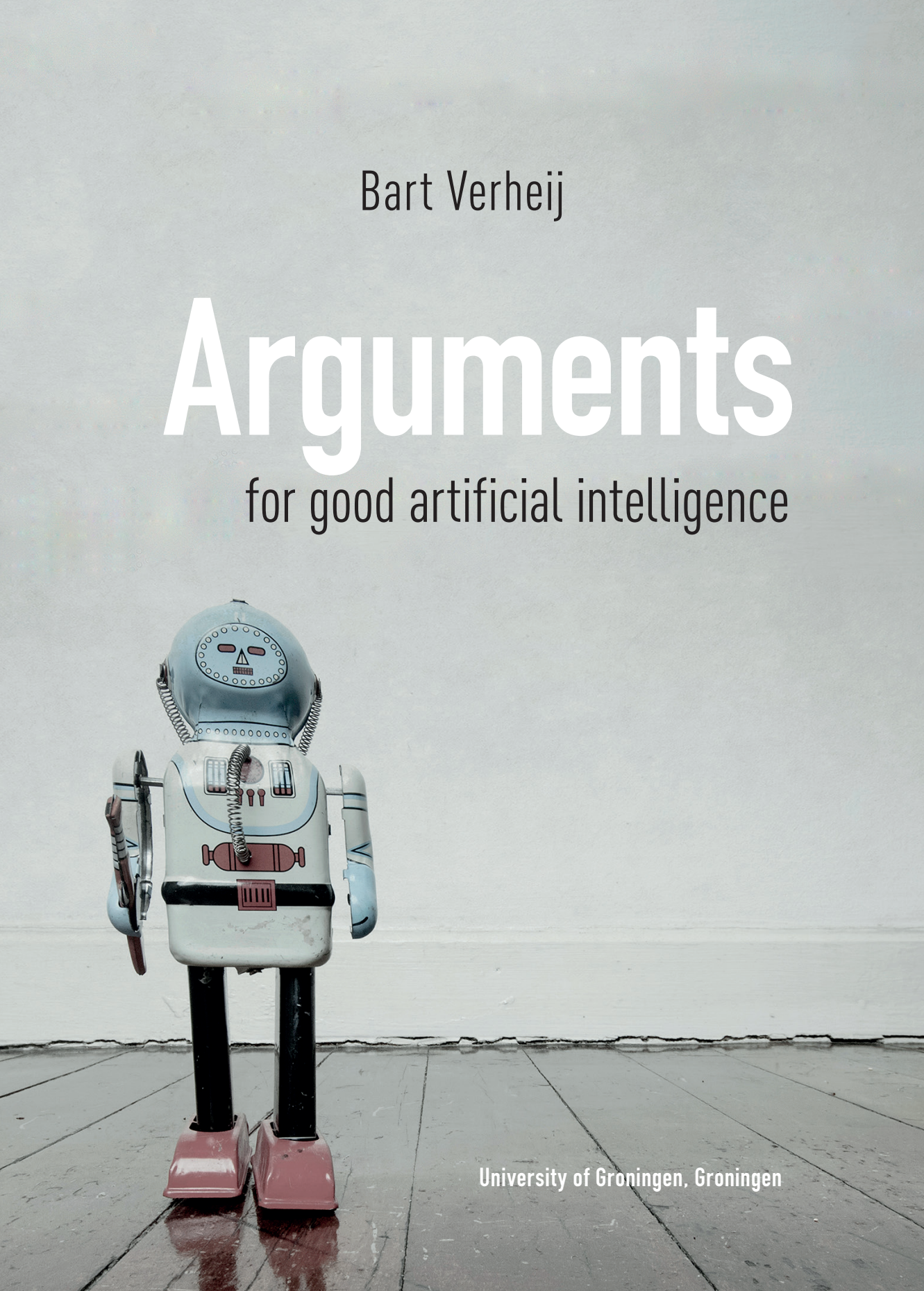
Arguments for Good Artificial Intelligence

Verheij, Bart

Bart Verheij

Arguments

for good artificial intelligence

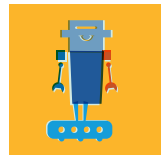


University of Groningen, Groningen

ARGUMENTS FOR GOOD ARTIFICIAL INTELLIGENCE

Bart Verheij

Arguments for good artificial intelligence



University of Groningen, Groningen

PUBLISHED BY UNIVERSITY OF GRONINGEN, GRONINGEN

The book has been typeset in LaTeX using the `tufte-book` class.

Cover design and good AI logo: Charlot Luiting

ISBN 978-94-034-1016-6
ISBN 978-94-034-1015-9 (e-boek)
NUR 984 Kunstmatige intelligentie

Contents

Preface 9

Argumenten voor goede kunstmatige intelligentie 11

Dromen en angsten 11
Kunstmatige intelligentie nu 13
Goede kunstmatige intelligentie 17
Argumentatiesystemen 21
Wiskundige grondslagen van argumentatie 23
Zoektocht naar de goede wiskunde 26
Over het vangen van een dief 30
Terug naar de wiskundige grondslagen van argumentatie 34
Argumenten, gevallen en regels 36
Argumenten overbruggen de kloof tussen kennis- en datasystemen 37

Arguments for good artificial intelligence 41

Dreams and fears 41
Artificial intelligence now 43
Good artificial intelligence 47
Argumentation systems 50
Mathematical foundations of argumentation 53

Search for the right mathematics 55
About catching a thief 60
Return to the mathematical foundations of argumentation 63
Arguments, cases and rules 65
Arguments bridge the gap between knowledge and data systems 66

Dankwoord 69

Curriculum vitae 73

Bibliography 75

List of Figures

1 'Kunstmatige intelligentie' in het NRC Handelsblad 11
2 Een zelfrijdende auto; AlphaGo v. Tang Weixing; robot Baxter 12
3 Humans need not apply; the frightful five; ban killer robots 12
4 Een splitsing in de kunstmatige intelligentie 15
5 Argumentstructuur van de onrechtmatige daad 16
6 Vormen van de hoofdletter A; structuur van een neuraal netwerk 16
7 Van Paterswolde naar Groningen; van Groningen naar Amsterdam 18
8 Poor Man's Watson 19
9 Mensen volgens Google 20
10 Argumentatie 22
11 Een splitsing in het argumentatieonderzoek 23
12 Abstracte argumentatie als een vorm van grafentheorie 24
13 Evaluatie van aanvalsgrafen 24
14 Het argumentatievoorbeeld in ArguMed 25
15 Argumentatiesemantieken 25
16 Drie hulpmiddelen om bewijs te ordenen en evalueren 26
17 Argumenten en scenarios 27
18 Een Bayesiaans netwerk 28
19 Scenarios en kansen 29
20 Argumenten en kansen 30
21 Het misdaadverhaal uit Alfred Hitchcock's 'To Catch A Thief' 32
22 Casusmodel voor Alfred Hitchcock's 'To Catch A Thief' 32
23 Gevalsmodel voor de onrechtmatige daad 36
24 Voorbij de splitsing in de kunstmatige intelligentie 37

1 'Artificial intelligence' in the NRC Handelsblad 41
2 A self-driving car; AlphaGo v. Tang Weixing; robot Baxter 42
3 Humans need not apply; the frightful five; ban killer robots 42
4 A gap in artificial intelligence 44
5 Argument structure for Dutch law of unlawful acts 46
6 Forms of the capital A; structure of a neural network 46
7 From Paterswolde to Groningen; from Groningen to Amsterdam 47
8 Poor Man's Watson 48
9 Human beings according to Google 50
10 Argumentation 51
11 A gap in argumentation research 52
12 Abstract argumentation as a form of graph theory 53
13 Evaluation of attack graphs 53
14 The argumentation example in ArguMed 54
15 Argumentation semantics 55
16 Three tools for organizing and evaluating evidence 56
17 Arguments and scenarios 57
18 A Bayesian network 58
19 Scenarios and probabilities 59
20 Arguments and probabilities 60
21 The crime story in Alfred Hitchcock's 'To Catch A Thief' 61
22 Case model of Alfred Hitchcock's 'To Catch A Thief' 61
23 Case model for Dutch law of unlawful acts 66
24 Beyond the gap in artificial intelligence 67

Preface

This book contains the text of my inaugural address ('oratie'), delivered on September 12, 2017, in the Aula of the Academiegebouw of the University of Groningen (Broerstraat 5, Groningen). The main claim is that by developing argumentation systems we can arrive at good artificial intelligence, i.e., artificial intelligence that provides good answers, has good reasons and makes good choices. Hence, argumentation systems research can bridge the gap between knowledge-based and data-driven systems and provides a specific path towards explainable, responsible and social artificial intelligence. The text appears twice: first in the original Dutch, and then in English translation. That version was presented as an Orient Forum lecture at Zhejiang University, Hangzhou, on April 17, 2018.

Groningen, September 2018

Argumenten voor goede kunstmatige intelligentie

Geachte Rector Magnificus,
zeer gewaardeerde toehoorders,

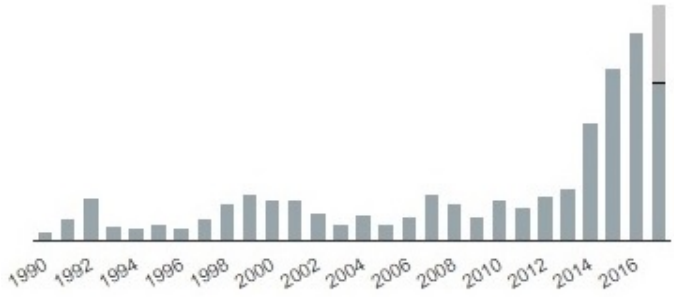
Dromen en angsten

Dezer dagen staan er regelmatig verhalen over kunstmatige intelligentie in de krant. Het worden er bovendien steeds meer (figuur 1). En passend bij het onderwerp zijn dat verhalen vol van dromen en van angsten.

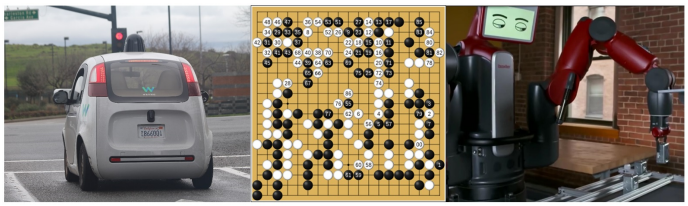
Soms lijkt het wel of computers al bijna alles kunnen: als we de krantenkoppen mogen geloven, kunnen machines intussen prima voetballen, vrachtwagens besturen, en adviseren over werk.[1]

En de kunstmatige intelligentie van nu spréekt ook tot de verbeelding. Er rijden al jaren zelfrijdende auto's op de openbare weg; bij het ene na het andere klassieke denkspel worden de beste mensen door computers verslagen; en op de werkvloer zien we al de humanoide robots uit sciencefictionverhalen (figuur 2).

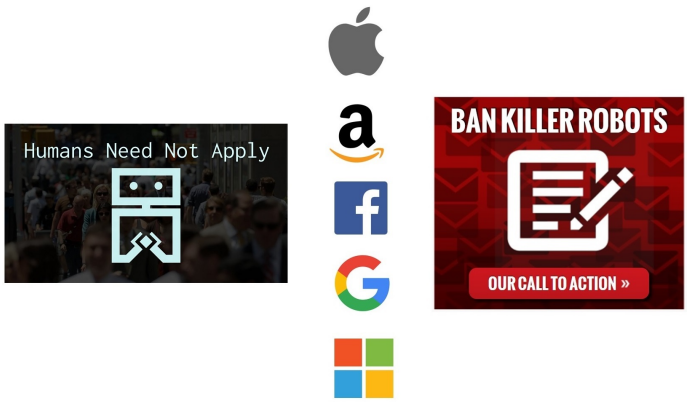
[1] Een paar recente koppen uit het NRC Handelsblad: Voetbalrobots hebben nu een loepzuiver en dodelijk schot, 21 juli 2017; Ruim baan voor de robottruck, 17 februari 2017; Algoritme geeft werkzoekende sollicitatie-advies, 27 juli 2017



Figuur 1: Aantal artikelen met trefwoord 'kunstmatige intelligentie' in het NRC Handelsblad (1990–2017). De laatste balk is een extrapolatie op basis van de eerste 8 maanden van 2017. Bron: www.nrc.nl



Figuur 2: Een zelfrijdende auto; AlphaGo v. Tang Weixing; robot Baxter. Bron: www.wikipedia.org



Figuur 3: Humans need not apply; the frightful five; ban killer robots. Bron: www.wikipedia.org

Met de dromen groeien ook de angsten, want steeds verder gaande automatisering van werk dat nu nog alleen door mensen wordt gedaan zal de arbeidsmarkt grondig veranderen; door automatische analyse van alle informatie die wij over onszelf online zetten krijgen bedrijven en overheden steeds meer controle over ons; en het idee van wapensystemen die zelfstandig besluiten of ze tot de aanval overgaan is al helemaal griezelig (figuur 3). Argumenten genoeg dus om voorzichtig te zijn met kunstmatige intelligentie.

