Department of Computer Science,
University of Otago

Technical Report OUCS-2002-07

Het zeven-kaartenprobleem: kennislogica en multi-agentsystemen (The seven cards problem: epistemic logic and multiagent systems)

Author:
Hans van Ditmarsch
Department of Computer Science, University of Otago

Status: to appear in
"Nieuw Archief voor Wiskunde" (New Archive of Mathematics)

Department of Computer Science,
University of Otago, PO Box 56, Dunedin, Otago, New Zealand

http://www.cs.otago.ac.nz/trseries/

Het zeven-kaartenprobleem
Kennislogica en multi-agentsystemen

Hans van Ditmarsch[*†]

Uit zeven kaarten $0, 1, 2, 3, 4, 5, 6$ *trekken Anna en Bert er ieder drie, en Cees krijgt de resterende kaart. Al het voorgaande is gemeenschappelijke kennis. Hoe kunnen Anna en Bert elkaar openlijk van hun kaarten op de hoogte brengen, zonder dat Cees van een van hun kaarten leert wie die kaart heeft?*

Dit 'zeven-kaartenprobleem' werd gesteld tijdens de Wiskunde-Olympiade in Moskou in 2000. Inmiddels is het ook verschenen in de Mathematical Intelligencer [12] en in Natuur&Techniek [18, 21]. Soortgelijke problemen circuleren onder coderingstheoretici [4]. Er zijn meerdere goede oplossingen. De Moskouse commissie werd destijds ook geconfronteerd met een aantal 'foute oplossingen' die op het eerste gezicht eigenlijk niet zo onredelijk waren. Om deze te analyseren blijkt een vrij recente tak van logica de helpende hand te bieden, namelijk *kennislogica* (epistemic logic). Kennislogica wordt momenteel veel gebruikt voor het specificeren van multi-agentsystemen [3]. We analyseren eerst een aantal 'foute oplossingen' in termen van wat spelers van elkaar weten. Daarna introduceren we de kennislogica die e.e.a. formaliseert. We besluiten met oplossingen van het probleem, een kort historisch overzicht, en enige andere toepassingen van de kennislogica.

1 Schijnbare oplossingen

De voorwaarden voor de oplossing van het probleem zijn dat:

- Anna kent Berts kaarten, ($aweetb$)
- Bert kent Anna's kaarten, ($bweeta$)
- Cees kent geen van Anna's of Berts kaarten. ($cweetniet$)

Neem nu verder aan dat de hand van Anna $\{0, 1, 2\}$ is (schrijf 012), dat de hand van Bert 345 is, en dat Cees kaart 6 heeft. De 'foute oplossing van de Olympiade' genoemd in [12] is:

[*]Computer Science, University of Otago, New Zealand, `hans@cs.otago.ac.nz`

[†]Met dank aan Barteld Kooi en Ger Koole voor hun detailcommentaar.

Anna zegt: "Als jij 0 *niet hebt, dan heb ik* 012*." en Bert zegt: "Als jij* 3 *niet hebt, dan heb ik* 345*."* (i)

Waarom 'lijkt' dit een oplossing? Stelt u zich het standpunt voor van een 'insider' Dirk, die in ieders kaarten kan kijken:

Dirk zegt: "Als Bert 0 *niet heeft, dan heeft Anna* 012*, en als Anna* 3 *niet heeft dan heeft Bert* 345*."* (ii)

We bereiken dan een informatietoestand waarin *aweetb*, *bweeta* en *cweetniet* allemaal het geval zijn; *cweetniet* geldt, omdat na Dirk's uitspraak Anna nog een van de handen $012, 345$ (en $134, 135, 234, 235$ en 123) kan hebben, dus Cees kan geen enkele kaart uitsluiten. Waarom verschillen (i) en (ii) van betekenis, en is (i) geen oplossing van het probleem? Dit is omdat Anna en Bert *minder* weten dan de virtuele Dirk. Hun uitspraken zijn daarom juist *informatiever*:

Cees redeneert als volgt: "Anna kent in deze spelsituatie alleen haar eigen kaarten. Ze kan dus niet weten *of Bert kaart 0 heeft, tenzij ze die zelf vasthoudt. Anna kan dus alleen* naar waarheid *haar uitspraak doen, als zij zelf* 012 *heeft."*

Het ligt voor de hand dat de uitspraken van de spelers gebaseerd moeten zijn op de informatie die hen ter beschikking staat. Maar hiermee zijn we er nog niet, gezien het volgende:

Anna zegt: "Ik heb kaart 6 niet" en Bert zegt: "Ik heb ook kaart 6 niet." (iii)

Na Anna's uitspraak, ook onder de aanname dat ze weet wat ze zegt, weet Cees geen van haar kaarten: *cweetniet* geldt. Echter, *Anna weet niet dat cweetniet*! Dit komt omdat na een andere uitvoering van het protocol dat Anna kennelijk hanteert, Cees wél een van haar kaarten geleerd had. Ze had immers ook kunnen zeggen "Ik heb kaart 4 niet." Om Anna zeker van haar zaak te laten zijn, moeten we dus kennelijk ook verlangen dat na haar uitspraak *Anna weet dat cweetniet geldt.* Ook dit blijkt niet genoeg, gezien het volgende:

