

De Opmars van Cognitieve Agenten in Kunstmatige Intelligentie

Stel je hebt geld nodig en er zijn twee dingen die je kan doen om aan het nodige geld te komen: *werken* of *meespelen met de lotto*. Waarop zou je een dergelijke beslissing dan baseren? Hoe zou je die beslissing beschrijven en welke concepten zou je gebruiken? Deze vragen zijn niet alleen relevant voor het bestuderen van menselijke beslissingen en het menselijke gedrag. Ze zijn ook van belang bij het bestuderen, beschrijven, en eventueel modelleren van het gedrag en beslissingen van andere organismen, systemen of organisaties.

Het autonoom nemen van beslissingen wordt over het algemeen gezien als een complex proces dat intelligentie vereist. De bestudering ervan houdt onderzoekers op het gebied van de kunstmatige intelligentie in het algemeen en agent technologie in het bijzonder dan ook al geruime tijd bezig. De verwachting is dat exacte beslistheorieën het realiseren van toekomstige generaties intelligente computersystemen een stap dichterbij brengt. Net als andere deelgebieden van kunstmatige intelligentie heeft het onderzoek naar het autonome gedrag van intelligente agenten zich laten motiveren door studies uit verschillende wetenschappelijke disciplines, zoals de neurologie, psychologie, filosofie, economie, logica, en wiskunde. Ieder van deze disciplines hanteert andere begrippen om het gedrag en het beslisproces te beschrijven. Bijvoorbeeld, neurologen proberen het gedrag in termen van hersenactiviteiten te beschrijven, een deel van de psychologen vindt aangeboren of aangeleerde eigenschappen relevant, terwijl andere psychologen (behavioristen) gedrag bij voorkeur beschrijven in termen van externe stimuli en respons.

Een onderzoekstraditie die grote invloed heeft op de ontwikkeling van autonome software agenten, is de rationele beslistheorie. In deze traditie wordt het gedrag van een agent gezien als de reeks van beslissingen die die agent achtereenvolgens neemt. Het nemen van een beslissing is gedefinieerd als het kiezen (prefereren) van een actie uit een verzameling beschikbare acties. Zo is het prefereren van bijvoorbeeld de actie *werken* (boven *meespelen met de lotto*) een beslissing. De vraag is dan hoe deze preferentie beschreven kan worden en het gebruik van welke concepten hierbij gepast is.

Rationele Beslistheorie

Eén van de grondleggers van de rationele beslistheorie is de gerenommeerde statisticus Leonard Savage (1917-1971). Hij liet zien dat de beslissingen en het gedrag van een rationele agent gepresenteerd kunnen worden in termen van twee numerieke agent-afhankelijke functies en een algemene agent-onafhankelijke beslisregel. De eerste van de twee functies betreft de waarschijnlijkheid van de mogelijke uitkomsten van acties, d.w.z. hoe waarschijnlijk een agent een situatie als de uitkomst van een actie acht. De tweede functie specificiert het nut die de mogelijke uitkomsten hebben voor een agent. De agent-onafhankelijke beslisregel levert de actie die het verwachte nut voor de agent maximaliseert. Een agent is dan rationeel als zijn (geprefereerde) gekozen acties volgens deze beslisregel bepaald worden. Merk op dat een agent een willekeurige actie kan prefereren, maar als de geprefereerde actie niet het maximale verwachte nut oplevert, is

de beslissing niet rationeel volgens de theorie van Savage. Hiermee kan aldus een onderscheid gemaakt worden tussen rationele agenten en irrationele agenten.

De rationele beslistheorie van Savage kan aan de hand van een voorbeeld geïllustreerd worden. In het bovengenoemde scenario beschouwen we alleen twee mogelijke uitkomsten: *rijk* en *arm*. De waarschijnlijkheidsfunctie van een agent kan bijvoorbeeld de getallen 0.9 en 0.1 toekennen aan respectievelijk de toestanden *rijk* en *arm* als gevolg van de actie *werken* (de agent denkt dat werken geld opbrengt) en de getallen 0.2 en 0.8 aan dezelfde toestanden als gevolg van de actie *meespelen met de lotto* (de agent denkt dus niet dat het meespelen met de lotto lucratief is). De nutsfunctie kan dan getal 5 toekennen aan de toestand *rijk* en getal 1 aan de toestand *arm*. In dat geval vindt de agent *rijk*-zijn vijf keer zo belangrijk als *arm*-zijn. De actie *werken* kan dus met de waarschijnlijkheid 0.9 het nut 5 (de toestand *rijk*) en met de waarschijnlijkheid 0.1 het nut 1 opleveren. Het verwachte nut van deze actie is daarom $(0.9 * 5) + (0.1 * 1) = 4.6$ en het verwachte nut van de actie *meespelen met de lotto* is $(0.2 * 5) + (0.8 * 1) = 1.8$. De beslisregel schrijft voor dat de geprefereerde actie het maximaal verwachte nut moet opleveren, hetgeen voor dit voorbeeld betekent dat de actie *werken* geprefereerd zou moeten worden.

