en Recht, nr. 18, Deventer: Kluwer 1997, http://law-and-ict.org/wp-content/uploads/2018/03/proefschrift-de-vey-mestdagh-RUG-printed-with-all-fonts-embedded.pdf, waarin de autonomie van een juridisch kennissysteem er toe blijkt te leiden dat 66% van door mensen genomen beslissingen in complexe milieuvergunningencasus (ernstige) fouten blijken te bevatten.
5. Ongevraagd advies van de Raad van State over de effecten van de digitalisering voor de rechtsstatelijke verhoudingen van 31 augustus 2018 (*Kamerstukken II* 2017/18, 26643, 557).
6. Zie onder andere J.J. Dijkstra, 'Legal knowledge-based systems: The blind leading the sheep?', in: *International Review of Law, Computers & Technology*. *15*, 2001, p. 119-128. Juristen beschouwen een juridisch kennissysteem als objectief en rationeel, ook als het in feite subjectief is en conclusies trekt op grond van eenzijdige en onvolledige kennis.

genomen dat zij volledig zijn.[7] Want welke jurist wil er nu een juridisch kennissysteem dat zegt 'Ik weet het niet', 'Zoek het zelf maar uit', 'Hier zijn verschillende mogelijkheden' of zelfs 'Hier kun je verschillend over denken'. Toch zou een computersysteem dit moeten kunnen zeggen, omdat hij anders geen betrouwbare beslissingen kan nemen. De gebruiker weet anders nooit of een gegeven onbekend is, dat er kennis ontbreekt of dat er geen zekerheid is over een gegeven of kennis en derhalve onzekerheden, argumenten en opvattingen concurreren. Gelukkig zijn er beproefde technieken om een beslissend of beslissingsondersteunend computersysteem wél in staat te stellen tot een dergelijke dialoog.[8] Ook is het bouwen van *objectief* transparante systemen geen enkel probleem. Hiermee zouden weliswaar de ongenuanceerdheid en het gebrek aan objectieve transparantie zijn opgeheven, maar niet het gebrek aan voorspelbaarheid en de *subjectieve* transparantie van autonome systemen. De beslissingen van computers zijn namelijk niet alleen afhankelijk van de van tevoren beschikbare gegevens en kennis die in hun algoritmen is opgenomen, maar ook van de *context* waarin ze deze beslissingen nemen. Deze context bestaat natuurlijk uit gegevens uit externe bronnen die gedurende het beslissingsproces worden verzameld, maar ook uit interacties met andere beslissende systemen. Zelfrijdende auto's verzamelen bijvoorbeeld externe gegevens met hun eigen sensoren en door interacties met andere gegevensverwerkende systemen (andere zelfrijdende auto's, weer- en verkeerssystemen, navigatiesystemen, etc.). Ook wordt de subjectieve transparantie zoals gezegd niet opgeheven. Een systeem kan objectief transparant zijn omdat er tenminste één expert is die het systeem kan doorgronden, maar het moet ook subjectief transparant zijn zodat ook de principalen en de subjecten van zijn beslissingen dat kunnen.

Het opereren in een *netwerk* van onzekere gegevens en kennisperspectieven maakt een computersysteem subjectief, dat wil zeggen zijn kennis perspectief gebonden en geeft ruimte voor twijfel. De zelfrijdende auto kan bijvoorbeeld vaststellen dat de weg over enkele kilometers nat zou kunnen worden door een komende regenbui en kan dan de keuze hebben om sneller te gaan rijden om de regen voor te zijn of langzamer te gaan rijden om veilig door de regenbui heen te komen. Hoe meer gegevens en interacties hoe meer van deze onzekerheden er ontstaan.