Anna zegt: "Ik heb 012 *of ik heb geen van die kaarten." en Bert zegt: "Ik heb* 345 *of ik heb geen van die kaarten."* (iv)

We bereiken nu een informatietoestand waarin *cweetniet* geldt. Ook weet Anna dat *cweetniet* geldt: het protocol dat ze hanteert heeft maar één uitvoering. Echter, *Cees weet niet dat Anna weet dat cweetniet*! De situatie blijkt opnieuw informatief voor Cees:

Cees redeneert als volgt: "Stel dat Anna kaart 0 **niet** *heeft. Dan weet Anna niet of ik 0 heb. Stel ik had 0. Dan was Anna's hand niet 012. Dan had Anna, volgens haar eigen uitspraak, geen van die kaarten. Ik had dan geleerd dat Anna 1 en 2 niet heeft. Dus dan had Anna niet gezegd wat ze zei. Maar ze zei het wel. Dus heeft ze kaart 0* **wel**. *Dus, volgens haar eigen uitspraak, heeft ze* 012*."*

Met andere woorden: Cees leert Anna's kaarten uit de aanname dat Anna alleen uitspraken doet waaruit Cees haar kaarten niet leert. En *zonder* die aanname had Cees Anna's kaarten *niet* geleerd. De tegenspraak is maar schijn, want de tegengestelde beweringen gelden in *verschillende* informatietoestanden. Het scenario lijkt op dat in het bekende 'modderige-kinderen'-probleem, waarin de modderige kinderen leren dat ze modderig zijn, nadat publiek bekend wordt dat geen van de kinderen weet of ze modderig zijn [3].

U vraagt zich inmiddels natuurlijk af waar dit stopt. Het stopt bij *gemeenschappelijke kennis* van de oplossingscriteria. Een bewering is gemeenschappelijk bekend als (en alleen als) deze het geval is en iedereen het weet dat het gemeenschappelijk bekend is (preciezer: als het kleinste dekpunt – 'least fixed point' – van deze operatie). Hieruit volgt dat Anna weet dat Cees weet dat Anna weet dat Bert het weet, etc.

Een eindig aantal uitspraken van Anna en Bert is een oplossing voor het zeven-kaartenprobleem, als na elk van deze *cweetniet* gemeenschappelijk bekend is en na de laatste eveneens *aweetb* en *bweeta* gemeenschappelijk bekend zijn.

In de context van ons probleem is een bewering gemeenschappelijk bekend als deze bekend is aan een virtuele *buitenstaander* ('outsider') Erna, die zelf geen kaarten vasthoudt maar alles hoort wat er gezegd wordt. Met nog andere woorden: als de bewering geldt onafhankelijk van de werkelijke kaartverdeling.

Dit vraagt natuurlijk allemaal om verdere precisering. Om de lezer te intrigeren, presenteren we alvast, zonder verdere toelichting, één oplossing van het probleem:

> *Anna zegt: "Ik hernoem kaart 0 tot 6, 6 tot 7, 1 tot 4, 4 tot 1, en de rest blijft hetzelfde. De som van mijn hand kaarten is nu 12." en Bert zegt: "Cees heeft kaart 6."* (v)

2 Kennislogica voor multi-agentsystemen

De in het voorgaande geïntroduceerde noties zoals 'weten dat' en 'informatietoestand' gaan we nu formaliseren. We beginnen met een nog eenvoudiger voorbeeld dan 'zeven kaarten': er zijn nu slechts drie kaarten 0, 1 en 2, waarvan Anna (a), Bert (b) en Cees (c) er ieder één vasthouden. Neem aan dat Anna kaart 0 heeft, Bert kaart 1, en Cees kaart 2. Noteer dit als 012. Er zijn zes mogelijke kaartverdelingen, nl. 012, 021, 102, 120, 201 en 210. Twee kaartverdelingen zijn voor een speler niet van elkaar te onderscheiden, als deze in beide dezelfde kaart heeft. Deze equivalentierelatie induceert een partitie op de verzameling kaartverdelingen. Bijvoorbeeld voor Anna is de partitie: $\{\{012, 021\}, \{102, 120\}, \{201, 210\}\}$. Met feiten identificeren we een deelverzameling van de verzameling kaartverdelingen. Bijvoorbeeld met 'Anna heeft kaart 0' correspondeert $\{012, 021\}$. De informatietoestand waarin iedere speler alleen zijn eigen kaarten kent en de kaartverdeling 012 is, kunnen we formeel representeren als een relationele structuur: deze bestaat uit het domein van de (zes) mogelijke kaartverdelingen, drie equivalenties daarop (corresponderend met de kennis van de spelers), negen deelverzamelingen ervan (corresponderend met

feiten), en een 'speciaal object' (de werkelijke kaartverdeling, in dit geval 012). Zie Figuur 1. We hebben nu alleen nog een formele taal nodig die beweringen zoals 'Anna weet niet dat Bert kaart 1 heeft' kan interpreteren op deze structuur. Deze introduceren we in zijn algemeenheid:

Definitie[Kennistoestand] Gegeven is een verzameling agents (actoren) N en een verzameling atomen (feiten) P. Een *mogelijke-werelden model* M (kennismodel, Kripke-model) is een drietal $M = \langle W, \sim, V \rangle$. *Domein* W is een verzameling abstracte objecten genaamd *werelden* of feitelijke toestanden. *Toegankelijkheid* $\sim$ is een functie van agents $n \in N$ naar equivalentierelaties $\sim_n \subseteq W \times W$. *Waardering* V is een functie van atomen $p \in P$ naar deelverzamelingen $V_p \subseteq W$. Een paar (M, w) bestaande uit een model M en een wereld $w \in \mathcal{D}(M)$ is een *informatietoestand* of *kennistoestand* (M, w).

Voor het interpreteren van gemeenschappelijke kennis, hierna, gebruiken we de reflexieve en transitieve afsluiting $\sim_N$ van alle equivalentierelaties: $\sim_N := (\bigcup_{n \in N} \sim_n)^*$.

Definitie[Taal van de kennislogica] Gegeven is een verzameling agents N en een verzameling atomen P. Kennislogica (met 'common knowledge' en 'updates') $\mathcal{L}_N(P)$ is de kleinste verzameling die alle atomen $p \in P$ bevat en voor alle φ, ψ die er reeds deel van uitmaken tevens $\neg\varphi, (\varphi \wedge \psi), K_n\varphi, C\varphi, [\psi]\varphi$.

Formule $\neg\varphi$ staat voor 'niet φ', $(\varphi \wedge \psi)$ staat voor 'φ en ψ', $K_n\varphi$ staat voor 'agent n weet dat φ', $C\varphi$ staat voor 'φ is gemeenschappelijk bekend' (bij groep N), $[\psi]\varphi$ staat voor 'na update met ψ geldt φ'. Behalve 'ψ is de updateformule' zeggen we ook '$[\psi]$ is de update'. Andere logische connectieven kunnen als afkortingen worden ingevoerd ($\rightarrow$ voor 'impliceert', $\vee$ voor 'of', $\leftrightarrow$ voor 'dan en slechts dan'), en desgewenst kan sequentie van of keuze tussen updates eveneens als afkorting worden ingevoerd. De updates zijn zogenaamde *publieke* en *waarheidsgetrouwe* updates: iedereen hoort wat er gezegd wordt en liegen is verboden. Andere vormen zijn ook voorstelbaar, maar vergen een complexere logica.

Definitie[Semantiek] De interpretatie van een formule $\varphi \in \mathcal{L}_N(P)$ in een wereld $w \in M = \langle W, \sim, V \rangle$ is als volgt inductief gedefinieerd:

$$
\begin{array}{lll}
M, w \models p & \text{iff} & w \in V_p \\
M, w \models \varphi \wedge \psi & \text{iff} & M, w \models \varphi \text{ en } M, w \models \psi \\
M, w \models \neg\varphi & \text{iff} & \text{niet } (M, w \models \varphi) \\
M, w \models K_n\varphi & \text{iff} & \forall v \in W : v \sim_n w \Rightarrow M, v \models \varphi \\
M, w \models C\varphi & \text{iff} & \forall v \in W : v \sim_N w \Rightarrow M, v \models \varphi \\
M, w \models [\psi]\varphi & \text{iff} & M, w \models \psi \Rightarrow M_\psi, w \models \varphi
\end{array}
$$

M_ψ is de restrictie van model M, inclusief equivalenties, tot de werelden waar ψ geldt, d.w.z. $\mathcal{D}(M_\psi) := \{w \in \mathcal{D}(M) \mid M, w \models \psi\}$.

Voor $M, w \models \varphi$ lezen we 'In wereld w van model M geldt formule φ' of 'In wereld w van model M is φ waar'.

Behalve logica's van kennis zijn er ook logica's van geloof, waarin wat je 'weet' niet noodzakelijk het geval is. Je kunt geloven dat φ, terwijl φ niet