Het is dan ook geen wonder dat er steeds meer aandacht is voor de vraag hoe we kunstmatige intelligentie op een verstandige manier in onze samenleving kunnen inpassen. In een recente opinie van het Europees Economisch en Sociaal Comité, een adviesorgaan van de Europese Unie, wordt bijvoorbeeld gepleit voor een *human-in-command* benadering van kunstmatige intelli-

gentie, waarbij machines machines blijven en mensen te allen tijde de controle over deze machines zullen behouden.[2] Nog recenter is de oproep in een open brief om 'killer robots' te verbieden, ondertekend door topmensen van ruim 100 bedrijven op het gebied van robotica en kunstmatige intelligentie, waaronder Tesla en Google.[3]

Ik zal u uitleggen dat er behalve menselijke controle en verbieden nog een derde weg is om kunstmatige intelligentie verstandig in te passen in onze samenleving, en dat is door de ontwikkeling van goede kunstmatige intelligentie. De weg daarnaartoe—zo zal ik betogen—is de ontwikkeling van argumentatiesystemen, dat wil zeggen systemen die een kritische discussie op basis van argumenten kunnen voeren.

[2] C. Muller. Kunstmatige Intelligentie – de Gevolgen van Kunstmatige Intelligentie voor de (Digitale) Eengemaakte Markt, de Productie, Consumptie, Werkgelegenheid en Samenleving. *Advies Europees Economisch en Sociaal Comité*, INT/806, 2017

[3] `futureoflife.org/autonomous-weapons-open-letter-2017`, 20 augustus 2017

Kunstmatige intelligentie nu

In dit verband is het goed te beseffen hoe het er voor staat in de kunstmatige intelligentie.[4] Het is dan handig om specialistische, algemene en superieure kunstmatige intelligentie te onderscheiden.

- *Specialistische kunstmatige intelligentie* is de tegenwoordig heel gewone vorm van kunstmatige intelligentie die specifiek afgebakende intelligente taken kan uitvoeren. Alle kunstmatige intelligentie die nu bestaat is specialistisch. Denk aan computerprogramma's die de expertise hebben om het invullen van een belastingaangifte kinderspel te maken en aan smartphone apps die in onze eindeloze fotoverzameling heel aardig een rijtje plaatjes met bomen kunnen vinden. De meeste onderzoekers houden zich bezig met specialistische kunstmatige intelligentie.
- Met *algemene kunstmatige intelligentie* wordt verwezen naar computerprogramma's of machines die zich goed redden onder een grote variatie aan omstandigheden en bij een breed palet aan problemen; net zoals we dat gewend zijn bij mensen. Ze kunnen bijvoorbeeld boeken begrijpen en ook verzinnen, en ze kunnen leren fietsen door een drukke straat, ook als ze niet

[4] Voor een introductie zie Russell and Norvig 2010 en ook de werken van Douglas Hofstadter, met name Hofstadter 1979, 1985, 2007.

in Nederland zijn geboren. Algemene kunstmatige intelligentie bestaat nu niet. Sommige onderzoekers denken na over de vraag hoe we algemene kunstmatige intelligentie kunnen bereiken, of anders waarom dan niet.

- *Superieure kunstmatige intelligentie* is de vorm van kunstmatige intelligentie waar sommige mensen zich grote zorgen over maken. Het idee van superieure kunstmatige intelligentie is dat, áls we eenmaal algemene kunstmatige intelligentie hebben bereikt, menselijke intelligentie onmiddellijk op onoverbrugbare achterstand staat. Is er nog wel plaats voor de mens na de uitvinding van superieure kunstmatige intelligentie? Niemand die het weet. Veel onderzoekers vinden het leuk om op feesten en recepties over superieure kunstmatige intelligentie van gedachten te wisselen, maar zijn in hun dagelijks werk druk met het aanpakken van de wetenschappelijke hordes die nog genomen moeten worden op hun specifieke deelgebied.

Een belangrijke horde die hoognodig genomen moet worden is het overbruggen van de kloof tussen kennis- en datasystemen. In kennissystemen wordt de kennis die nodig is om een complexe taak uit te voeren rechtstreeks in de computer ingevoerd; de kennis wordt gerepresenteerd en daarmee wordt automatisch geredeneerd. In datasystemen wordt de aanpak van een probleem geleerd door de automatische analyse van een databank met voorbeelden.

Aanvankelijk bestond de kloof tussen de twee typen intelligente systemen niet (figuur 4). Midden twintigste eeuw toen kunstmatige intelligentie als vakgebied ontstond—de term 'artificial intelligence' is verzonnen in 1955—was het programmeren van computers zelf nog nieuw en alle mogelijke aanpakken van kunstmatige intelligentie werden geprobeerd. Geleidelijk aan groeiden het onderzoek naar kennissystemen—gebaseerd op

representeren en redeneren—en dat naar datasystemen—gebaseerd op leren van voorbeelden—uit elkaar. Ook de gebruikte wiskunde is verschillend. Kennissystemenonderzoek gebruikt vaak de logica; datasystemenonderzoek vaak de kansrekening.

Hier ziet u een wetsartikel over schadevergoedingsplicht op grond van onrechtmatige daad, bijvoorbeeld als door uw schuld iemands telefoon kapot valt:

> *Artikel 6:162 lid 1.* Hij die jegens een ander een onrechtmatige daad pleegt, welke hem kan worden toegerekend, is verplicht de schade die de ander dientengevolge lijdt, te vergoeden.

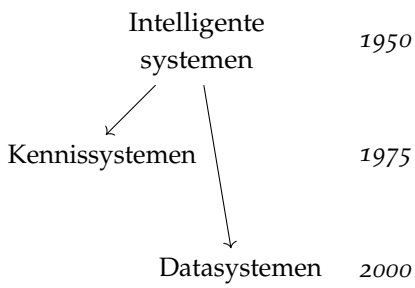
In een representatie van dit leerstuk worden vier voorwaarden onderscheiden voor zo'n schadevergoedingsplicht. ALS er schade is EN de gedraging is onrechtmatig EN die is toerekenbaar EN er is causaal verband tussen de gedraging en de schade DAN is er schadevergoedingsplicht. U ziet hiervan ook een logisch geformaliseerde versie (die regel hiernaast met drieletterige afkortingen en symbolen), en het pijlendiagram (figuur 5) laat een uitgebreidere representatie van dit leerstuk zien met meer regels en uitzonderingen.[5]

```
ALS schade
  EN onrechtmatig
  EN toerekenbaar
  EN causaal-verband
DAN schadevergoedingsplicht
```

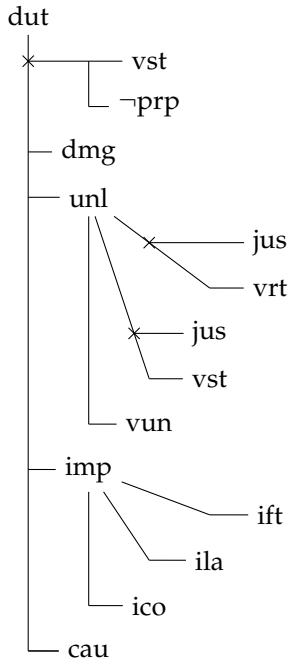
$$\mathrm{dmg} \wedge \mathrm{unl} \wedge \mathrm{imp} \wedge \mathrm{cau} \rightsquigarrow \mathrm{dut}$$

[5] Zie ook Verheij et al. 1997, Verheij 2017b

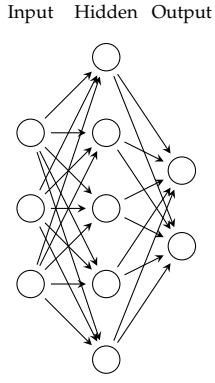
Door een computerprogramma te vullen met dit soort, vaak regelachtige gerepresenteerde kennis kunnen allerlei intelligente systemen worden gebouwd. Studenten kunstmatige intelligentie in Groningen hebben bijvoorbeeld systemen gebouwd voor huisartsentriage, poëzieclassificatie, een digitale dominee—zelfs de exper-



Figuur 4: Een splitsing in de kunstmatige intelligentie



Figuur 5: Argumentstructuur van de onrechtmatige daad



Figuur 6: Vormen van de hoofdletter 'A' (Hofstadter 1995); structuur van een neuraal netwerk

tise van een kruidendokter is eens in een kennissysteem gevangen.

In datasystemen wordt de aanpak van een probleem geleerd door de analyse van een databank met voorbeelden. Aan bijvoorbeeld een neuraal netwerk worden aan de 'input'-kant allerlei vormen van de hoofdletter A getoond (figuur 6), zodat geleidelijk aan de interne structuur van het netwerk—met hier een 'hidden layer' in het midden—kan worden aangepast om zulke letters correct te herkennen aan de 'output'-kant. Een tijdlang werd gedacht dat neurale netwerken fundamentele beperkingen zouden hebben, terwijl die juist zijn uitgegroeid tot een heel krachtige data-analysetechniek.

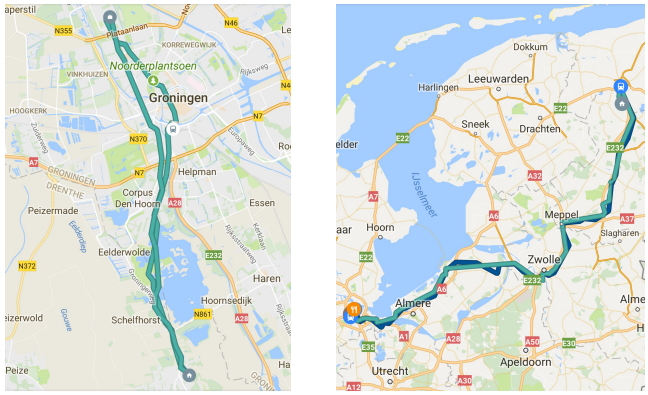
Goede kunstmatige intelligentie

Ik stelde al dat we toe moeten naar de ontwikkeling van goede kunstmatige intelligentie. Dan denk ik aan drie noodzakelijke kenmerken:

- Ten eerste moet een intelligent systeem de *goede antwoorden* kunnen geven op problemen;
- ten tweede moet een intelligent systeem daarvoor *goede redenen* kunnen geven; en
- ten derde moet een intelligent systeem de *goede keuzes* kunnen maken.

Eerst over de goede antwoorden. Datasystemen zijn niet ontworpen voor het geven van goede antwoorden, maar voor het *zo vaak mogelijk* geven van goede antwoorden, en op allerlei terreinen zijn ze daar heel goed in. Regelmatig geven ze ook váker goede antwoorden dan mensen dat doen. Dat zijn knappe prestaties. Tegelijk is het zo dat datasystemen vergissingen maken die een mens niet snel zal maken.