Bij het maken van afwegingen tussen bestaande alternatieven behoren ook belangen in het beslissingsproces meegewogen te worden. Deze belangen kunnen worden behartigd door toepassing van onze formele en informele normen (rechtsregels en sociale normen). In beide gevallen begeven we ons dan op het gebied van de e*thiek* (het vaststellen van de criteria om te kunnen beoordelen of een beslissing en de daarop gebaseerde handeling als goed of fout kan worden gekwalificeerd). Deze normen en hun onderliggende waarden vergroten de twijfel en maken het beslissingsproces nog complexer. De zelfrijdende auto zou bijvoorbeeld mee kunnen wegen welk verschil in gevaarzetting, vervuiling en economische gevolgen er is tussen beide genoemde alternatieven (sneller dan wel langzamer gaan rijden).[9]

Kort gezegd verdelen steeds complexere netwerken van gegevens- en kennisverwerking de verantwoordelijkheid over steeds meer verwerkingen en verwerkingseenheden, waardoor er mogelijk geen verantwoordelijke meer overblijft. Verantwoordelijkheid betreft immers het (al dan niet juridisch) instaan voor de gevolgen van het eigen handelen (inclusief nalaten). Het handelen is in deze situatie zo versnipperd over het netwerk en bovendien in zo een grote mate overgenomen door machines en onvoorspelbaar geworden, dat het individuele menselijke handelen geen voor de consequenties van het handelen van het netwerk (al dan niet juridisch) relevant gegeven meer is. Deze situatie valt beter te begrijpen als men zich voorstelt dat een onderdeel van de hersenen (een neuronaal netwerk) van een persoon verantwoordelijk zou worden gesteld voor de handelingen van de persoon. Betreft het een gebrekkig onderdeel dan wordt de persoon als geheel handelingsonbekwaam of ontoerekeningsvatbaar verklaard, betreft het een niet gebrekkig onderdeel dan wordt de persoon als geheel aansprakelijk gesteld of gestraft.

Ter onderbouwing van het voorgaande wordt in dit artikel eerst een casus behandeld die laat zien dat computersystemen autonoom beslissingen kunnen nemen. Vervolgens wordt bekeken welke rechtsgevolgen het autonoom beslissen van computers kan hebben. De vraag is waar de verantwoordelijkheid

7. Er zijn nog steeds veel databanken en kennissystemen gebaseerd op de drogreden Argumentum ad ignorantiam. De hierbij gehanteerde Closed World Assumption en de inference rule Negation As Failure maken doorredeneren met incomplete kennis mogelijk zonder dat deze systemen stoppen. Dit is alleen geen probleem als de gegevens en kennis volledig zijn, hetgeen vrijwel nooit het geval is.
8. Zie hierover het werk van M. Burgin (Dept. Of Mathematics, UCLA) en C.N.J. de Vey Mestdagh op het gebied van inconsistente juridische kennis en logical varieties. Een overzicht daarvan is te vinden in de literatuurlijst van C.N.J. de Vey Mestdagh, 'A Model of Complexity for the Legal Domain', in: *Proceedings of the IS4SI 2017 Summit Digitalisation for a sustainable society 1*, 192, Gothenburg, Sweden 12–16 June 2017. Issue Editor: Gordana Dodig Crnkovic. ISSN 2504-3900, http://www.mdpi.com/2504-3900/1/3/192.
9. Zie voor een uitgebreide toelichting op het autonome karakter van de zelfrijdende auto, het geldende recht en het tekortschieten daarvan C.N.J. de Vey Mestdagh & J. Lubbers, 'Nee hoor, u wilt helemaal niet naar Den Haag..., Over de techniek, het recht en de toekomst van de zelfrijdende auto', *Ars Aequi*, april 2015, p. 267-280, http://law-and-ict.org/wp-content/uploads/2019/10/Nee-hoor-u-wilt-helemaal-niet-naar-den-Haag-AA20150267-over-de-techniek-het-recht-en-de-toekomst-van-de-zelfrijdende-auto.pdf.

(zoals hiervoor beschreven) ligt voor beslissingen die door autonome systemen worden genomen, in gevallen waarin de menselijke principalen het gedrag van hun computer-agents niet meer kunnen voorspellen en zich evenmin bewust zijn van de risico's. Ten slotte wordt een voorstel gedaan om dit probleem op te lossen.