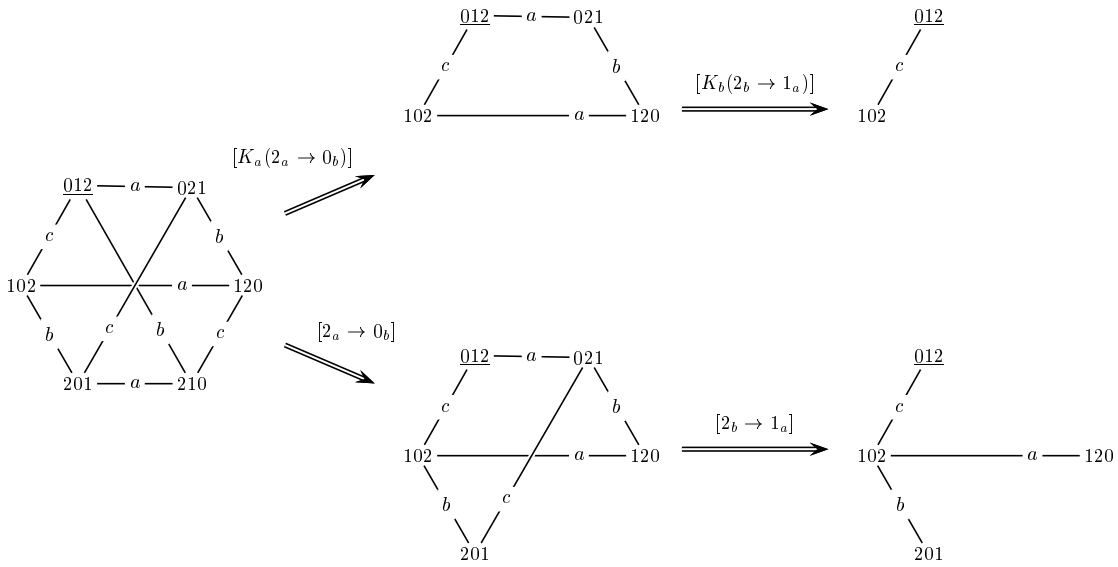

Figure 1: Updates in de informatietoestand $(Hexa, 012)$. De actuele toestanden zijn onderlijnd. Toestanden in dezelfde n-equivalentieklasse zijn verbonden, en de verbinding is gelabeld met die speler n.

waar is. Dit betekent dat de 'werkelijke toestand' van de wereld voor jou niet voorstelbaar ('toegankelijk') is, met andere woorden: de in definitie 2 met de operator K_n corresponderende toegankelijkheidsrelatie $\sim_n$ is niet reflexief, dus geen equivalentierelatie.

Het zal duidelijk zijn dat de hiervoor beschreven structuur voor kaartverdeling 012 een informatietoestand is in de hier gedefinieerde zin. Het model noemen we $Hexa$, het betreft dus de informatietoestand $(Hexa, 012)$. In dit geval zijn de werelden kaartverdelingen en atomen / feiten q_n drukken uit dat kaart q in handen is van speler n. We kunnen nu de geldigheid van allerlei beweringen over deze kennistoestand uitrekenen:

'Anna houdt kaart 0 vast in kaartverdeling 012' correspondeert met

$$Hexa, 012 \models 0_a$$

en dit geldt omdat $012 \in V_{0_a} = \{012, 021\}$.

'Anna weet dat ze kaart 0 heeft' correspondeert met

$$Hexa, 012 \models K_a 0_a$$

en dit geldt omdat er een een a-verbinding is tussen 012 en 021, en omdat zowel $Hexa, 012 \models 0_a$ als $Hexa, 021 \models 0_a$. Merk op dat er kennistoestanden voorstelbaar zijn waarin Anna wel kaart 0 heeft maar dit nog niet weet, bijvoorbeeld

als de kaarten al wel verdeeld zijn over de spelers maar deze de kaarten nog niet opgepakt hebben van de tafel.

'Anna kan zich voorstellen dat Bert zich kan voorstellen dat zij kaart 1 heeft' (terwijl ze in werkelijkheid kaart 0 heeft) correspondeert met

$$Hexa, 012 \models \neg K_a \neg \neg K_b \neg 1_a$$

en dit geldt omdat 012—a—021 en 021—b—120 en $Hexa, 120 \models 1_a$ (Anna heeft kaart 1 in kaartverdeling 120). 'Zich kunnen voorstellen dat' is de duale van 'weten dat'. 'Agent n kan zich voorstellen dat φ' is gedefinieerd als $\neg K_n \neg \varphi$.

'Het is gemeenschappelijk bekend dat Anna haar eigen kaart kent' correspondeert met

$$Hexa, 012 \models C(K_a 0_a \vee K_a 1_a \vee K_a 2_a)$$

en dit geldt omdat alle kaartverdelingen in Hexa met 012 verbonden zijn, en omdat een van de disjuncten dan altijd geldt.

Beweringen over het gevolg van updates kunnen we ook berekenen. Als Anna tegen Bert zegt: "Als ik kaart 2 heb dan heb jij kaart 0" dan komt dit overeen met update $[K_a(2_a \to 0_b)]$. Bewering $2_a \to 0_b$ is alleen onwaar als Anna 2 heeft en Bert niet 0, d.w.z. als Bert 1 heeft: kaartverdeling 210. Maar dit kan ze alleen *weten* als ze zelf niet 2 heeft, dus $K_a(2_a \to 0_b)$ is overal waar behalve in (Anna's equivalentieklasse) 201 en 210. Zie Figuur 1.

'Nadat Anna dit gezegd heeft is gemeenschappelijk bekend dat Anna kaart 2 niet heeft' correspondeert met

$$Hexa, 012 \models [K_a(2_a \to 0_b)] C \neg 2_a$$

en bewering $C\neg 2_a$ is eenvoudig te verifiëren in het uit de update resulterende informatietoestand.