Een voorbeeld. Doordat ik mijn telefoon meestal op zak heb, kan Google goed bijhouden waar ik ben. Hier is bijvoorbeeld een Google-plaatje te zien van mijn dagelijkse route van huis naar werk en terug (figuur 7, links). De route is groen gekleurd. Die kleur betekent een

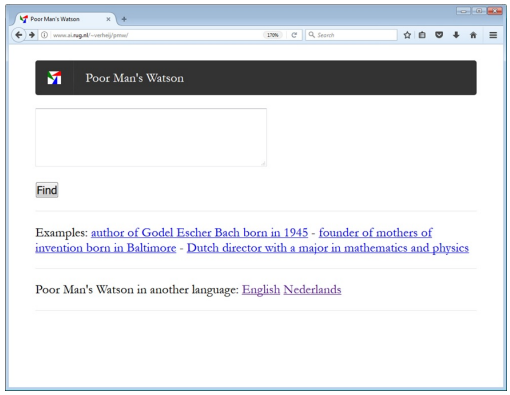


Figuur 7: Van Paterswolde naar Groningen en terug; en van Groningen naar Amsterdam en terug. Bron: Google Maps, tijdlijn

fietstocht en het datasysteem van Google heeft dan ook terecht aangenomen dat ik op de fiets was.

Kijk nu naar het plaatje rechts van een reis van Groningen naar Amsterdam en terug. Google heeft de heenreis blauw gekleurd—wat klopt want ik was met de trein. Maar de terugreis is groen, wat betekent dat Google denkt dat ik terug ben gefíetst, wat op een normale werkdag toch wat veel van het goede zou zijn. Je zou eigenlijk het datasysteem willen corrigeren en zeggen: 'Mensen fietsen niet in twee uur van Amsterdam naar Groningen'. Maar zo werkt het niet bij datasystemen. Zo'n correctie kan wel in een kennissysteem.

Dan de goede redenen. Aan datasystemen kun je niet vragen waarom ze tot een bepaald antwoord komen. Een datasysteem heeft namelijk geen expliciete redenen. De data als geheel is de onderbouwing van het antwoord. Dat die treinreis per ongeluk als een fietstocht werd geclassificeerd komt door de analyse van de beschikbare data, die over mij en die over anderen. Volgens die data-analyse lijkt mijn treinreis net wat meer op de fietstochten in de dataset dan op de treinreizen. Waar de uitkomst precies op gebaseerd is is lastig vast te stellen. Datasystemen worden daarom wel als 'black boxes', 'zwarte dozen' gezien. De binnenkant van een datasysteem is niet goed te zien, de werking van het systeem is niet transparant.

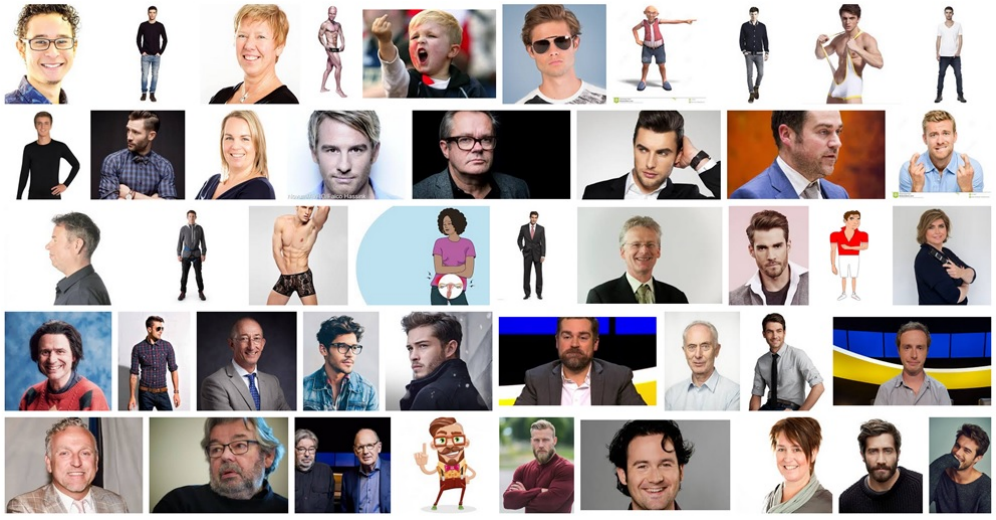


Figuur 8: Poor Man's Watson. Bron: www.ai.rug.nl/~verheij/pmw/

Een voorbeeld is mijn 'Poor Man's Watson' (figuur 8). Het is geïnspireerd door IBM's Watson dat in 2011 de Amerikaanse kenniskwis Jeopardy! won van de beste menselijke spelers.[6] Dat was de spectaculaire uitkomst van een forse onderzoeksinvestering, terwijl 'Poor Man's Watson' het resultaat is van een middag programmeren door één onderzoeker; vandaar de naam. Ik bouwde het vooral om te laten zien wat er met een eenvoudig script dat gebruik maakt van Google en Wikipedia kon en daardoor beter te begrijpen hoe bijzonder de Jeopardy!-prestatie was. En in sommige opzichten lijkt 'Poor Man's Watson' al behoorlijk intelligent.

[6] en.wikipedia.org/wiki/Watson_(computer)

Als auteur van 'Gödel, Escher, Bach' geboren in 1945 wordt netjes geantwoord met 'Douglas Hofstadter'. En ook het antwoord op de 'founder of the mothers of invention born in Baltimore' is correct. Tenminste dat was zo toen ik het gisteren nog testte. Omdat het systeem aan het internet hangt en afhankelijk is van de informatie die daar te vinden is kan er van alles gebeuren. Soms gaat iets wat lang goed ging opeens niet meer goed. Een tijdje werkte bijvoorbeeld 'born in 1940' ook goed voor de in Baltimore geboren Frank Zappa maar dat werkt nu niet meer. Wat de reden daarvoor is kan niet aan het systeem worden gevraagd want het systeem heeft geen redenen. Bij het reisvoorbeeld zou je eigenlijk aan het datasysteem willen vragen:



Figuur 9: Mensen volgens Google. Bron: Google afbeeldingen, zoekvraag 'mens'

'Waarom denk je dat ik van Amsterdam naar Groningen ben gefietst?' Maar daar heeft een datasysteem geen antwoord op. Dat heeft een kennissysteem wel.

Ten derde goede keuzes. Datasystemen kiezen niet zelf. Ze zijn naar hun aard beschrijvend. Ze beschrijven de data waarop ze gebaseerd zijn. Die objectiviteit is tegelijk hun kracht en hun zwakte. Als de data goed zijn, zijn de keuzes goed. Maar de data zijn niet altijd goed. Een bekend voorbeeld is Tay, een chatbot van Microsoft.[7] Tay ging op 23 juni 2016 online. Tay leerde van de voorbeelden van gesprekken met gebruikers. En dus—internet is internet—werden de gesprekken met Tay steeds vreemder, onplezieriger en soms beledigend, want zo zijn de gesprekken op internet vaak. Na 16 uur haalde Microsoft Tay weer offline.

[7] en.wikipedia.org/wiki/Tay_(bot)

Een ander voorbeeld is het vermeende racisme van zoekmachines dat zo nu en dan in het nieuws komt. Ik was benieuwd hoe het er voor stond, tikte van de week het woord 'mens' en kreeg inderdaad heel veel plaatjes van mensen te zien (figuur 9). Van de 36 mensen die ik in beeld kreeg waren er vier vrouw, waarvan 1 getekend.

Die getekende vrouw was de enige duidelijk niet-witte mens. Kennelijk is dit het beeld van mensen zoals dat uit Google's data oprijst.

De les is duidelijk: data beschrijven is niet genoeg om een complexe wereld te begrijpen. Heel vaak moet er van de data worden afgeweken om de gewenste doelen te bereiken. Je zou eigenlijk een datasysteem willen opvoeden en zeggen: 'Wil je nou wel eens ophouden met die beledigingen?' Maar daar luistert een datasysteem niet naar. Bij een kennissysteem kan dat wel.

Argumentatiesystemen

Kennissystemen kunnen wél goede antwoorden geven, kunnen wél goede redenen hebben en kunnen wél goede keuzes maken. Precies de eigenschappen van goede kunstmatige intelligentie die ik noemde. Maar kennissystemen kunnen weer niet goed wat datasystemen wel goed kunnen. Het is al lang bekend: kennissystemen zijn niet ontworpen om te leren van data, kennissystemen zijn moeilijk schaalbaar, en kennissystemen kunnen niet goed omgaan met visuele en andere meetkundig gestructureerde informatie. En daar zijn datasystemen vaak juist goed in.

Waar we dan ook naar toe moeten is een combinatie van de goede eigenschappen van kennissystemen en van datasystemen. Probleem is alleen dat we vooralsnog niet goed weten hoe de verschillende gebruikte technieken bij elkaar passen.

Om daarmee vooruitgang te boeken stel ik voor argumentatiesystemen te ontwikkelen, oftewel systemen die een kritische discussie op basis van argumenten kunnen voeren. Vandaag gebruik ik deze definitie:

```
Argumentatiesystemen zijn systemen die een kritische
discussie kunnen voeren waarin hypothesen worden
geconstrueerd, getoetst en gewaardeerd op basis
van redelijke argumenten.
```

Een voorbeeld van argumentatie. Als een getuige heeft verklaard dat de verdachte ergens anders was tijdens het

misdrijf, geeft dat een reden dat hij een alibi heeft, wat weer een reden is voor de onschuld van de verdachte. Als dit de enige informatie is kunnen we in de onschuld van de verdachte geloven (figuur 10, links). Maar vervolgens kan blijken dat de getuige de partner is van de verdachte wat een reden is dat diens verklaring een leugen is, wat weer een reden is om de getuige niet te geloven over het alibi (figuur 10, rechts). Het voorbeeld laat zien dat meer informatie het perspectief kan veranderen. Eerst geloven we in de onschuld van de verdachte, later krijgt de twijfel de overhand. Formeel hebben we het hier over een niet-monotoon logisch systeem.[8]

Argumentatie wordt al sinds de oudheid bestudeerd en kwam als wetenschapsgebied volop in ontwikkeling vanaf het midden van de twintigste eeuw—toevallig in dezelfde tijd als de kunstmatige intelligentie.[9] Terwijl daar de kracht van wiskundige, formele methoden werd geëxploiteerd, ging het er in de argumentatietheorie juist vaak over dat de formele methoden van die tijd niet passend waren om argumentatie te begrijpen zoals die 'in het wild' voorkomt (dus in de politiek, in de rechtspraak of thuis).[10] Niet toevallig voor de loop van dit verhaal zijn dat trouwens dezelfde formele methoden waar ik het al over had bij het onderscheid tussen kennis- en datasystemen: de logica en de kansrekening. Er ontstond dan ook een splitsing tussen formeel en informeel argumentatieonderzoek (figuur 11).[11]

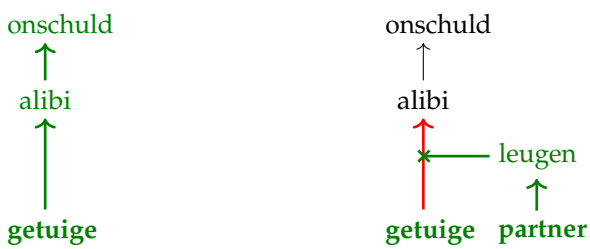
Sinds de jaren zeventig is de informele argumentatietheorie goed op stoom, en voor de formele argumentatie-

[8] Reiter 1980, Pollock 1987, 1995, Gabbay et al. 1994

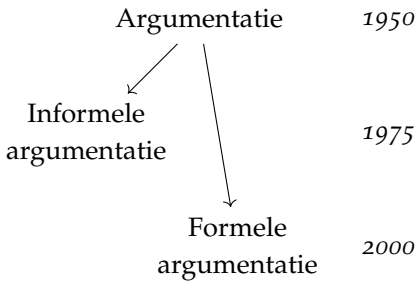
[9] Zie van Eemeren et al. 2014

[10] Toulmin 1958. Zie ook Hitchcock and Verheij 2006, Verheij 2009

[11] Al is die splitsing niet altijd scherp. Zie bijvoorbeeld Barth and Krabbe 1982, Freeman 1991, Walton and Krabbe 1995, Reed and Grasso 2007



Figuur 10: Argumentatie



Figuur 11: Een splitsing in het argumentatieonderzoek

theorie geldt dat vooral vanaf het begin van deze eeuw. Langzaamaan groeien de formele en informele methoden weer naar elkaar toe—juist ook onder invloed van het onderzoek naar argumentatiesystemen in de kunstmatige intelligentie.[12]

Ik zal u een aantal recente ontwikkelingen laten zien aan de hand van de wiskundige grondslagen van argumentatie, correct redeneren met forensisch bewijs en de verbanden tussen regels en casus in het recht.