2. Casus Flash Crash[10]

6 mei 2010, 14:32 u

Een grote trader besluit om 75.000 future contracten[11] ter waarde van ca. $4,1 miljard[12] aan te bieden op een Amerikaanse beurs. Deze trader gebruikt daarbij de volgende strategie: verkoop 9% van het handelsvolume in de voorafgaande minuut, ongeacht de prijs en zonder beperking van de verkoopsnelheid. De ratio achter deze strategie is dat de verkoop slechts een beperkt percentage van het bestaande handelsvolume betreft, zodat de prijs niet te veel wordt gedrukt door overaanbod en de trader vervolgens ook de rest van zijn contracten tegen een redelijke prijs kan verkopen. Op deze dag beginnen andere traders, zoals gebruikelijk, te kopen als de prijs daalt door het grote initiële aanbod van onze trader, maar ook weer te verkopen omdat de prijs weer stijgt door hun eigen vraag. Als de prijs vervolgens weer daalt door het grote aanbod (van henzelf en van onze trader) wordt de verkoopgolf en de daarop volgende prijsdaling nog verder aangewakkerd. Bij kleinere volumes is dit geen probleem, er ontstaat een evenwichtsprijs. Bij het grote volume en het continue aanbod van 6 mei verloopt het anders.

Hot Potato

Tussen 14:45:13 en 14:45:27 (14 seconden) worden bruto meer dan 27.000 contracten verhandeld, terwijl het netto slechts ging om circa 200 contracten (het verschil tussen aantallen verkoop- en koopcontracten). Tussen 14:41:00 en 14:45:27 (4,5 minuten) daalden de prijzen op de E-Mini meer dan 5%. Onze trader houdt het vuurtje aan de gang en gaat volgens zijn verkoopstrategie steeds meer verkopen omdat het handelsvolume (schijnbaar) stijgt. In 20 minuten (in plaats van de gebruikelijke 5 uur), verkoopt onze trader alle 75.000 contracten. Cross-market trading firms verkopen onderliggende waarden en trekken zich vanwege de grote volatiliteit terug van andere markten waardoor ook daar de indices dalen (de Dow Jones Index daalde in een half uur met 9%, dat wil zeggen met circa duizend miljard dollar). De traditionele verklaringen voor deze zogeheten hot potato zijn de al grote volatiliteit van de markt voordat onze trader deze betrad en een lage buy side liquidity, dat wil zeggen beschikbare middelen om fondsen aan te kopen, vanwege de Europese (Griekse) schuldencrisis. De buy side liquidity *voor alle fondsen* bedroeg op deze dag $2,65 miljard. Het aanbod van onze trader overtrof deze ruimschoots.

Flash Crash

De werkelijke verklaring was echter dat onze trader en de meeste andere traders zogenaamde High Frequency Traders zijn, oftewel machines die zelf beslissingen nemen! Het gevolg daarvan is deze Flash Crash. De oplossing die in dit geval werd gekozen was een aanpassing van de zogeheten circuitbreakers van 60% naar 10%. Circuit breakers pauzeren de handel in securities gedurende 5 minuten als de prijs gedurende de voorafgaande vijf minuten met het ingestelde percentage verandert. Maar machines kunnen natuurlijk ook worden geprogrammeerd om met deze circuit breakers rekening te houden.[13] Als de circuitbreaker werkt (de markt stabiliseert, maar ook spanning oplevert omdat verkoop of kooporders moeten worden opgeschort) dan is het gunstig om vlak voor en na de break te kopen of te verkopen. Met een selffulfilling prophecy ten gevolge, want de koersen dalen of stijgen na de break verder! Ook dan zullen High Frequency Traders een grotere marktonbalans opleveren door hun snelheid en grotere volumes.

3. Netwerkcomputers kunnen autonoom handelen met rechtsgevolg

De vraag is hoe het samen handelen van de principalen en de autonome computers in het netwerk juridisch kan worden gekwalificeerd. Van welke *rechtsfeiten (rechtshandelingen, feitelijke handelingen of blote rechtsfeiten)* is er sprake?