Kwalitatieve Beslistheorie

Een punt van kritiek dat van toepassing is op deze beslistheorie is dat het ondoenlijk en onrealistisch zou zijn om voor concrete agenten deze numerieke waarschijnlijkheids- en nutsfuncties te bepalen. Als reactie hierop zijn er verschillende kwalitatieve versies van beslistheorie geformuleerd, onder andere door computerwetenschappers als Judea Pearl en Craig Boutilier. Kwalitatieve beslistheorie onderscheidt zich doordat de twee numerieke functies worden vervangen door een aantal mentalistische concepten die afkomstig zijn uit *folk psychology*, een stroming waarbij er vanuit wordt gegaan dat het gedrag van mensen uitgelegd en voorspeld kan worden in termen van alledaagse concepten als geloof, wens, angst, en pijn. In de kwalitatieve beslistheorie wordt de waarschijnlijkheidsfunctie vervangen door geloofstoestanden, en de nutsfunctie door wenstoestanden. De beslissingen van een rationele agent worden beschreven in termen van haar mentale toestand als vastgelegd door haar geloof en haar wensen. In dit raamwerk wordt er verondersteld dat een agent op basis van haar geloof en wensen haar doelen vaststelt alvorens zij een reeks acties, een plan, genereert om de vastgestelde doelen te bereiken.

In ons voorbeeld heeft de agent de wens om rijk te worden. Stel dat de agent ook gelooft dat zij rijk kan worden, omdat zij bijvoorbeeld de actie *werken* tot haar beschikking heeft. In dat geval kan de agent de wens rijk te worden tot doel stellen en vervolgens een plan vormen om dit doel te bereiken, bijvoorbeeld door de actie *werken* te kiezen. Stel nu dat de agent niet in de positie is te *werken* maar alleen mee kan spelen in de *lotto*. Dan heeft de agent de wens rijk te worden zonder dat zij gelooft dat zij rijk kan worden. Zij gelooft tenslotte niet dat meespelen in de lotto tot grote rijkdom leidt. Merk op dat de agent in beide gevallen *rijkdom* als doel kan hebben. In het eerste geval gelooft zij dat zij haar

doel kan bereiken terwijl dat in het tweede geval niet zo is. Toch zou je het gedrag van de agent in het eerste geval realistischer kunnen noemen, omdat zij dan alleen doelen heeft die zij gelooft te kunnen realiseren. Om dit onderscheid in het beslis- of gedragmodel van een agent te integreren wordt het concept *agent type* geïntroduceerd. Een *realistische agent* heeft alleen doelen die zij gelooft te kunnen realiseren terwijl een *wishful thinking agent* doelen kan hebben waarvan zij niet gelooft ze te kunnen bereiken. Het is belangrijk dit onderscheid niet te verwarren met dat tussen rationele en niet rationele agenten dat ten grondslag ligt aan het model van Savage. Het gaat hier immers niet om het maximale verwachte nut van de acties.

De kwalitatieve versies van de beslistheorie zijn voornamelijk populair onder logici omdat mentale concepten als geloof en wens zich goed lenen tot formalisatie. De logische theorieën die cognitieve concepten zoals geloof en wens formaliseren stellen je in staat de logische consequenties van een bepaalde geloven en wensen af te leiden. Stel bijvoorbeeld dat je gelooft dat je familie rijk is en dat jezelf rijk bent als je een rijke familie hebt. Uit deze twee geloofsuitingen kun je volgens de meeste gangbare theorieën over geloof afleiden dat je gelooft dat je rijk bent. Een centraal thema is dan de interactie die tussen de verschillende cognitieve concepten bestaat in kaart te brengen. Het is bijvoorbeeld de vraag of wat je gelooft een gevolg te zijn van een wens ook als een wens gezien moet worden. Dergelijke vormen van interactie tussen geloof en wens kunnen geïllustreerd worden aan de hand van het volgende voorbeeld. Stel dat je wenst naar de tandarts te gaan en dat je bovendien gelooft dat als je naar de tandarts gaat, je ook pijn zult hebben. Is het dan ook zo dat je wenst pijn te hebben? Dergelijke kwesties spelen een belangrijke rol in de verschillende logische formalisering van geloof en wens.