[12] Zie hoofdstuk 11 in van Eemeren et al. 2014 en Chesñevar et al. 2000, Reed and Norman 2004, Bench-Capon and Dunne 2007, Rahwan and Simari 2009, Atkinson et al. 2017. Zie ook Bondarenko et al. 1997, Vreeswijk 1997, Grasso et al. 2000, Rahwan et al. 2003, García and Simari 2004, Chesñevar et al. 2006, Amgoud and Caminada 2007, Besnard and Hunter 2008, Modgil 2009, Brewka and Woltran 2010, Prakken 2010, Thimm 2012, Brewka et al. 2013, Baroni et al. 2014, Cerutti et al. 2017

Wiskundige grondslagen van argumentatie

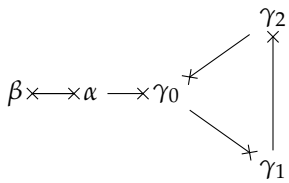
Om te beginnen de wiskundige grondslagen van argumentatie.[13] Zowel in de informele als in de formele literatuur is er veel aandacht besteed aan de structuur van argumentatie. We zagen al een voorbeeld met de argumentatie op basis van een getuigenverklaring.

[13] Simari and Loui 1992. Zie Baroni et al. 2018

Het blijkt dat de wiskunde van de evaluatie van argumenten die elkaar aanvallen en verdedigen door tegenaanvallen verrassend interessant en gevarieerd is. Deze wiskunde wordt sinds het midden van de jaren negentig bestudeerd als een vorm van grafentheorie (figuur 12).[14] In de figuur zijn vijf argumenten afgebeeld en de pijlen geven aan hoe argumenten elkaar aanvallen. De argumenten α en β vallen bijvoorbeeld elkaar aan en argument α ook nog γ_0.

[14] Dung 1995

Evaluatie van aanvalsgrafen is gebaseerd op het idee dat argumenten elkaar kunnen aanvallen maar ook verdedigen. Als argument α niet wordt aangevallen, is het argument onweerlegd en komt het als winnaar uit een argumentatief debat (aangegeven met de groene



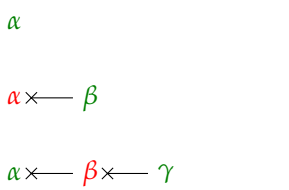
Figuur 12: Abstracte argumentatie als een vorm van grafentheorie

kleur in figuur 13). Als argument β α aanvalt, wint de aanvaller β en verliest α (aangegeven met de rode kleur). Als β zelf weer wordt aangevallen door argument γ, dan verliest β. Argument α verliest nu níet meer door de succesvolle verdediging tegen β door γ. Het argument is als het ware in ere hersteld, het is 'reinstated'.

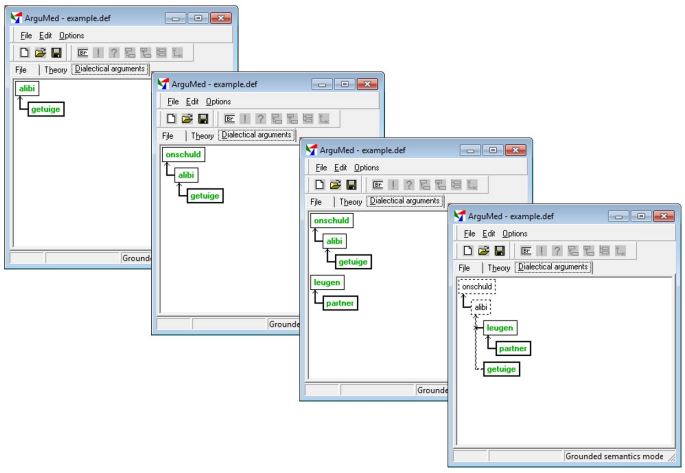
De zo ontwikkelde wiskunde kan als grondslag dienen voor het ontwerpen van computerprogramma's voor het construeren en evalueren van argumentatie. In figuur 14 wordt het voorbeeld dat ik net gaf in mijn ArguMed-programma[15] opgebouwd en geëvalueerd. Een formele bijzonderheid aan deze software is dat de grafische structuur van de diagrammen isomorf is aan de logische structuur van de wiskunde. Een pijl uit een plaatje correspondeert met een conditionele zin uit de logica.

[15] Verheij 2003a, 2005a. Zie ook Kirschner et al. 2003, van Gelder 2003, Verheij 2007a, Scheuer et al. 2010

Het blijkt dat er in het algemeen geen eenduidige manier is om zo aanvalsgrafen te evalueren. In vet staan in figuur 15 (links) de vier originele semantieken en hun relaties zoals onderscheiden door de uitvinder van abstracte argumentatie Phan Minh Dung: de stabiele, geprefereerde, gegronde en complete semantiek. In de tijd van mijn promotieonderzoek ontdekte ik daar nog de semi-stabiele en de stadiumsemantiek bij. Dat



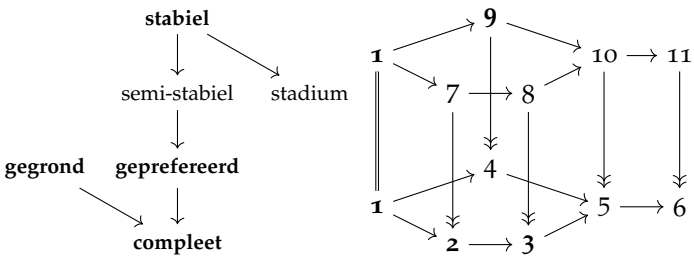
Figuur 13: Evaluatie van aanvalsgrafen



Figuur 14: Het argumentatievoorbeeld in ArguMed

zijn al zes mogelijkheden. Als behalve aanvallen door argumenten ook ondersteuning is toegestaan zijn er nog meer mogelijkheden, in figuur 15 (rechts) zijn het er 11.

Het gekke is nu dat het bij échte argumentatie nooit gaat over de vraag of de semantiek gegrond, compleet, geprefereerd of stabiel is. De vraag is daarmee of deze wiskunde die weliswaar mooi en interessant is, wel precies past bij het behandelde verschijnsel, namelijk argumentatie. Anders gezegd: is het wel de goede wiskunde?



Figuur 15: Argumentatiesemantieken. Links: aanvankelijk **vier**, later zes semantieken (voor grafen met alleen aanval, Dung 1995, Verheij 1996b). Rechts: elf semantieken, hier alleen genummerd weergegeven (voor grafen met zowel ondersteuning als aanval, Verheij 2003b). De nummers **1**, **2**, **3** en **9** corresponderen respectievelijk met de stabiele, semi-stabiele, geprefereerde en stadiumsemantieken uit de figuur links.

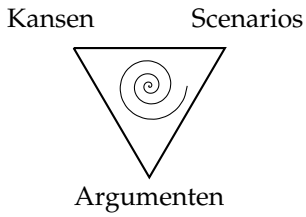
Zoektocht naar de goede wiskunde

Om de goede wiskunde op het spoor te komen zijn mijn collega's en ik gaan kijken naar goed redeneren met forensisch bewijs in het strafrecht. Er blijkt daarbij iets geks aan de hand te zijn. Er zijn namelijk drie theoretische stromingen om redeneren met bewijs te ordenen en te evalueren (figuur 16).[16] De eerste stroming is gebaseerd op argumentatie. Daarvan zagen we net al een voorbeeld toen we het hadden over een getuige die een alibi verschafte aan de verdachte, maar die diens partner blijkt te zijn. Bij argumentatieve analyse is het verzamelen en afwegen van argumenten en tegenargumenten van belang.

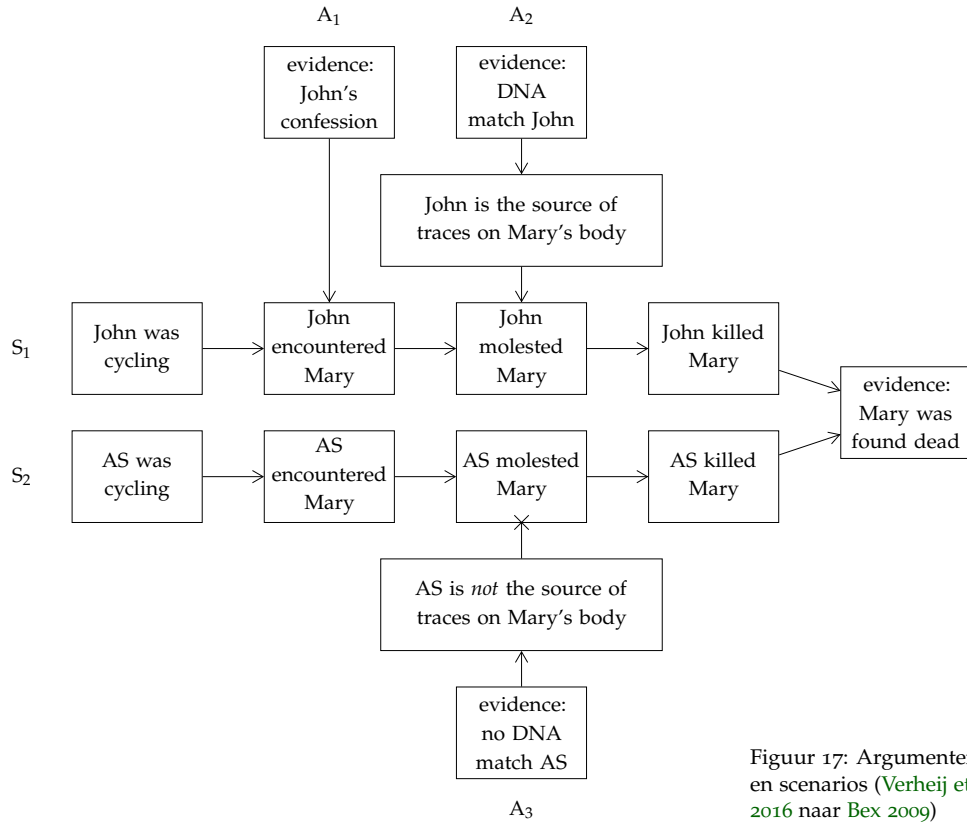
De tweede stroming gebruikt scenarios. Bij een scenarioanalyse van bewijs worden verschillende scenarios geconstrueerd en vergeleken in relatie tot het bewijs. Het alibiscenario kan bijvoorbeeld vergeleken worden met het schuldscenario zoals dat door het openbaar ministerie wordt gepresenteerd. Bij scenarioanalyse speelt de onderlinge samenhang, de coherentie van de scenarios een rol. Een moordscenario zonder motief of zonder moordwapen is bijvoorbeeld niet compleet.