De casus Flash Crash laat het volgende zien. Autonoom beslissende systemen bestaan, want het was niet de wil van de principaal van het systeem om een crash te veroorzaken en hij herkende dat risico evenmin. Anders had hij het wel gelaten om een

10. De beschrijving van de Flash Crash van 6 mei 2010 is geconstrueerd aan de hand van een groot aantal internetbronnen, in het bijzonder de *Findings Regarding the Market Events of May 6, 2010* (Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues, September 30, 2010), https://www.sec.gov/news/studies/2010/marketevents-report.pdf.
11. Een future contract of termijncontract is een financieel contract waarbij partijen zich verbinden om op een bepaald moment in de toekomst een bepaalde hoeveelheid van een product of financieel instrument, bijvoorbeeld aandelen, te verhandelen tegen een vooraf bepaalde prijs. Futures worden verhandeld op een (effecten)beurs. In deze casus betrof het de E-Mini S&P 500, een future contract op de Standard & Poor 500 aandelenindex, dat wordt verhandeld op het Globex platform van de Chicago Mercantile Exchange group.
12. Dit is een groot, maar zeker niet bijzonder groot aanbod. De E-Mini S&P 500 heeft een gebruikelijke dagomzet van circa $100 miljard.
13. Dit levert een aardig voorbeeld van een Prisoner's dilemma op, waarbij twee HFT's door samen te werken beiden winst kunnen maken.

zo grote schade te veroorzaken dan wel op de koop toe te nemen. Het staat vast dat de beslissingen van autonome systemen ongewenste effecten kunnen hebben. In dit geval een ernstige marktverstoring en in andere gevallen verkeerde beslissingen van zelfrijdende auto's of beslissende systemen in het openbaar bestuur, de medische praktijk, etc. Ook is duidelijk dat de beslissingen niet in overeenstemming hoeven te zijn met de wil van de (menselijke) principalen, d.w.z. zij hoeven geen rechtshandelingen te zijn, maar kunnen feitelijke handelingen zijn (het in werking stellen van het systeem). Een feitelijke handeling kan een onrechtmatig karakter hebben. In de meeste gevallen is dan schuld vereist. In het geval van autonome systemen is daar per definitie geen sprake van (het gedrag van deze systemen is immers onvoorspelbaar voor de principaal). Ook in de gevallen waar geen schuld vereist is, is het bijvoorbeeld noodzakelijk dat de principaal wel het risico op bepaalde mogelijke gevolgen neemt. Daarvoor is bewustzijn van de risicodragende hoedanigheid of van het feitelijke risico vereist. Zo is bijvoorbeeld de werkgever aansprakelijk voor de gedragingen van de werknemer (art. 6:170 BW), de opdrachtgever voor gedragingen van de opdrachtnemer (art. 6:171 BW), de eigenaar van een roerende zaak voor gebreken daaraan, die schade veroorzaken (art. 6:173 BW) en de automobilist voor het toebrengen van schade aan zwakkere verkeersdeelnemers (art. 185 WvW). De werkgever, opdrachtgever, eigenaar en automobilist behoeven geen schuld te hebben, maar zij moeten wel het risico kunnen voorzien. Ook daarvan is in de behandelde casus en daarmee in gelijksoortige casus waarin autonome systemen worden toegepast geen sprake. De samenloop van omstandigheden (van gegevens en gegevensverwerkende systemen) is daarvoor te onwaarschijnlijk.

Als de principaal geen schuld heeft en zich niet bewust is van het risico is er sprake van een bloot rechtsfeit met rechtsgevolg, maar zonder aansprakelijkheid.[14] Er ontstaat een feitelijk *en* juridisch autonoom systeem, waarvan de consequenties van bepaalde beslissingen niet (met een bepaalde zekerheid) voorspelbaar zijn en waarvan de juridische consequenties niet met menselijke handelingen zijn verbonden.

Autonoom beslissende systemen kunnen rechtsfeiten met rechtsgevolgen creëren.

De gevolgen van de beslissingen van de computersystemen in de casus zijn niet alleen onwenselijk maar ook onrechtmatig. Zij veroorzaken een grote schade doordat zij een ernstige marktverstoring teweegbrengen. Nu er geen verwijtbare handeling (met schuld of risicobewustzijn) aan vooraf is gegaan is er geen sprake van wanprestatie, onrechtmatige daad of van een strafbaar feit. Er lopen tientallen rechtszaken, maar tot nu toe zonder duidelijk resultaat.[15] De feitelijke oorzaken zijn meervoudig en staan onvoldoende vast (complexiteit) en individuele personen hebben veelal als onderdeel van een handelingsnetwerk (te) weinig individuele verantwoordelijkheid voor de gevolgen. In dit geval worden de rechtsgevolgen van onrechtmatig handelen door de toezichthoudende instanties (SEC en de CFTC) geaccepteerd. Er is geen enkele transactie teruggedraaid en er zijn geen civiele of strafrechtelijke gevolgen voor de principaal geweest.