Belief-Desire-Intention (BDI) Model

In het bovengenoemde werden geloof en wens beschouwd als primaire concepten in termen waarvan het gedrag van agenten beschreven en gemodelleerd kan worden. De filosoof Bratman wees echter op een nadeel van deze modellen. Volgens hem zou het definiëren van beslissingen op basis van de begrippen geloof en wens tot een instabiele notie van gedrag kunnen leiden. Een stabiel gedrag zou volgens hem ook afhankelijk zijn van de intenties van de betreffende agent. Er dienen verschillende vormen van intentie onderscheiden te worden, waaronder de intentie die gevormd wordt op basis van eerder gestelde doelen. Stel bijvoorbeeld dat de agent in ons voorbeeld, behalve *rijk*, ook *hoogleraar* wenst te worden. Stel verder dat zij gelooft dat zij met hard studeren *hoogleraar* zou kunnen worden, maar dat zij onmogelijk tegelijk *hoogleraar* en *rijk* kan zijn. Deze agent kan realistisch zijn ondanks het feit dat zij haar wens om een hoogleraar te worden tot doel stelt. Zij heeft immers een doel waarvan zij gelooft het te kunnen realiseren, namelijk hoogleraar te worden, arm of rijk. Het probleem is dat door dit doel het gedrag van de agent instabiel zou kunnen worden, aangezien de acties van de agent gericht zouden kunnen worden op het bewerkstelligen van twee tegenstrijdige doelen. Bratman stelt voor dat het gedrag van een agent stabiel kan zijn als die enkel wensen tot doel stelt die niet tegenstrijdig zijn met haar intenties. Dit model van het gedrag, waarin zowel geloof, wens en intentie een rol speelt, wordt vaak het Belief-Desire-Intention

(BDI) model genoemd. Een agent die in termen van een BDI model gedefinieerd is, wordt ook wel een BDI of cognitieve agent genoemd.

In het BDI model beperken de intenties welke wensen tot doel gesteld kunnen worden. De hierdoor verkregen stabiliteit zou ten koste kunnen gaan van de flexibiliteit van het gedrag. Stel dat de hierboven genoemde agent de intentie heeft gevormd hoogleraar te worden. Stel verder dat deze agent betrapt wordt op plagiaat waardoor het hem onmogelijk wordt ooit een leerstoel te bekleden. In zo een geval zou je willen toelaten dat de agent deze intentie opgeeft om bijvoorbeeld een carrière in de IT na te streven. Daarom onderscheidt Bratman een aantal agenttypen die ieder verschillend omgaan met hun intenties. Deze agenttypen bepalen onder welke voorwaarden een intentie opgegeven dient te worden. Deze condities worden ook wel *commitment strategieën* genoemd. Een dergelijke strategie zou bijvoorbeeld kunnen voorschrijven dat een intentie pas opgegeven kan worden als de agent gelooft dat het doel bereikt is. Deze strategie wordt ook de *blindly-minded commitment* strategie genoemd. Een andere strategie, die milder met de intenties omgaat, is de *single-minded commitment* strategie. Volgens deze strategie kan een intentie ook opgegeven worden wanneer de agent gelooft dat de intentie niet meer gerealiseerd kan worden. Dus, als de agent gelooft dat zij nooit meer een hoogleraar kan worden, dan kan zij deze intentie opgeven om vervolgens een carrière in de IT te maken. Tot slot, de *open-minded* strategie laat een agent zijn intenties opgeven wanneer de agent een wens tot doel stelt dat tegenstrijdig is met deze intenties. Voor onze agent betekent dit dat zij de intentie om hoogleraar te worden opgeeft om een IT carrière te maken terwijl zij hoogleraarschap wel mogelijk acht.

Norm en Verplichting

In het BDI model worden beslissingen en gedrag gedefinieerd in termen van de mentalistische concepten geloof, wens, en intentie. De doelen van een agent vloeien enkel voort uit de wensen van de agent, waarna de intenties gevormd worden. Een vraag is nu of de doelen van een agent alleen tot stand komen op basis van haar wensen. Of zijn er ook andere aspecten die bijdragen tot de totstandkoming van doelen. Het zou bijvoorbeeld betoogd kunnen worden dat verplichtingen en normen ook een belangrijke rol spelen bij het totstandkoming van doelen. Deze concepten zijn essentieel wanneer een agent zich in een sociale gemeenschap bevindt, waarin regels of normen gelden. Het deelnemen aan een sociale gemeenschap en het maken van onderlinge afspraken verplicht agenten tot het nemen van bepaalde beslissingen en het uitvoeren van bepaalde acties. Stel bijvoorbeeld dat je je partner beloofd hebt om samen op vakantie te gaan. Op basis van de bij deze belofte behorende verplichting zou je je ook tot doel moeten stellen op vakantie te gaan.