Met de insider Dirk correspondeert een (virtuele) agent d wiens toegankelijkheid op het model de identiteit is ($\sim_d := \{(x, x) \mid x \in \mathcal{D}(Hexa)\}$). Met de outsider Erna correspondeert een (virtuele) agent e wiens toegankelijkheid op het model de universele relatie is ($\sim_e := \mathcal{D}(Hexa) \times \mathcal{D}(Hexa)$). De insider Dirk 'weet alles wat het geval is', met andere woorden $K_d\varphi \leftrightarrow \varphi$ geldt. De outsider Erna 'weet alles wat openbaar is' met andere woorden $K_e\varphi \leftrightarrow C\varphi$ geldt. Als $C\varphi$ geldt dan is φ geldig op het hele model, of met nog andere woorden: voor alle kaartverdelingen $d \in \mathcal{D}(Hexa)$ geldt $Hexa, d \models \varphi$.

Stel dat niet Anna maar insider Dirk had gezegd "Als Anna kaart 2 heeft dan heeft Bert kaart 0." Dit komt overeen met de update $[K_d(2_a \to 0_b)]$ die dus hetzelfde effect heeft als $[2_a \to 0_b]$. Nu geldt de update formule overal behalve in kaartverdeling 210. Zie opnieuw Figuur 1.

De analyse van 'foute oplossingen' verduidelijken we aan updates in de informatietoestand $(Hexa, 012)$. We willen dat:

$K_a 1_b$	Anna weet dat Bert 1 heeft
$K_b 0_a$	Bert weet dat Anna 0 heeft
$\neg K_c 0_a \wedge \neg K_c 0_b \wedge \neg K_c 1_a \wedge \neg K_c 1_b$	Cees weet niet wie van Anna of Bert 0 of 1 heeft

Maar in feite moet dit natuurlijk onafhankelijk van de werkelijke kaartverdeling 012 geformuleerd worden en krijgen we dus:

$$\begin{array}{ll} \mathsf{aweetb} & \bigwedge_{q=0,1,2}(q_b \to K_a q_b) \\ \mathsf{bweeta} & \bigwedge_{q=0,1,2}(q_a \to K_b q_a) \\ \mathsf{cweetniet} & \bigwedge_{q=0,1,2} \bigwedge_{n=a,b}(q_n \to \neg K_c q_n) \end{array}$$

Na de sequentie $[2_a \to 0_b][2_b \to 1_a]$ van twee updates ontstaat een informatietoestand waarin Anna en Bert elkaars kaart weten maar Cees dit niet van hen weet. Deze informatietoestand is weinig stabiel (er is geen gemeenschappelijke kennis), want als Anna nu zou zeggen "Ik weet Berts kaart" dan resulteert deze update $[K_a(K_a 0_b \vee K_a 1_b \vee K_a 2_b)]$ in het (niet gevisualiseerde) model dat bestaat uit niet verbonden werelden 012 en 201. Daarin is de kaartverdeling 012 (alsnog) gemeenschappelijk bekend.

Na de sequentie $[K_a(2_a \to 0_b)][K_b(2_b \to 1_a)]$ van twee updates ontstaat een informatietoestand waarin $C(aweetb \wedge bweeta \wedge cweetniet)$ geldt. Helaas geldt na uitsluitend de update $[K_a(2_a \to 0_b)]$ wel $cweetniet$ maar niet $Ccweetniet$. Daarom is dit geen oplossing van het 'drie-kaartenprobleem'. Als Anna en Bert beide hadden gezegd "Ik heb kaart 2 niet" had dit tot dezelfde informatietoestand geleid. De situatie is dus analoog aan die in voorbeeld (*iii*).

Overigens is hier interessant om op te merken dat voor het geval van 'e'en kaart per speler geen oplossing bestaat (wat niet lastig is in te zien binnen de context van epistemische logica, maar zie ook [4] voor een bewijs van zes pagina's waarvan dit een karakteristiek geval is).

U kunt dit soort publieke updates in het model $Hexa$ ook online zelf uitvoeren, op de webpagina `http://www.science.uva.nl/projects/opencollege/cognitie/-hexagon/` [13].

We vatten de resultaten nog even samen, waarbij we over wat technische hindernissen en impliciete generalisaties zonder verdere uitleg heenlopen (voor details, zie [20]). Als een speler n in een kaartenprobleem een uitspraak φ doet, dan is dit geen update $[\varphi]$ en ook geen update $[K_n\varphi]$ en ook geen update $[K_n\varphi \wedge [K_n\varphi]cweetniet]$, maar een update $[K_n\varphi \wedge [K_n\varphi]Ccweetniet]$. We noemen dit een veilige kennis-update. Zo'n veilige kennis-update is gelijk aan de sequentie van twee updates $[K_n\varphi][Ccweetniet]$. Na het uitvoeren hiervan geldt (nog steeds) $Ccweetniet$. Dit is niet triviaal, want we hebben een voorbeeld gezien (*iv*) van een informatietoestand waarin $[K_n\varphi][cweetniet]\neg cweetniet$ geldt (m.a.w. $cweetniet$ geldt na update met $[K_n\varphi]$ en $\neg cweetniet$ geldt na update met $[K_n\varphi][cweetniet]$). Een oplossing voor het kaartenprobleem is een sequentie van veilige kennis-updates waarna tevens $C(aweetb \wedge bweeta)$ geldt.