De vraag is welke juridische of – ruimer – ethische aansprakelijkheidsnormen van toepassing zijn als dit soort gevallen (technologie die onvoorspelbaar is waardoor verantwoordelijkheid van de principaal ontbreekt) zich (steeds vaker) voordoen. Er zijn hiervoor drie mogelijke denkrichtingen:

1. De mens moet worden weerhouden van het gebruik van bepaalde technologie (verbod), of
2. De mens moet worden gedwongen tot beter gebruik van deze technologie (verplichting), of
3. De technologie zelf moet beter worden ingericht (toestemming tot gebruik onder voorwaarden, dat wil zeggen toestemming om systemen te gebruiken die aantoonbaar in staat zijn om zelf juridische of ruimer ethische normen toe te passen).

4. De ethiek van autonome systemen: verbod, verplichting of toestemming onder voorwaarden?

Het hiervoor geïdentificeerde probleem is: wie (of wat) is aansprakelijk in geval van het veroorzaken van schade door autonoom beslissende systemen zonder verwijtbare gedraging of bewustzijn van het risico bij de principaal? De conclusie dat het hier een beperkt aantal systemen betreft (zoals het behandelde trading systeem) is onjuist. De autonomie van het systeem wordt weliswaar bepaald door zijn complexiteit, maar deze complexiteit is niet objectief, maar subjectief. Met andere woorden: de complexiteit van het systeem is afhankelijk van de mate van (dan wel het gebrek aan) complexiteit van de principaal.[16] Om dit te verduidelijken: het is niet waarschijnlijk dat de Nederlandse wetgever besluit dat objectief steeds complexere vervoerssystemen, zoals de zelfrijdende auto, slechts gebruikt mogen worden door een deskundige elite, met een beter voorspellingsvermogen dan gemiddeld en het is zelfs uitgesloten dat de grote en toenemende aantallen autonome systemen bij de overheid en in het

14. Er zijn meer situaties waarin dit het geval is. Bijvoorbeeld als de principaal een rechtspersoon is die ontbonden is of een natuurlijke persoon is die gestorven is of handelingsonbekwaam is.

15. In een enkel enigszins vergelijkbaar geval is een veroordeling gevolgd. Er moet dan anders dan in de *Flash Crash*-casus wel sprake zijn van *spoofing*: bewuste marktmanipulatie met behulp van algoritmen. Zie 7th U.S. Circuit Court of Appeals, 7-8-2017, United States v Coscia, No. 16-3017.

16. Deze wijze van denken is juristen niet vreemd. Zie bijvoorbeeld het *Haviltex*-arrest (HR 13 maart 1981, *NJ* 1981/63), waarin de uitleg van overeenkomsten afhankelijk wordt gemaakt van de (relatieve) deskundigheid van de partijen.

bedrijfsleven bij elke beslissing voldoende deskundige principalen hebben (hiervoor ontbreken de deskundigheid bij de ambtenaren en werknemers en de financiële middelen om die deskundigheid voldoende te verhogen).[17]

Samengevat: voor de principaal voorspelbare (gedetermineerde) systemen leveren geen nieuwe ethische- of rechtsvragen op omdat de principaal dan schuld- of risicoaansprakelijk is. Autonome systemen zijn voor de principaal echter niet voorspelbaar. De principaal mist de cognitieve capaciteiten en kennis om het systeem in zijn context te doorgronden en commerciële verdienmodellen verzetten zich tegen de voor voorspelbaarheid vereiste transparantie.