De derde stroming is gebaseerd op kansrekening. Daar kun je vandaag de dag bij bewijs in het strafrecht niet omheen door de belangrijke rol van DNA-bewijs, waarvan de zeldzaamheid statistisch wordt geanalyseerd. De kansen die bij een goed spoor worden gerapporteerd zijn zó miniem dat het welhaast geen toeval kán zijn als een gevonden DNA-spoor past bij het DNA van de verdachte. Redeneren met kansen wordt gezien als een geduchte bron van denkfouten. Met collega's zijn we de

[16] Zie Anderson et al. 2005, Kaptein et al. 2009, Dawid et al. 2011, Di Bello and Verheij 2018. Voor specifieke bijdragen zie Wigmore 1913, Tribe 1971, Bennett and Feldman 1981, Kaye 1986, Thompson and Shumann 1987, Crombag et al. 1992, Pennington and Hastie 1993a,b, Wagenaar et al. 1993, Kadane and Schum 1996, Cook et al. 1998, Schum and Starace 2001, Thagard 2004, Zabell 2005, Keppens and Schafer 2006, Taroni et al. 2006, Hepler et al. 2007, Mortera and Dawid 2007, Pardo and Allen 2008, Tillers 2011, Fenton 2011, Keppens 2012, Fenton et al. 2013



Figuur 16: Drie hulpmiddelen om bewijs te ordenen en evalueren: argumenten, scenarios en kansen



Figuur 17: Argumenten en scenarios (Verheij et al. 2016 naar Bex 2009)

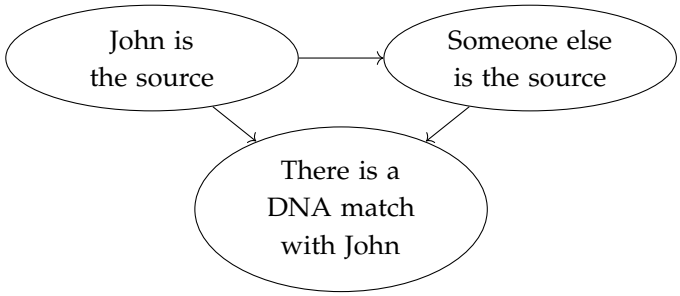
relaties tussen de drie stromingen gaan onderzoeken.

Eerst hebben we dankzij steun van NWO nagedacht over de relaties tussen argumenten en scenarios.[17] Dat leidde tot het proefschrift van Floris Bex,[18] onder begeleiding van Henry Prakken, Peter van Koppen en mijzelf. Daarin wordt een nieuwe, formeel uitgewerkte theorie op de relaties tussen argumenten en scenarios ontwikkeld. Hier zijn—horizontaal—twee scenarios over wat er in een misdrijf gebeurd kan zijn afgebeeld en—verticaal—de argumenten voor en tegen de onderdelen van die scenarios (figuur 17).

[17] Bex et al. 2007, 2010, Bex and Verheij 2012, 2013

[18] Bex 2009

Daarna zijn we ons dankzij een subsidie in het NWO Forensic Science programma gaan richten op kanstheoretisch modelleren. We hebben gekeken naar Bayesiaanse



John is the source

John is the source = false	8000/8001
John is the source = true	1/8001

Someone else is the source

John is the source	false	true
Someone else is the source = false	0	1
Someone else is the source = true	1	0

There is a DNA match with John

John is the source	false		true	
Someone else	false	true	false	true
DNA match = false	0.5*	$1 - 0.66 \cdot 10^{-21}$	0	0.5*
DNA match = true	0.5*	$0.66 \cdot 10^{-21}$	1	0.5*

Figuur 18: Een Bayesiaans netwerk (Verheij et al. 2016)

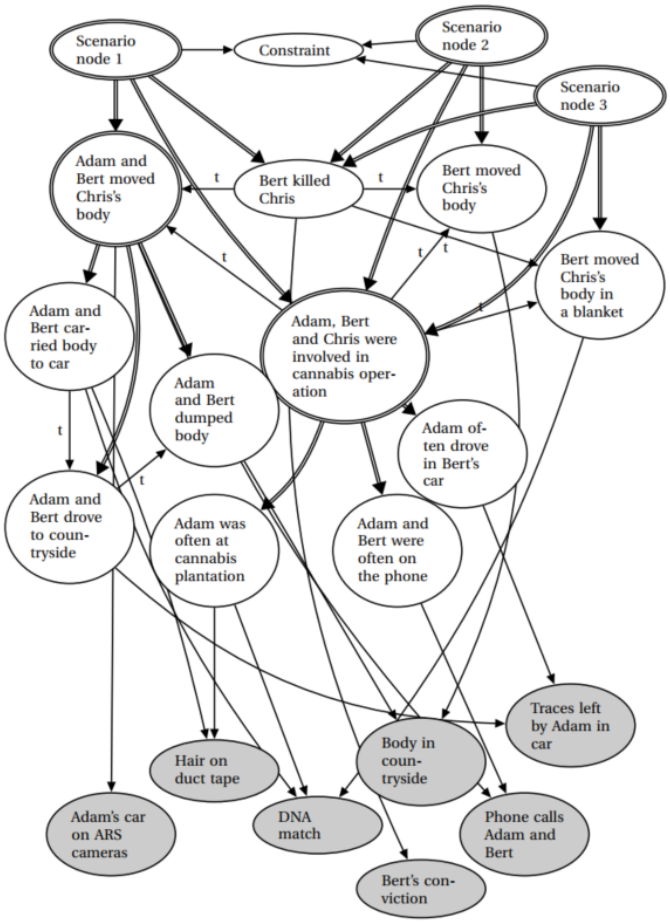
netwerken, een nuttige modelleertechniek in de kunstmatige intelligentie die een grafische netwerkstructuur koppelt aan kansen.[19]

[19] en.wikipedia.org/wiki/Bayesian_network

In figuur 18 een voorbeeld met drie knopen en de bijbehorende tabellen met kansen en conditionele kansen. Een belangrijke eigenschap van Bayesiaanse netwerken is dat kansfuncties er efficiënt mee gemodelleerd kunnen worden door gebruik te maken van onafhankelijkheden tussen variabelen.

Charlotte Vlek onderzoekt in haar proefschrift de relaties tussen scenarios en kansen.[20] Ze heeft onder andere een methode ontwikkeld om de scenarios over wat er in een strafzaak gebeurd is in een Bayesiaans

[20] Vlek 2016



Figuur 19: Scenarios en kansen (Vlek 2016)

netwerk te modelleren (figuur 19).[21] In de figuur worden scenarios gerepresenteerd door clusters knopen. Het bewijs is zichtbaar in grijs. Ze heeft ook laten zien hoe je een Bayesiaans netwerk met scenarios kunt uitleggen.[22]

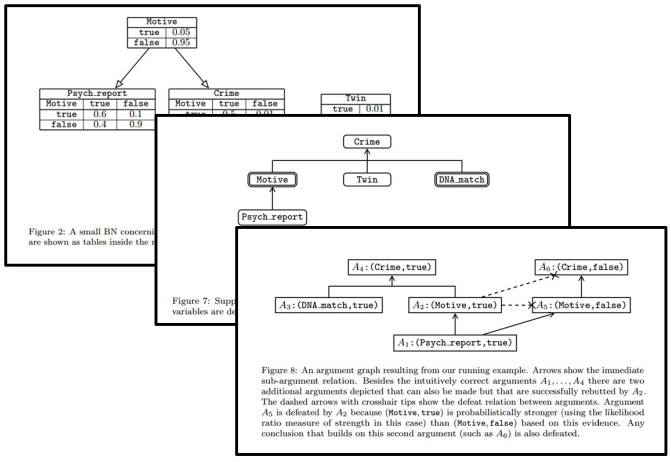
Sjoerd Timmer bestudeert in zijn proefschrift de relaties tussen argumenten en kansen.[23] Hij heeft onder andere een algoritme ontwikkeld dat argumenten en tegenargumenten uit een Bayesiaans netwerk kan genereren (figuur 20).[24] In de figuur is op de achtergrond een Bayesiaans netwerk zichtbaar, in het midden de eruit gegenereerde ondersteuningsgraaf en op de voorgrond

[21] Vlek et al. 2014

[22] Vlek et al. 2016

[23] Timmer 2017

[24] Timmer et al. 2017



Figuur 20: Argumenten en kansen (Timmer 2017)

de daarop gebaseerde argumenten. Hij heeft ook laten zien hoe je argumentatieschema's[25] in een Bayesiaans netwerk kunt modelleren.

[25] Walton et al. 2008, Garssen 2001

Deze twee proefschriften kwamen tot stand in een vruchtbare samenwerking met Henry Prakken, Silja Renooij, John-Jules Meyer en Rineke Verbrugge. En Floris Bex die ik al noemde was als adviseur ook steeds dichtbij. Het werken met de kunstmatige-intelligentietechniek van Bayesiaanse netwerken levert zo veel inzicht op over de relaties tussen het gebruik van argumenten, scenarios en kansen als hulpmiddelen om bewijsredeneringen te evalueren. Ons team drukt zo in internationaal verband zijn stempel op de discussie over veilige manieren van bewijsredeneren met argumenten, scenarios en kansen.

Over het vangen van een dief

De pogingen om argumenten, scenarios en kansen aan elkaar te koppelen leverden uiteindelijk een nieuw type wiskunde voor argumentatie op. De eerste versie van dat formalisme—ontwikkeld onder de Californische zon—bestaat uit een grafische 'taal' waarin het kritische proces wordt weergegeven van de constructie, toetsing en waardering van hypothesen over wat er gebeurt is in een strafzaak.[26] U herkent de woorden waarmee ik net argumentatiesystemen definieerde.

[26] Verheij 2014

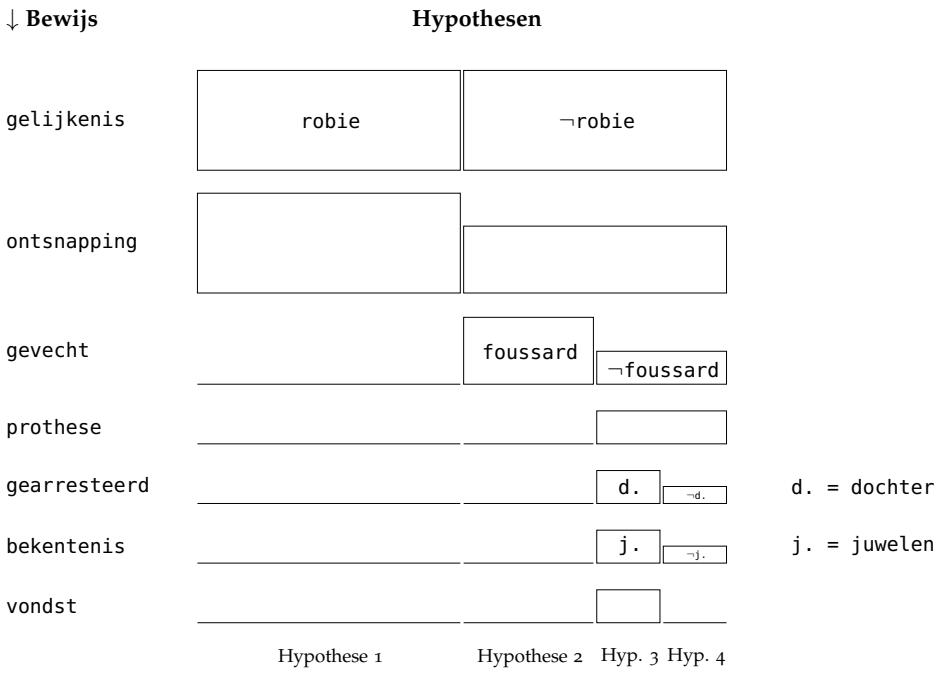
Als voorbeeld bespreek ik het misdaadverhaal uit Alfred Hitchock's mooie film 'To Catch A Thief' uit 1955 dat zich afspeelt in het zuiden van Frankrijk. Robie—gespeeld door Cary Grant—is een voormalige dief die beroemd en berucht is door zijn acrobatische klauterpartijen.

Op elke regel in figuur 21 wordt nieuw bewijs verzameld, worden hypothesen geconstrueerd en getoetst, geselecteerd. De blokken stellen de mogelijke hypothesen voor, en op elke regel blijft er minder van de ruimte aan mogelijkheden over.

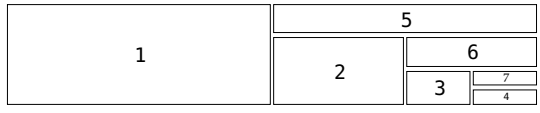
Bovenaan worden twee hypothesen geconstrueerd: het blok links dat Robie net als vroeger spectaculaire diefstallen pleegt, het blok rechts dat hij niet de dief is. Bewijs daarvoor is de gelijkenis van een nieuwe reeks diefstallen met Robie's acrobatische stijl van jaren terug. Eerst heeft de politie geen voorkeur voor één van de hypothesen. Daarom zijn de twee blokken op de eerste regel even groot. Op de tweede regel is het blok rechts dat Robie's onschuld representeert kleiner want na zijn ontsnapping aan de politie is dat minder geloofwaardig geworden. Na een nachtelijke hinderlaag onstaat een gevecht, waarbij Foussard, een vriend uit de tijd van de *résistance*, van het dak valt en sterft.