De verschillende kwalitatieve beslistheorieën en de BDI modellen lenen zich er evenwel goed toe uitgebreid te worden met dergelijke sociale concepten. Deze uitgebreide modellen zijn erg belangrijk voor agenten die tezamen een multi-agent systeem constitueren. Het idee is dan dat de doelen van een agent die ingebed is in een sociale omgeving tot stand komen op basis van de sociaal-cognitieve concepten van geloof,

wens, intentie, norm en verplichting. De toename van de concepten die de doelen van een agent beïnvloeden maakt dat de formele interactie tussen deze concepten veel complexer wordt. De vraag is hoe de sociaal-cognitieve concepten geformaliseerd kunnen worden zonder dat dit ten koste gaat van de inzichtelijkheid van het formele systeem in zijn geheel. Stel, bijvoorbeeld dat je rijk wenst te worden maar verplicht bent om op vakantie te gaan en dat je bovendien gelooft dat het rijk worden en op vakantie gaan niet samen mogelijk is. Welk doel zou je je in een dergelijke situatie moeten stellen? Leg je de prioriteit bij het rijk worden en laat je die vakantie voor wat hij is, of ga je op vakantie en geef je de wens om rijk te worden op? Op dit punt zijn er verschillende mogelijkheden, geen van welke op het eerste gezicht duidelijk beter is dan de andere. De oplossing lijkt weer te liggen in het onderscheiden van agent typen. Afhankelijk van hun type genereren agenten het doel om rijk te worden of het doel hun verplichtingen na te komen. Men zou bijvoorbeeld een *egoïstische* en een *sociale agent* kunnen onderscheiden. Zo zou een sociale agent zich tot doel stellen om haar verplichtingen na te komen, wanneer zij geconfronteerd wordt met tegenstrijdige wensen en verplichtingen, terwijl een egoïstische agent zich juist tot doel zou stellen haar wensen te realiseren. De verschillende typen kunnen bovendien gecombineerd worden, hetgeen leidt tot een groot aantal agenttypen zoals *sociaal realistisch* en *egoïstische wishful thinking*.

De Rol van Emoties

De vraag blijft natuurlijk bestaan of deze cognitieve concepten toereikend zijn om het gedrag van agenten uitputtend te beschrijven en te modelleren. Psychologen als Frijda en Arnold betwisten dit. Volgens hen zouden ook emoties een belangrijke rol spelen in het menselijk beslisproces en gedrag. Ook zijn er neurologische studies en observaties die aantonen dat mensen met specifieke emotionele stoornissen moeite hebben met het nemen van beslissingen of zelfs daartoe niet in staat zijn. Damasio concludeert dan dat de afwezigheid van emoties bij mensen tot een eindeloze afweging van acties kan leiden zodat het nemen van beslissingen bemoeilijkt wordt. Deze redenering lijkt te suggereren dat de emoties ook een belangrijke rol kunnen vervullen in de praktische beslismodellen en voor het ontwerp van software agenten. Welke rol emoties kunnen spelen bij het ontwerpen van de software agenten is een opkomend onderwerp in het huidige onderzoek. Een mogelijke rol van emoties bij het ontwerpen van software agenten is een functionele rol waarbij emoties beschouwd worden als labels die het beslisproces van een agent gedeeltelijk bepalen. Volgens deze benadering, geformuleerd door John-Jules Meyer, hoogleraar aan de Universiteit van Utrecht, is een specifieke emotie het effect van voorafgaande beslissingen. De (gegenereerde) emotie bepaalt vervolgens het beslisproces en daarmee het gedrag van de agent. Deze rol van emoties kan aan de hand van het volgende voorbeeld geïllustreerd worden. Stel dat alle pogingen van een agent tot het realiseren van het doel *rijk worden* tot mislukken gedoemd bleken te zijn. De agent zou hierdoor in de *boos* toestand terecht kunnen komen. Hierdoor zou zij nog meer en eenzijdiger gaan inspanssen het doel *rijk worden* te bereiken. Als gevolg hiervan zouden andere doelen en de bijbehorende plannen opgegeven kunnen worden.

Tot Slot

Er zijn nog altijd andere concepten denkbaar die de beslissingen en het gedrag van de agenten kunnen beïnvloeden, beschrijven en modelleren. Denk hierbij aan concepten als vertrouwen en macht. Vele uitbreidingen van BDI systemen zijn dan ook voorgesteld. Momenteel wordt er veel multidisciplinair onderzoek gedaan op dit gebied. Het theoretische belang van deze modellen is dat ze inzicht geven in menselijk gedrag. Maar ze zijn vooral populair vanwege hun toepassingen bij het specificeren, verifiëren en implementeren van intelligente computer systemen. Deze autonome computer systemen hebben applicaties in sociale simulatie, in e-commerce, and als personal assistants. Voor ieder van deze toepassingen is het essentieel dat de computer systemen, die geacht worden met elkaar interacteren, zelfstandig beslissingen kunnen nemen en aldus de belangen van hun ontwerpers of opdrachtgevers te behartigen. De verwachting is dan ook dat toekomstige generaties intelligente systemen direct ontworpen en geïmplementeerd zullen worden op basis van sociaal-cognitive modellen.