Wat betekent dit nu voor de drie genoemde oplossingsrichtingen? Zolang de deskundigheid van de principalen en subjecten van autonome systemen niet voldoende is zouden deze systemen eigenlijk moeten worden verboden. Dit is echter een ouroboros. De wetgever kan niet zomaar specifieke systemen verbieden, maar zal bepaalde klassen van systemen verbieden. De bestuurder kan evenmin de taak gedelegeerd krijgen om van elke klasse de extensie te beschrijven (de concrete systemen die daaronder vallen). Het gaat namelijk om veel systemen, waarvoor veel expertise vereist is om dit oordeel te kunnen geven. Ook kan niet van de ontwikkelaars (makers) verwacht worden dat zij hun eigen systemen aan deze toets onderwerpen. Hiervoor is immers niet alleen technische (makers)kennis vereist, maar ook kennis van de omgeving waarin het systeem zal functioneren (bijvoorbeeld het verkeer, de markt en de daarin gebruikelijke processen zoals in de behandelde casus het geval is) en kennis van het deskundigheidsniveau van de toekomstige gebruikers. De gebruiker moet daarom bepalen of het systeem onder een van deze klassen valt. Hoe stel je als principaal echter vast dat een verboden technologie wordt gebruikt (er sprake is van een autonoom systeem)? Het systeem is immers verboden als het te complex is, om de minder complexe principaal te beschermen, terwijl deze wel moet vaststellen of het systeem niet te complex is om te kunnen bepalen of het gebruik verboden is. Eenzelfde argument geldt voor de tweede oplossingsrichting (de verplichting tot het verantwoordelijk gebruiken van autonome systemen). Die verplichting richt zich immers rechtstreeks tot de gebruiker (de principaal).

Kortom de toepassing van verboden of verplichtingen in verband met autonome systemen vergen een (nog) niet haalbaar deskundigheidsniveau bij de adressanten ervan.[18] Zelfs als dit niet zo zou zijn dan is het de vraag of de markt bereid is de vereiste transparantie te leveren. De mate van transparantie die vereist is om een deskundig oordeel over de autonomie van een systeem te vellen is zo groot (vrijwel volledig) dat concurrenten daardoor moeiteloos bij je bedrijfsgeheimen kunnen komen. Het overlaten van het ontwikkelen van autonome systemen aan de markt is daarom ook prohibitief voor het invoeren van verboden of verplichtingen. Een stelsel waarin de overheid/overheden autonome systemen ontwikkelen zou interessant zijn, maar lijkt onhaalbaar. Er blijft dan nog maar één mogelijkheid over. Een volledig moratorium voor autonome systemen. Een dergelijk moratorium zou moeten voortduren tot het moment dat het systeem niet langer autonoom is, dat wil zeggen door verdere ontwikkeling wel voorspelbaar is geworden voor de principalen. Voor een aantal (niet-deterministische of chaotische) systemen werkt het moratorium als een permanent verbod. Het verbieden van deze technologie of het voorschrijven van een beter gebruik zijn daarom geen oplossing. Het beter inrichten van de technologie is de overblijvende mogelijkheid.

Toestemming onder voorwaarden: ingebouwde juridische/ethische normen

Als autonome systemen in staat zijn om juridische of ruimer ethische gedragsnormen in hun beslissingsproces te verwerken dan kan een autonoom systeem zelf voor elke (op zich onvoorspelbare) beslissing, als deze zich feitelijk voordoet, de afwegingen maken die de principaal had behoren te maken.[19] Dit veronderstelt wel dat deze ethische gedragsnormen kunnen worden geformuleerd zonder dat alle mogelijke systeembeslissingen bekend zijn. Daarvoor is een universele ethiek vereist, omdat een praktische ethiek voorspelbaarheid vereist om niet ernstig onvolledig te zijn. De speculaties over en de ervaringen met een universele ethiek in computersystemen geven weinig hoop dat dit een vruchtbare benadering is.[20] U kent ongetwijfeld de speculaties van de natuurkundige en schrijver Isaac Asimov over het inbouwen van drie ethische wetten van de robotica in robots. Het is vrij gemakkelijk om aan te tonen dat dergelijke universele ethische normen niet werken. Om er maar een paar