3 Oplossingen van het zeven-kaartenprobleem

Omdat updates restricties zijn op de verzameling mogelijke kaartverdelingen, zal duidelijk zijn dat iedere uitspraak van een speler equivalent is met keuze tussen alternatieve kaartverdelingen. Tevens is te bewijzen dat een uitspraak zelfs equivalent is met een aantal alternatieven voor de hand kaarten van die

speler (in het model komt dit overeen met een vereniging van equivalentieklassen kaartverdelingen voor die speler). Dit maakt het mogelijk om systematisch naar oplossingen te zoeken. Alle mij bekende oplossingen voor het zeven-kaartenprobleem bestaan uit één uitspraak van Anna en één van Bert (of zijn daartoe te reduceren). Daarom is na die van Anna reeds gemeenschappelijk bekend dat Bert haar kaarten kent. Dus weet Bert de kaart van Cees en kan zijn uitspraak die kennis openbaar maken. We laten Berts deel dus verder achterwege. Voor de aangenomen kaartverdeling waarin Anna 012 heeft, Bert 345 en Cees 6, is dit dus altijd:

Bert zegt: "Cees heeft kaart 6."

Berts uitspraak is uiteraard eveneens equivalent met een – steeds wisselend, afhankelijk van wat Anna zegt – aantal alternatieven voor zijn eigen hand.

Voor het zeven-kaartenprobleem bestaan alle oplossingen die uit twee uitspraken bestaan uit vijf, zes of zeven alternatieve handen kaarten voor Anna (het bewijs dat het niet in vier of acht kan, laat ik achterwege). Oplossing (v) die hiervoor genoemd is, is een variant waarin *Anna de som modulo 7 van haar kaarten zegt.* In dit geval dus:

Anna zegt: "De som modulo 7 van mijn kaarten is 3."

Dit is eigenlijk:

Anna zegt: "Mijn hand kaarten is een van 012, 046, 136, 145, 235."

Bert heeft 345, behalve in 012 komen een of meer van Berts kaarten in de alternatieven voor, dus deze informatie is voor Bert voldoende om Anna's kaarten te leren en daarmee de kaart van Cees: dus hij kan zijn bewering naar waarheid doen. Tevens moeten we aantonen dat na Anna's bewering gemeenschappelijk bekend is dat Cees geen van Anna's of Berts kaarten kent. Met andere woorden, Cees leert geen van Anna's of Berts kaarten, wat ook de werkelijke kaartverdeling was:

Als Cees kaart 0 had gehad, dan had Anna nog 136, 145 en 235 kunnen hebben. Cees leert nu geen van Anna's of Berts kaarten, want elk van $1, 2, 3, 4, 5, 6$ komt in ten minste een van die drie handen *wel* en in ten minste een van die drie handen *niet* voor.

Als Cees kaart 1 had gehad, dan ..., etc. Nadat Bert heeft gezegd dat Cees kaart 6 heeft, is gemeenschappelijk bekend dat *aweetb*, *bweeta* en *cweetniet*. De eerste twee zijn eenvoudig in te zien, voor de laatste volstaat de bovenstaande redenering voor Cees te herhalen voor kaart 6.

Deze oplossing is zowel minimaal als maximaal. Minimaal, omdat uitspraken van vier handen Cees altijd informatie geven. Het bewijs van de maximaliteit laten we zien, omdat het typerend is voor andere maximaliteits- en ook minimaliteitsbewijzen. Stel we breiden de vijf alternatieven voor Anna met een andere hand h uit. Twee van de in deze zesde hand voorkomende kaarten moeten eveneens in ten minste een van die vijf handen voorkomen (ga na!). Noem die hand h'. Als Bert precies de drie kaarten had gehad die niet in h of h'

voorkomen, dan had hij niet op grond van Anna's uitspraak haar hand kaarten kunnen bepalen: hij kan dan immers niet bepalen of ze h of h' heeft. Maar hij kon het *wel*. Dus heeft Anna noch h noch h'. De vier resterende handen zijn dan weer te informatief voor Cees.

Het volgende is een hele andere oplossing, die uit zeven handen kaarten bestaat:

Anna zegt: "Ik heb een van 012, 034, 056, 135, 146, 236, 245."

Deze oplossing kan gezien worden als de zeven lijnen in een projectief vlak van zeven punten. De oplossing is maximaal maar niet minimaal: een willekeurige van die zeven handen, behalve de werkelijke, kan weggelaten worden en ook *dat* is een oplossing. Die is dan wel minimaal.