Gevolg is dat Robie niet meer wordt verdacht, en op de derde regel is daarom het blok voor Robie's schuld verdwenen. Nu is Foussard de verdachte, wel met de mogelijkheid dat hij onschuldig is. Al snel vervalt de verdenking als bedacht wordt dat Foussard met zijn prothese nooit de benodigde capriolen kan hebben begaan. Later in de film wordt Foussard's dochter op heterdaad betrapt en gearresteerd. In haar bekentenis vertelt ze waar de gestolen juwelen gevonden kunnen worden. Na de vondst van de juwelen op de betreffende plek twijfelt niemand er meer aan dat zij de dief is. Op de onderste regel is dan ook nog maar één blok over dat de schuld van Foussard's dochter voorstelt.

De blokken representeren alle mogelijkheden, alle



Figuur 21: Model van het misdaadverhaal uit Alfred Hitchcock's 'To Catch A Thief' (Verheij 2017a)



Figuur 22: Casusmodel voor Alfred Hitchcock's 'To Catch A Thief' (Verheij 2017a)

gevallen die voor mogelijk worden gehouden. Dat wordt extra duidelijk als we de blokken op alle regels over elkaar heen schuiven. Dan krijgen we het model in figuur 22 bestaand uit 7 blokken, ieder een mogelijkheid representerend van wat er zou kunnen zijn gebeurd. Blok 1 staat voor de mogelijkheid dat Robie inderdaad weer dief was geworden. Blok 3 staat voor de mogelijkheid die uiteindelijk wordt geloofd—dat de dochter van verzetsvriend Foussard de dief is.

De relatieve grootte van de blokken stelt hun relatieve waarschijnlijkheid, hun relatieve geloofwaardigheid, voor. In het verhaal worden deze mogelijkheden langzaamaan opgebouwd, steeds met onzekerheid over wat geloofd moet worden. Totdat uiteindelijk elke onzekerheid is weggenomen door overduidelijk bewijs. Er is dan nog maar één blok over. Althans: elke redelijke onzekerheid is weggenomen. Er is namelijk aan het eind geen concrete reden voor onzekerheid meer over. Binnen het in een kritisch proces opgebouwde model is er geen twijfel meer mogelijk, is het zeker dat Foussard's dochter de dief is. Alleen een vergroting van de voorstelbare wereld van mogelijkheden kan daar verandering in brengen.

Een drietal in deze grafische taal verwerkte inzichten zijn technisch van belang.

- Ten eerste kan de geldigheid van argumenten worden 'afgelezen' uit de verzameling mogelijke hypothesen. Het formalisme geeft zo een semantiek voor argumenten en tegenargumenten.
- Ten tweede zijn er—in tegenstelling tot bij Bayesiaanse netwerken—weinig getallen nodig, want de aanpak werkt vooral met de relatieve verhouding, de ordening van de getallen.
- Ten derde en tot slot is er een natuurlijke koppeling met de logica en de kansrekening. En dat terwijl zoals ik al vertelde zowel informeel als formeel argumentatieonderzoek—ook het mijne—vaak is geïnspireerd door contrasten met logica en kansrekening.

Terug naar de wiskundige grondslagen van argumentatie

De formele uitwerking van deze grafische taal heeft geleid tot een formalisme waarin gevallen, in het Engels 'cases' centraal staan. De blokken uit de figuren corresponderen met gevallen. Elk geval representeert een mogelijkheid, een cluster eigenschappen dat samen kan voorkomen.

Hier volgt de wiskundige definitie van gevalsmodellen, in het Engels 'case models'. Gevallen worden gerepresenteerd door logisch consistente, logisch verschillende zinnen, die paarsgewijs incompatibel zijn. Hun ordening is een totale preordening, oftewel totaal en transitief, maar niet per se antisymmetrisch.

Definition (Verheij 2016b,a, 2017a)**.** *A* case model *is a pair* $(C, \geq)$ *with finite* $C \subseteq L$*, such that the following hold, for all* φ*,* ψ *and* $\chi \in C$*:*

1. $\not\models \neg\varphi$*;*
2. *If* $\not\models \varphi \leftrightarrow \psi$*, then* $\models \neg(\varphi \wedge \psi)$*;*
3. *If* $\models \varphi \leftrightarrow \psi$*, then* $\varphi = \psi$*;*
4. $\varphi \geq \psi$ *or* $\psi \geq \varphi$ *;*
5. *If* $\varphi \geq \psi$ *and* $\psi \geq \chi$*, then* $\varphi \geq \chi$*.*

Gevalsmodellen kunnen gebruikt worden om drie typen logische geldigheid van argumenten mee te definiëren. Coherente argumenten ondersteunen een mogelijk geval; presumptieve, 'veronderstellende' argumenten ondersteunen een maximaal geprefereerd geval; en conclusieve, 'beslissende' argumenten laten bovendien geen coherente weerlegging toe.

Definition (Verheij 2016b,a, 2017a)**.** *Let* $(C, \geq)$ *be a case model. Then we define, for all* φ, ψ *and* $\chi \in L$*:*

1. $(C, \geq) \models (\varphi, \psi)$ *if and only if* $\exists\omega \in C$*:* $\omega \models \varphi \wedge \psi$*.*

   We then say that the argument from φ *to* ψ *is* coherent *with respect to the case model.*

2. $(C, \geq) \models \varphi \Rightarrow \psi$ *if and only if* $\exists \omega \in C$*:* $\omega \models \varphi \wedge \psi$ *and* $\forall \omega \in C$*: if* $\omega \models \varphi$*, then* $\omega \models \varphi \wedge \psi$*.*

   We then say that the argument from φ *to* ψ *is* conclusive *with respect to the case model.*

3. $(C, \geq) \models \varphi \rightsquigarrow \psi$ *if and only if* $\exists \omega \in C$*:*

   (a) $\omega \models \varphi \wedge \psi$*; and*

   (b) $\forall \omega' \in C$ *: if* $\omega' \models \varphi$*, then* $\omega \geq \omega'$*.*

   We then say that the argument from φ *to* ψ *is* presumptively valid *with respect to the case model. Such an argument is* properly defeasible*, when it is not conclusive. Circumstances* χ *are* defeating *or* successfully attacking *when* $(\varphi \wedge \chi, \psi)$ *is not presumptively valid. Defeating circumstances are* rebutting *when* $(\varphi \wedge \chi, \neg\psi)$ *is presumptively valid; otherwise they are* undercutting*. Defeating circumstances are* excluding *when* $(\varphi \wedge \chi, \psi)$ *is not coherent.*

Wiskundig precies en filosofisch relevant is dat totale preordeningen die ordeningen zijn die numeriek gerealiseerd kunnen worden. Daarmee zijn het de ordeningen die tegelijk kwalitatief en kwantitatief zijn; ze zijn tegelijk met en zonder getallen (vgl. de titels van Verheij 2014, 2017a). En dus is de preferentieordening van gevalsmodellen ook numeriek realiseerbaar en het blijkt dat dat zelfs compatibel met de kansrekening kan.

Zo kunnen de drie definities van typen geldige argumenten ook kwantitatief worden herschreven. Coherente argumenten corresponderen met een positieve conditionele kans van conclusies gegeven premissen; presumptieve argumenten met een kans groter dan een drempelwaarde; en conclusieve argumenten met een kans gelijk aan 1. Op deze manier slaat argumentatie een formele brug tussen de logica en de kansrekening. De logica beschrijft de eigenschappen van de gevallen, en de kansrekening hun preferentieordening.

ZOALS GEZEGD IS de wiskunde van gevalsmodellen ontstaan door na te denken over correct redeneren met forensisch bewijs, en dan vooral door de puzzel hoe argumenten, scenarios en kansen samengaan zonder elkaar in de weg te zitten. De spannende vraag is nu of de geldigheid van een complexe argumentstructuur ook met gevalsmodellen is te reconstrueren. Preciezer: gegeven een regelstructuur van regels met hun uitzonderingen, is er dan een gevalsmodel waarin die regels geldig zijn?

Dat blijkt te kunnen. Een goed voorbeeld kan gegeven worden door te kijken naar de relaties tussen argumenten, casus en regels in het recht,[27] een onderwerp waar ik al eerder naar had gekeken met de eerste promovendus die ik mocht begeleiden, Bram Roth, eerst met Jaap Hage, later met Hans Crombag.[28] Eerder zegen we de formele basisstructuur van het Nederlandse recht over schadevergoedingen door onrechtmatige daad (figuur 5, blz. 16) die ik al gebruikte om kennissystemen te illustreren. En hier ziet u een gevalsmodel—bestaande uit 16 gevallen—waarin deze structuur geldig is (figuur 23). Het bijzondere is dat een verzameling casus met ogenschijnlijk weinig structuur de complexe argumentstructuur geldig maakt.

Zo is de kennis zoals die is gerepresenteerd in de argumentstructuur van figuur 5 (blz. 16) gegrond in de data van de verzameling casus uit figuur 23.

Argumenten, gevallen en regels

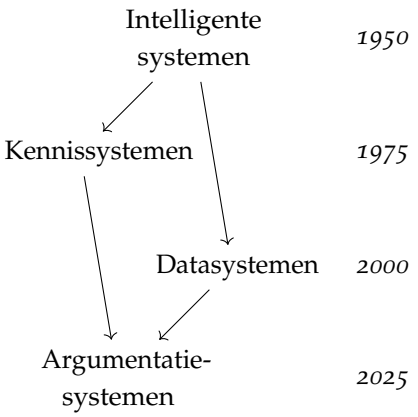
[27] Een kernthema in het veld AI & Law, zie Bench-Capon et al. 2012 voor een historisch perspectief en Gardner 1987, Rissland and Ashley 1987, Ashley 1990, Branting 1991, Berman and Hafner 1995, Gordon 1995, Loui and Norman 1995, Aleven and Ashley 1995, Prakken and Sartor 1996, Hage 1997, McCarty 1997, Prakken and Sartor 1998, Bench-Capon and Sartor 2003 voor specifieke bijdragen

[28] Roth 2003, Roth and Verheij 2004

Figuur 23: Gevalsmodel voor de onrechtmatige daad (Verheij 2017b)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	¬dut	¬dut	¬dut	dut	dut	dut	dut	dut	dut	dut	dut	dut	¬dut	¬dut	¬dut
¬dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg
	¬unl	unl	unl	unl	unl	unl	unl	unl	unl	unl	unl	unl	¬unl	¬unl	unl
		¬imp	imp	imp	imp	imp	imp	imp	imp	imp	imp	imp			imp
			¬cau	cau	cau	cau	cau	cau	cau	cau	cau	cau			cau
	¬vrt			vrt	vrt	vrt	¬vrt	¬vrt	¬vrt	¬vrt	¬vrt	¬vrt	vrt	¬vrt	
	¬vst			¬vst	¬vst	¬vst	vst	vst	vst	¬vst	¬vst	¬vst	¬vst	vst	vst
	¬vun			¬vun	¬vun	¬vun	¬vun	¬vun	¬vun	vun	vun	vun			
		¬ift		ift	¬ift	¬ift	ift	¬ift	¬ift	ift	¬ift	¬ift			
		¬ila		¬ila	ila	¬ila	¬ila	ila	¬ila	¬ila	ila	¬ila			
		¬ico		¬ico	¬ico	ico	¬ico	¬ico	ico	¬ico	¬ico	ico			
				¬jus	¬jus	¬jus	¬jus	¬jus	¬jus	¬jus	¬jus	¬jus	jus	jus	
				prp	prp	prp									¬prp

1 > 2 > 3 > 4 > 5 ~ 6 ~ 7 ~ 8 ~ 9 ~ 10 ~ 11 ~ 12 ~ 13 > 14 ~ 15 ~ 16



Figuur 24: Voorbij de splitsing in de kunstmatige intelligentie

Argumenten overbruggen de kloof tussen kennis- en datasystemen

NU ZIJN WE TERUG bij de overbrugging van de kloof in de kunstmatige intelligentie, de kloof tussen kennis- en datasystemen (besproken op blz. 15). Want het is geen grote stap om de gevallen in een gevalsmodel te zien als de data waarvan geleerd kan worden—zoals bij datasystemen—en de regels die er in gelden als de kennisstructuur van kennissystemen. De ontwikkelde wiskunde kan daarom de grondslag zijn voor argumentatiesystemen die de brug slaan tussen kennis en data (figuur 24).