17. In het geval van bedrijven geldt bovendien dat de (internationale) concurrentieverhoudingen dit onmogelijk maken, zelfs als de middelen er zouden zijn.
18. Zie bijvoorbeeld over het deskundigheidsniveau van rechters en rechtbankmedewerkers C.N.J. de Vey Mestdagh & T. van Zuijlen, 'Deskundigheidsbevordering ICT in de Rechtspraak', *Raad voor de Rechtspraak, Research Memorandum* Nummer 1 / 2018, https://www.rechtspraak.nl/SiteCollectionDocuments/RM-2018-1.pdf.
19. We moeten in ieder geval af van achterhaalde mogelijkheden als het verantwoordelijk maken van de maker of de eigenaar van het systeem. Het onvoorspelbare gedrag van autonome systemen levert al een onoverbrugbare kloof op met de mogelijke verantwoordelijkheid van de principaal. De maker of de eigenaar hebben nog minder inzicht in de omgeving waarin deze systemen worden ingezet dan de principaal.
20. Zie voor een recent overzicht van voorstellen voor AI ethics A.F.T Winfield, 'A Round Up of Robotics and AI ethics', Alan Winfield's Web Log, December 23, 2017, http://alanwinfield.blogspot.com/2017/12/a-round-up-of-robotics-and-ai-ethics.html.

te noemen: de robot moet de betekenis van alle gebruikte begrippen kennen (robot, mens, letsel, etc.), de robot moet weten dat hij een robot is en de robot moet alle consequenties van zijn handelen kennen (voorspelbaar zijn). De beperkte ervaringen die er zijn met het inbouwen van ethische normen in robots geven voorlopig ook weinig hoop.[21] Autonome robots in (zeer eenvoudige) kunstmatige werelden staan bijvoorbeeld na verloop van tijd stil (bereiken een evenwichtstoestand) als ze dezelfde ingebouwde ethisch normen zouden volgen.[22] Verder onderzoek zou moeten aantonen of het mogelijk is om dit probleem te vermijden door de verschillende robots licht verschillende ethische normen te geven of de robot een praktische ethische norm te laten formuleren op grond van hem bekende universele ethische normen en de praktische situatie waarin hij zich bevindt of in het geval van autonome systemen te eisen dat er slecht gebruik gemaakt mag worden van hybride systemen, waarbij het menselijke deel de ethische normen toepast. Deze mogelijkheden leveren evenzovele interessante politieke problemen op, maar zijn ook realistisch omdat deze meer lijken op de menselijke conditie.

De technische realisatie van deze mogelijke oplossing is overigens geen probleem. Ethische normen verschillen in kennistheoretisch opzicht niet van rechtsnormen en andere gedragsinstructies (inclusief machine instructies) en kunnen derhalve op dezelfde wijze worden gerepresenteerd in juridische kennissystemen. Hierover meer in een volgend artikel Calculo Ergo Sum (3). Maar zoals gezegd, ook deze technische oplossing (het inbouwen van ethische normen) faalt voorlopig.

5. Conclusies: moratorium, deskundigheidsbevordering, transparantie

Autonome netwerkcomputers leveren een juridisch/ethisch probleem op omdat hun gedrag voor hun principalen (opdrachtgevers) per definitie niet voorspelbaar is.[23] Dit niet voorspelbare gedrag wordt veroorzaakt doordat netwerkcomputers niet alleen hun eigen gegevens en kennis gebruiken, maar ook die van de (vele) andere computers waar ze mee verbonden zijn (de context waarin ze opereren). Dit kan grote schade veroorzaken zonder dat de rechtsgevolgen die aan deze schade ten grondslag liggen kunnen worden teruggedraaid. In de casus *Flash Crash* zou het terugdraaien van transacties ter waarde van vele miljarden dollars de beurs volledig ontwrichten. De principalen hebben geen schuld en zijn zich van te voren niet bewust van het risico. Er is sprake van een bloot rechtsfeit met rechtsgevolg, maar zonder aansprakelijkheid. De vraag is of de mens moet worden weerhouden van het gebruik van deze technologie of worden gedwongen tot beter gebruik ervan of – als dat niet mogelijk is – de technologie dan beter moet worden ingericht door het inbouwen van juridische of ruimer ethische normen. In het laatste geval moeten we accepteren dat het ethisch handelen van computers tenminste even feilbaar is als dat van de mensen die de ethische normen bij hen hebben ingevoerd.[24]