Door variatie op deze oplossingen is te bewijzen dat er 102 niet-equivalente oplossingen zijn voor deze gegeven kaartverdeling (namelijk 6 van zeven handen, 36 van zes, en 60 van vijf). Dit zijn *alle* manieren om met twee communicaties het probleem op te lossen.

Een interessante zijweg, waarmee we deze rondgang langs oplossingen afsluiten, is te bewandelen als we alleen eisen dat Cees niet de *hele* hand van Anna en/of van Bert te weten komt. Dus een of zelfs twee van hun kaarten leren mag wel. Alle voorgaande oplossingen lossen uiteraard ook dit eenvoudiger probleem op, maar tevens is nu een simpeler oplossing voorhanden, namelijk:

Anna zegt: "Mijn hand is een van 012, 034, en 056."

Hierna is gemeenschappelijk bekend dat Anna kaart 0 heeft en dat Bert Anna's hand kent, waarna Bert dus wederom de kaart van Cees bekend maakt.

4 Geschiedenis en relevantie van de kennislogica

Kennislogica is een zogenaamde *modale logica*. In zekere zin begint de modale logica al bij Aristoteles. De relationele semantiek voor modale logica komt van Kripke [10], kennislogica (voor individuele agents) van Hintikka [9], de uitbreiding met gemeenschappelijke kennis van Lewis [11] en daarna Aumann [1], met verschillende belangrijke bijdragen van auteurs van [3]. Dynamische kennislogica, dus met update-operatoren, maar dit kan nog veel wilder, is van meer recente datum, zie [14, 7, 2, 17]. Dit is eigenlijk de integratie van kennislogica met dynamische logica. De laatste heeft weer een geheel eigen voorgeschiedenis, zie [16, 8].

Multi-agentsystemen dateren van ruwweg de afgelopen 25 jaar. Dit is de verzamelnaam voor distributieve computersystemen die 'net als mensen' doelgericht gedrag vertonen op basis van interactie met hun omgeving. Voor een recente introductie in dit gebied, zie [22]. Kennislogica wordt gebruikt voor het specificeren van de abstracte architectuur van zo'n multi-agentsysteem. Hiervoor is een heel elegant model voorhanden [3]. Een (voor een computer) redelijke aanname is dat iedere agent of processor $n \in N$, en ook de omgeving, een eindig aantal lokale toestanden $l_n \in L_n$, respectievelijk $o \in O$, kan aannemen. Een

globale toestand van een multi-agentsysteem voor $|N| = m$ agents is dan simpelweg een punt $(l_1, \cdots, l_m, o)$ in het Cartesisch product $L_1 \times \cdots \times L_m \times O$. Een andere redelijke aanname is dat iedere processor zijn eigen toestand kent (net zoals iedere kaartspeler zijn eigen kaarten kent). Dit induceert equivalentierelaties op deze waardenverzameling, namelijk tussen globale toestanden met dezelfde n-coordinaat: $(l_1, \cdots, l_n, \cdots, l_m, o) \sim_n (l'_1, \cdots, l_n, \cdots, l'_m, o')$, zodat zo'n systeem formeel volgens definitie 2 als informatietoestand te representeren is.

Het zeven-kaartenprobleem is dus maar een van de multi-agentsystemen waarvan het gedrag zich goed met kennislogica laat specificeren. Kennislogica wordt tevens toegepast voor het correctbewijzen van protocollen voor informatieoverdracht, zoals het alternating-bit protocol [3] en het TCP/IP protocol [15]. En behalve publieke updates, zijn ook ingewikkelder updates voorstelbaar. Een voorbeeld binnen de huidige setting is de actie van het laten zien van je kaart aan (alleen) een andere speler [19]: alle kaartverdelingen blijven nu actueel, maar alleen de equivalentierelaties ertussen veranderen.

De studie van het uitwisselen van geheimen tussen kaartspelers is tevens relevant voor het ontwerpen van cryptografische protocollen. Een interessant verschil met 'public / private key encryption' is dat het geheim niet gegarandeerd is door de complexiteit van een berekening (namelijk het ontbinden in priemfactoren van een groot getal), maar dat het geheim fundamenteel (dat wil zeggen, ook voor 'computationally unlimited agents') niet achterhaalbaar is. Dit lichten we toe aan een simpele maar niet helemaal serieuze analogie:

Stel dat Anna de bank is die de credit card van Bert wil identificeren, en Cees de boef die geld van Bert wil stelen. Anna weet dat Cees en Bert beide slecht in hoofdrekenen zijn. Dus ter identificatie vraagt Anna wie van hen het eerst 161393 in priemfactoren kan ontbinden. (En kunt u dit?) Bert beschikt over 'private key' 643, voert een staartdeling uit, en geeft als antwoord: 643×251. Cees druipt af. (In termen van de veel grotere priemgetallen waar het in werkelijkheid om gaat: Cees heeft meer tijd nodig om de priemfactorontbinding te maken dan de ouderdom van het heelal.) In ons huidige scenario houdt Anna voor mogelijk dat Cees een 0 is en zegt daarom dat haar hand kaarten een is van 251, 203, 246, 504, 536, 106, 134. Wederom kan Bert zich als eerste identificeren met gebruikmaking van de informatie ('hand kaarten') 643 waar hij over beschikt, door Anna mede te delen dat zij 251 heeft, of, veiliger, door en public Cees te ontmaskeren als een 0. Anders dan hiervoor kan Cees zich nu onmogelijk zonder risico als Bert doen voorkomen, hoe goed hij ook kan hoofdrekenen.