Zulke argumentatiesystemen kunnen uitgroeien tot de goede kunstmatige intelligentie die ontwikkeld moet worden: intelligente systemen die de *goede antwoorden* geven, *goede redenen* hebben en *goede keuzes* maken.

1. Ten eerste de goede antwoorden. In een argumentatiesysteem wordt in een kritische discussie gezocht naar het goede antwoord bij een complex probleem—en als dat niet te vinden is naar het beste antwoord. Dat gaat verder dan het maximaliseren van het aantal goede antwoorden zoals we dat kennen van datasystemen. Tegelijk kan de kennis in een argumentatiesysteem voortdurend ontwikkeld worden door de constructie van nieuwe hypothesen en de toetsing aan beschikbare data.

2. Dan de goede redenen. Bij argumentatiesystemen staan redenen sowieso voorop. Een argumentatiesysteem kan een waarom-vraag beantwoorden met een geschakeerd lijstje redenen voor en tegen een gegeven antwoord, soms aangevuld met mogelijke alternatieven. Zo geeft de argumentstructuur een inkijkje in de kennisstructuur 'aan de binnenkant' met behoud van de koppeling aan data.

3. Tot slot de goede keuzes. Argumentatiesystemen kunnen zich aan de regels houden, ze kunnen de geldende normen volgen, rekening houdend met de specifieke omstandigheden. Door hun vermogen tot kritische discussie kunnen ze wel weerwoord geven, precies zoals dat gewenst is bij een serieuze discussie over wat de goede keuze is.

Door de ontwikkeling van argumentatiesystemen komen zo de dromen over kunstmatige intelligentie dichterbij, en angsten worden beheersbaarder. Want denk nog eens aan de voorstellen om kunstmatige intelligentie te reguleren door het verplichten van menselijke controle en door het verbieden van 'killer robots'. Mijn voorstel opent een andere route om kunstmatige intelligentie te reguleren want met argumentatiesystemen kan een kritische discussie worden gevoerd op basis van redelijke argumenten. Het winnen van zo'n discussie zal niet altijd makkelijk zijn. We zullen zelfs zo nu en dan verliezen. Wat dat betreft is het niet anders dan wat we gewend zijn in wetenschap, politiek en dagelijks leven. Maar dat is ook de kern van de zaak: het gaat bij intelligent gedrag niet om het winnen van de discussie, maar om het vinden van de goede antwoorden op de lastige problemen die het leven in een complexe, dynamische wereld stelt. Zoals we uit de wetenschap, de politiek en het dagelijks leven weten is een kritische discussie daarvoor onontbeerlijk. De ontwikkeling van goede kunstmatige intelligentie die ons daarbij helpt zal voorlopig nog een boeiende en uitdagende klus zijn die alleen mensen kunnen volbrengen.

Het moge duidelijk zijn dat vóór het zover is—en computers en robots serieuze gesprekspartners zijn in een kritische discussie—nog veel onderzoek nodig is en het is een genoegen om daar hier in Groningen met onze toegewijde staf en ruim 400 studenten kunstmatige intelligentie aan te kunnen werken.

IK HEB GEZEGD en ben benieuwd naar úw argumenten voor goede kunstmatige intelligentie.

Arguments for good artificial intelligence

Dear Rector Magnificus,
highly valued audience,

Dreams and fears

These days newspapers often publish stories about artificial intelligence. And numbers are rising (Figure 1). Fitting the subject these stories are full of dreams and of fears.

Sometimes it seems that computers already are capable of almost anything: if we can believe the headlines, machines can by now play football, drive trucks, and provide advise about jobs.[1]

And today's artificial intelligence does stimulate the imagination. Self-driving cars have been driving on the public road for years; at one after the other classic board game the best human players are beaten by computers; and at work we see the humanoid robots of science fiction stories (Figure 2).

[1] A few recent headlines in the NRC Handelsblad: Soccer robot's kicks now clean and deadly, July 21, 2017; Here comes the robot truck, February 17, 2017; Algorithm advises job seeker about applications, July 27, 2017

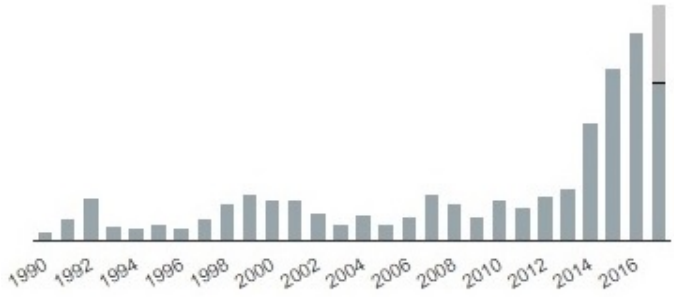


Figure 1: Number of articles with keyword 'artificial intelligence' in the NRC Handelsblad (1990–2017). The final column is an extrapolation of the first 8 months of 2017. Source: www.nrc.nl

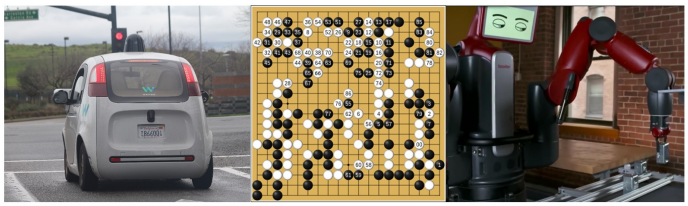


Figure 2: A self-driving car; AlphaGo v. Tang Weixing; robot Baxter. Source: www.wikipedia.org

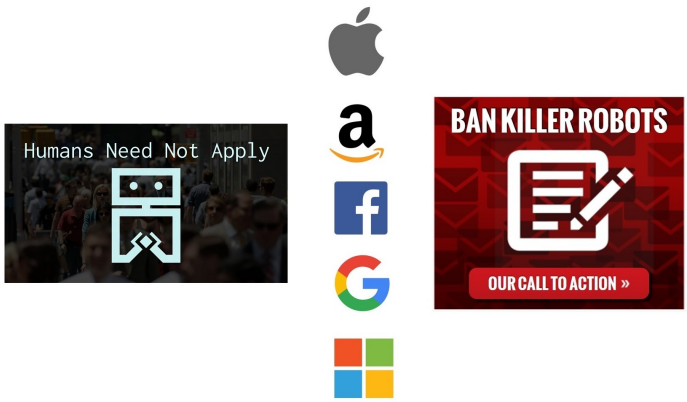


Figure 3: Humans need not apply; the frightful five; ban killer robots. Source: www.wikipedia.org

With the dreams also the fears are growing, because the continuously expanding automation of work that is now still only done by humans will profoundly change the labor market; by automatic analysis of all the information about ourselves that we put online companies and governments gain ever more control over us; and the idea of weapon systems that autonomously decide on whether to go on the attack is altogether scary (Figure 3). Arguments enough to be cautious with artificial intelligence.

Hence it is no surprise that there is ever more attention for the issue how to embed artificial intelligence sensibly in our society. In a recent opinion of the European Economic and Social Committee, an advisory institution of the European Union, a plea is made for a *human-in-command* approach to artificial intelligence, in which machines stay machines and humans will at all

times keep control over these machines.[2] Even more recent is the call in an open letter to prohibit 'killer robots', signed by leaders of over a 100 robotics and artificial intelligence companies, among them Tesla and Google.[3]

I will explain to you that—next to human control and prohibition—a third way exists to embed artificial intelligence sensibly in our society, and that is by the development of good artificial intelligence. The road towards that—so I will argue—is the development of argumentation systems, i.e., systems that can conduct a critical discussion on the basis of arguments.

[2] C. Muller. Artificial Intelligence – the Consequences of Artificial Intelligence for the (Digital) Single Market, Production, Consumption, Employment and Society. *Opinion European Economic and Social Committee*, INT/806, 2017

[3] `futureoflife.org/autonomous-weapons-open-letter-2017`, August 20, 2017

Artificial intelligence now

In this connection it is good to be aware of how things are in artificial intelligence.[4] For that it is useful to distinguish between specialized, general and superior artificial intelligence.

[4] For an introduction, see Russell and Norvig 2010 and also Douglas Hofstadter's works, in particular Hofstadter 1979, 1985, 2007.

- *Specialized artificial intelligence* is the nowadays very common kind of artificial intelligence that can perform specifically delineated intelligent tasks. All artificial intelligence that exists today is specialized. Think of computer programs that have the expertise to turn the completion of a tax return form into child's play and of smartphone apps that are rather good at finding a list of tree pictures in our endless collection of photos. Most researchers work on specialized artificial intelligence.
- *General artificial intelligence* refers to computer programs or machines that manage themselves well under a wide variety of circumstances and with a broad palette of problems; just like what we are used to with humans. They can for instance understand books and also create them, and they can learn how to ride a bike in a busy street, also when not born in the Netherlands. General artificial intelligence does not exist today. Some researchers think about the issue how to arrive at general artificial intelligence, or else why that is not possible.

- *Superior artificial intelligence* is the form of artificial intelligence that worries some people very much. The idea of superior artificial intelligence is that—once we have reached general artificial intelligence—human intelligence will immediately fall behind irrecoverably. Will there still be a place for humans after the invention of superior artificial intelligence? Nobody knows. Many researchers enjoy the exchange of views about superior artificial intelligence at parties and during receptions, but are in their daily work busy addressing scientific hurdles that have to be overcome in their specific specialization.

An important hurdle that must be overcome is the bridging of the gap between knowledge and data systems. In knowledge systems, the knowledge that is needed to perform a complex task is directly entered into a computer; knowledge is represented and is used for automated reasoning. In data systems the handling of a problem is learnt by the automatic analysis of a database of examples.

Initially there was no gap between the two types of intelligent systems (Figure 4). Mid twentieth century when artificial intelligence originated as a research area—the term 'artificial intelligence' was coined in 1955—the programming of computers was new in itself and all possible approaches to artificial intelligence were tried. Gradually the research into knowledge systems—

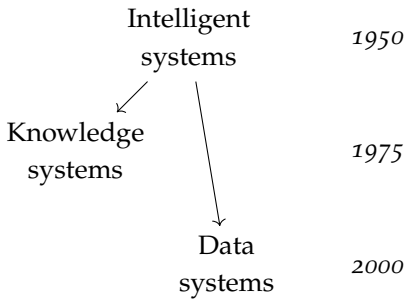


Figure 4: A gap in artificial intelligence

based on representation and reasoning—and into data systems—based on learning from examples—grew apart. Also the mathematics used differs. Knowledge systems research often uses logic; data systems research probability theory.

Here you see a Dutch statutory provision on compensation for damages on the grounds of an unlawful act, for instance when by your fault someone's phone falls and breaks:

> *Article 6:162.1 of the Dutch Civil Code*. A person who commits an unlawful act toward another which can be imputed to him, must repair the damages which the other person suffers as a consequence thereof.