In het eerste geval is er sprake van een verbod. De kennis die vereist is om een dergelijk algemeen verbod voor de vele bestaande en nog te ontwikkelen specifieke autonome systemen te formuleren en te handhaven overstijgt op dit moment het kennisniveau van de overheid en de principalen ruimschoots. Het opleggen van verplichtingen die de mens dwingen tot beter gebruik zou op hetzelfde neerkomen. De autonome systemen moeten daarvoor eerst voorspelbaar worden gemaakt anders is nakoming van de verplichtingen niet mogelijk. Het beter inrichten van autonome systemen kan door het inbouwen van de nodige juridische of ruimer ethische normen in de systemen zelf. Deze kunnen dan de verantwoordelijkheid van de principaal overnemen door het nemen van ethische beslissingen op het moment dat de behoefte aan een concrete beslissing zich voordoet. Het inbouwen van ethische normen is technisch geen probleem, maar de toepassing ervan door autonome systemen leidt nog niet tot het gewenste resultaat. Systemen met dezelfde ethiek bereiken na enige tijd een evenwichtstoestand, dat wil zeggen: ze staan stil. Voor de realisatie van deze oplossing zijn ingrijpende politieke besluiten vereist. Bijvoorbeeld het loslaten van een level playing field (het toestaan van verschillende (ethische) normen voor de deelnemende systemen aan een markt) of het overlaten van het formuleren van praktische ethische normen gebaseerd op door de wetgever geformuleerde algemene ethische normen door de autonome systemen zelf of het stellen van de eis dat een mens (principaal) deel uitmaakt van elk autonoom systeem, waar-

21. Zie voor een recent overzicht van voorstellen voor AI ethics A.F.T Winfield, 'A Round Up of Robotics and AI ethics', Alan Winfield's Web Log, December 23, 2017, http://alanwinfield.blogspot.com/2017/12/a-round-up-of-robotics-and-ai-ethics.html De op deze webpagina genoemde projecten laten zien dat er met het inbouwen van complexere ethische normen nog weinig ervaring is.
22. A.F.T. Winfield, C. Blum, and W. Liu, 'Towards an ethical robot: internal models, consequences and ethical action selection', in *Advances in Autonomous Robotics Systems*, eds M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish, Springer 2014, p. 85–96; en C. Blum, A.F.T. Winfield and V.V. Hafner (2018), 'Simulation-Based Internal Models for Safer Robots' in Frontiers in Robotics and AI, 2018 (Online https://www.frontiersin.org/articles/10.3389/frobt.2017.00074/full).
23. Dit geldt overigens ook en zelfs in versterkte mate voor hun makers en eigenaars, die veelal de omgeving (context) waarin autonome systemen opereren nog slechter kennen dan de principaal.
24. Ik wil nog niet speculeren over computersystemen die hun ingebouwde ethische normen op grond van hun ervaringen gaan veranderen en uitbreiden. Dit is technisch wel mogelijk, maar zou het geschetste probleem van onvoorspelbaarheid nog vergroten.

door de bestuurlijke en economische betekenis van het inzetten van autonome systemen waarschijnlijk vrijwel volledig verdampt.

De enige haalbare oplossing op korte termijn is daarom het instellen van een moratorium, desnoods voor de gevaarlijkste categorieën van autonome systemen,[25] en het inzetten op deskundigheidsbevordering bij de overheid, het parlement (dat uiteindelijk over de toelaatbaarheid moet beslissen) en de potentiële principalen.[26] Om niet onder het moratorium te vallen zouden principalen eerst moeten aantonen dat hun systeem wat betreft het potentiële risico wel voorspelbaar en minder gevaarlijk is.[27] Voor minder risicovolle categorieën kan het voorschrijven van een *verzekering* en vrijwel volledige *transparantie*, waardoor het reguliere maatschappelijke *kwaliteit bewakingssysteem* (journalisten, wetenschappers, juristen, overheid) zijn werk kan doen, werken.

25. Bijvoorbeeld handelssystemen die tot grote financiële schade kunnen leiden; vervoerssystemen en medische systemen die tot letselschade kunnen leiden; bestuurlijke systemen die tot onrechtmatig overheidshandelen kunnen leiden.
26. Als wij kiezen voor een toekomst met autonome systemen, en dat is vrijwel onvermijdelijk, dan betreft dit de gehele bevolking en zal er nieuw verplicht onderwijs 'werken met autonome systemen' voor iedereen moeten worden ingevoerd.
27. Niet het volledige gedrag is dan voorspelbaar, maar wel de (ernst van de) potentiële consequenties van het gedrag.