De ontwikkeling van de kennislogica lijkt deels te drijven op de analyse van puzzels waarin kennis een rol speelt. Het 'Modderige-kinderenprobleem' is al genoemd [3]. Een van de andere klassiekers is 'Som en Product':

A zegt tot S en P: Ik heb twee gehele getallen x, y gekozen met $1 < x < y$ en $x + y \leq 100$. Straks deel ik $s = x + y$ aan S alleen mee, en $p = xy$ aan P alleen. Deze mededelingen blijven geheim. Maar jullie moeten je inspannen om het paar (x, y) uit te rekenen.

Hij doet zoals aangekondigd. Nu volgt dit gesprek:

1. *P zegt: Ik weet het niet.*
2. *S zegt: Dat wist ik al.*
3. *P zegt: Nu weet ik het.*
4. *S zegt: Nu weet ik het ook.*

Bepaal het paar (x, y).

Dit probleem is door John McCarthy, Martin Gardner, en anderen gepopulariseerd vanaf de jaren zeventig, en is onlangs door Rineke Verbrugge met veel verve tijdens de Nationale Wiskundedagen 2002 opnieuw ten tonele gebracht. Internationaal lijkt minder bekend dat de oudste publicatie over dit raadsel van Hans Freudenthal is, in het Nieuw Archief [5, 6]. De formulering van het probleem hiervoor is het letterlijke citaat. De lijst van goede inzenders in [6] is onthullend voor de huidige stand van wiskunde en informatica in Nederland.

References

[1] R.J. Aumann. Agreeing to disagree. *Annals of Statistics*, 4(6):1236–1239, 1976.

[2] A. Baltag. A logic for suspicious players: Epistemic actions and belief updates in games. *Bulletin of Economic Research*, 54(1):1–45, 2002.

[3] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge MA, 1995.

[4] M.J. Fischer and R.N. Wright. Bounds on secret key exchange using a random deal of cards. *Journal of Cryptology*, 9(2):71–99, 1996.

[5] H. Freudenthal. (formulering van het 'som-en-product'-probleem). *Nieuw Archief voor Wiskunde*, 17:152, 1969.

[6] H. Freudenthal. (oplossing van het 'som-en-product'-probleem). *Nieuw Archief voor Wiskunde*, 18:102–106, 1970.

[7] J.D. Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, University of Amsterdam, 1999. ILLC Dissertation Series DS-1999-01.

[8] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge MA, 2000. Foundations of Computing Series.

[9] J. Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, NY, 1962.

[10] S. Kripke. A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24:1–14, 1959.

[11] D. Lewis. *Convention, a Philosophical Study*. Harvard University Press, Cambridge MA, 1969.

[12] K.S. Makarychev and Yu.S. Makarychev. The importance of being formal. *Mathematical Intelligencer*, 23(1):41–42, 2001.

[13] van Ditmarsch hexagon. Open college UvA 'Hoe Wiskunde Werkt, van Natuur tot Cognitie', `http://www.science.uva.nl/projects/opencollege/cognitie/-hexagon/`, 2002.

[14] J.A. Plaza. Logics of public communications. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic, and Z.W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216, 1989.

[15] F. Stulp and L.C. Verbrugge. A knowledge-based algorithm for the internet transmission control protocol (tcp). *Bulletin of Economic Research*, 54(1):69–94, 2002.

[16] J.F.A.K. van Benthem. *Exploring logical dynamics*. CSLI Publications, 1996.

[17] H.P. van Ditmarsch. *Knowledge games*. PhD thesis, University of Groningen, 2000. ILLC Dissertation Series DS-2000-06.

[18] H.P. van Ditmarsch. Killing cluedo. *Natuur & Techniek*, 69(11):32–40, 2001.

[19] H.P. van Ditmarsch. Descriptions of game actions. *Journal of Logic, Language and Information*, 11:349–365, 2002.

[20] H.P. van Ditmarsch. Keeping secrets with public communication. In *Proceedings of LOFT5*, Turin, Italy, 2002. ICER. `http://www.cs.otago.ac.nz/staffpriv/hans/hvdLOFT02.ps`.

[21] H.P. van Ditmarsch. Oplossing van het mysterie (solution of the murder mystery). *Natuur & Techniek*, 70(2):17, 2002.

[22] M. Wooldridge. *An Introduction to Multiagent Systems*. John Wiley & Sons Ltd, Chichester, UK, 2002.