In a representation of this legal doctrine four conditions are distinguished that together determine the duty to repair the damages. IF there are damages AND the act is unlawful AND it is imputable AND there is a causal connection between action and damages THEN there is a duty to repair the damages. You can also see a logically formalised version (the line to the right with three letter abbreviations and symbols), and the arrow diagram (Figure 5) shows a more extensive representation of the doctrine with more rules and exceptions.[5]

```
IF damages
  AND unlawful
  AND imputable
  AND
causal-connection
 THEN duty-to-repair
```

$$\mathrm{dmg} \wedge \mathrm{unl} \wedge \mathrm{imp} \wedge \mathrm{cau} \rightsquigarrow \mathrm{dut}$$

[5] See also Verheij et al. 1997, Verheij 2017b

By filling a computer program with this kind of, often rule-like, represented knowledge a diversity of intelligent systems can be built. Students of artificial intelligence in Groningen have for instance built systems for a general practitioner's triage, poetry classification, a digital reverend—and even the expertise of a herbalist was once captured in a knowledge system.

In data systems the handling of a problem is learnt by the analysis of a database with examples. For instance, at the 'input' side of a neural network all kinds of forms of the capital A are shown (Figure 6), so that gradually the internal structure of the network—here with a 'hidden layer' in the middle—can be adapted to correctly recognize such letters at the 'output' side. For a while it was believed that neural networks had fundamental

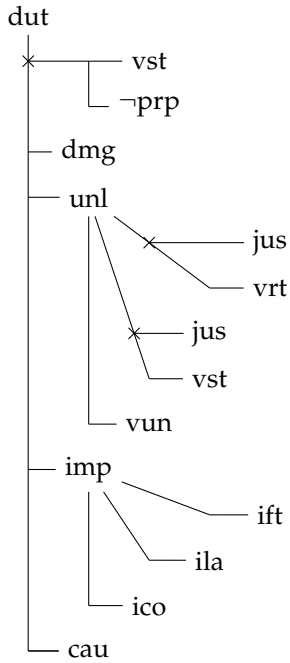


Figure 5: Argument structure for Dutch law of unlawful acts

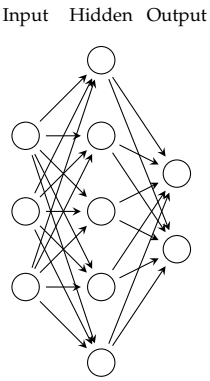


Figure 6: Forms of the capital 'A' (Hofstadter 1995); structure of a neural network

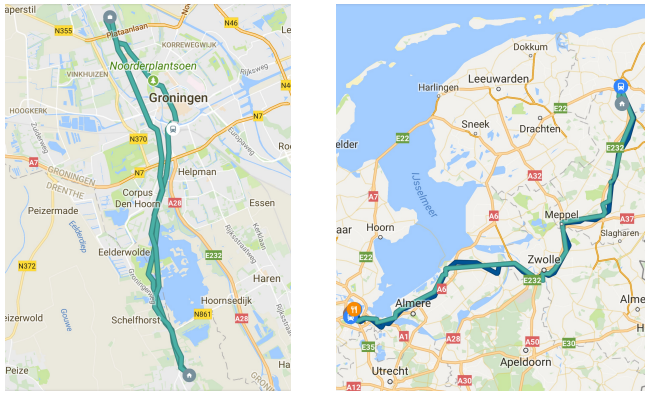


Figure 7: From Paterswolde to Groningen and back; and from Groningen to Amsterdam and back. Source: Google Maps, timeline

limitations, while in fact these have grown to become a very powerful data analysis technique.

Good artificial intelligence

I ALREADY STATED that we need to develop good artificial intelligence. I then think of three necessary characteristics:

- First an intelligent system must be able to provide *good answers* to problems;
- second an intelligent system must be able to provide *good reasons* for these; and
- third an intelligent systems must be able to make *good choices*.

First about the good anwers. Data systems are not designed to provide good answers, but to give good answers *as often as possible*, and in all kinds of domains they are very good at that. Regularly they give good answers more often than humans do. Such performance is impressive. At the same time data systems make mistakes that a human will not quickly make.

An example. Since I usually carry my phone, Google can track my whereabouts well. Here is for instance a Google picture of my daily commute from home to work and back (Figure 7, left). The route has a green color,

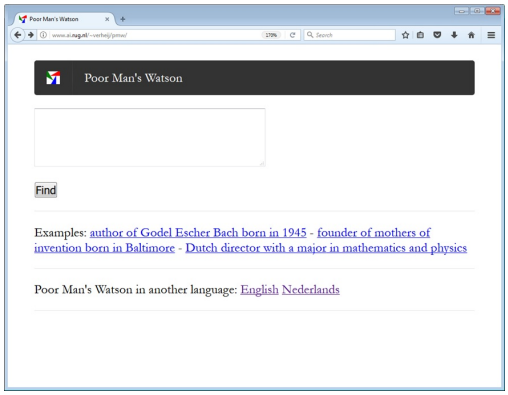


Figure 8: Poor Man's Watson. Bron: www.ai.rug.nl/∼verheij/pmw/

meaning that this is a biking trip, and indeed Google's data system has correctly assumed that I was on my bike.

Now look at the picture on the right of a trip from Groningen to Amsterdam and back. Google has colored the outward journey blue—which is correct as I went by train. But the inward journey is green, so Google thinks that I biked back—but such a 190 km biking trip seems a bit too much for a normal working day. One would like to correct the data system and say: 'People do not bike from Amsterdam to Groningen in two hours'. But that is not how things work in data systems. Such a correction is possible in a knowledge system.

Then the good reasons. A data system cannot be asked why it came to a certain answer, because a data system does not have explicit reasons. All data as a whole supports the answer. That my train trip was by accident classified as a biking trip was the result of the analysis of the available data, about me and about others. According to that data analysis my train trip was a bit more like the biking trips in the dataset than like the train trips. On what exactly the outcome is based is hard to determine. Data systems are therefore considered to be 'black boxes'. One cannot look well into the insides of a data system, its inner workings are not transparent.

An example is my 'Poor Man's Watson' (Figure 8). It is inspired by IBM's Watson that in 2011 won the

American quiz show Jeopardy! against the best human players.[6] That was the spectacular outcome of a significant research effort, while 'Poor Man's Watson' is the result of an afternoon's programming by a single researcher; hence the name. I built it in particular to show what is possible with a simple script using Google and Wikipedia and thereby understand better how special the Jeopardy! result was. And in some respects 'Poor Man's Watson' already seems quite intelligent.

[6] en.wikipedia.org/wiki/Watson_(computer)

Asked for the author of 'Gödel, Escher, Bach' born in 1945 it properly answers 'Douglas Hofstadter'. And also the answer to the 'founder of the mothers of invention born in Baltimore' is correct. At least that was the case when I tested yesterday. Because the system is connected to the internet and depends on the information that can be found there, anything can happen. Sometimes something that went well for a long time suddenly stops working. For a while 'born in 1940' also worked well for the Baltimore born Frank Zappa but that no longer works. The system cannot be asked what the reason for this is since the system does not have reasons. For the trip example one would like to ask the data system: 'Why do you think I biked from Amsterdam to Groningen?' But a data system does not have an answer to such a question. A knowledge system would have.

Third the right choices. Data systems don't choose themselves. They are descriptive by their nature. They describe the data on which they are based. That objectivity is at the same time their strength and their limitation. When the data is good, the choices are good. But the data isn't always good. A well-known example is Tay, a chatbot developed by Microsoft.[7] Tay went online on June 23, 2016. Tay learnt from the examples of conversations with users. And hence—internet is internet—dialogues with Tay grew ever stranger, more unpleasant and sometimes offensive, because that is what internet dialogue is often like. After 16 hours Microsoft took Tay offline.

[7] en.wikipedia.org/wiki/Tay_(bot)

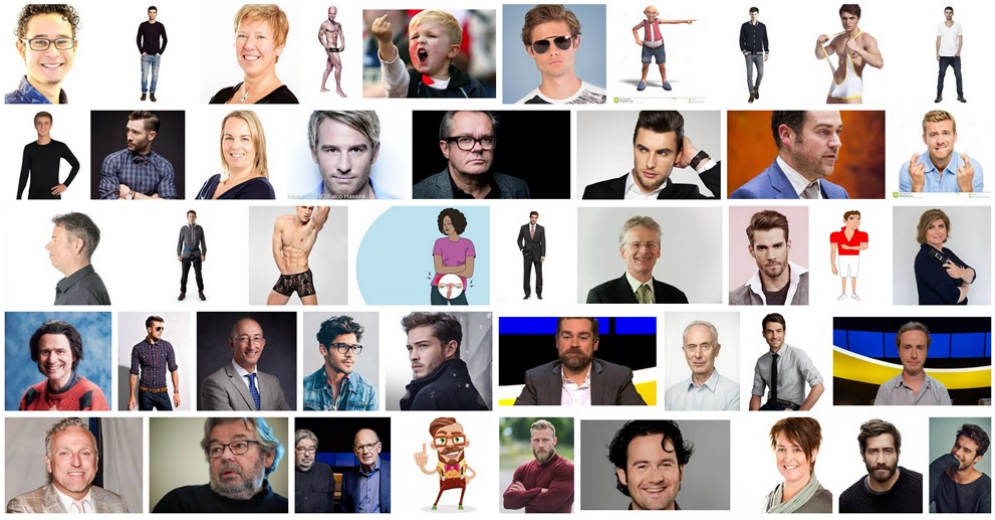


Figure 9: Human beings according to Google. Source: Google images, query 'mens' ('human being' in Dutch)

Another example is the alleged racism of search engines that is in the news every now and then. I was curious what the situation was, typed 'human being' the other day and indeed saw many pictures of human beings (Figure 9). Out of the 36 human beings on my screen four were female, one of them as a drawing. The drawn woman was the only clearly non-white human being. Apparently this is the image of human beings as it emerges from Google's data.

The lesson is clear: describing data is not enough to understand a complex world. Very often one has to move away from the data to achieve the desired effects. One actually would like to educate a data system and say: 'Would you please stop now with these insulting remarks?' But a data system cannot follow such advice. A knowledge system can.

Argumentation systems

KNOWLEDGE SYSTEMS CAN give good answers, can provide reasons and can make good choices. Exactly the properties of good artificial intelligence that I mentioned. But knowledge systems are not good at what data sys-

tems are good at. It has been known for long: knowledge systems have not been designed to learn from data, knowledge systems are hard to scale, and knowledge systems cannot handle visual and other geometrically structured information. And data systems are often good at all that.

What is hence needed is a combination of the good properties of knowledge systems and of data systems. The only problem is that we as yet do not know well how the different techniques used go together.

To make progress with that I propose to develop argumentation systems, that is systems that can conduct a critical discussion based on arguments. Today I use this definition:

```
Argumentation systems are systems that can conduct
a critical discussion in which hypotheses can be
constructed, tested and evaluated on the basis of
reasonable arguments.
```

An example of argumentation. When a witness has testified that the suspect was somewhere else during the crime, that gives a reason that he has an alibi, which on its turn is a reason for the innocence of the suspect. If that is the only information, we can believe in the suspect's innocence (Figure 10, left). But next it can turn out that the witness is the suspect's partner, which is a reason that the testimony is a lie, which in turn is a reason to not believe the witness about the alibi (Figure 10, right). The example shows that more information can change the perspective. At first we believe in the

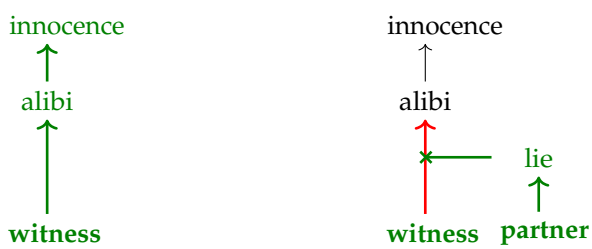


Figure 10: Argumentation

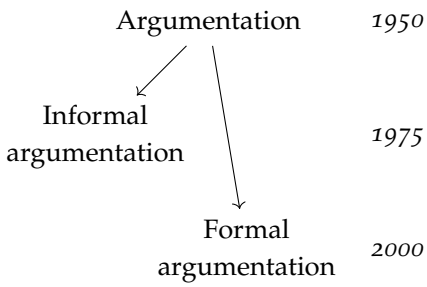


Figure 11: A gap in argumentation research

suspect's innocence, later doubt creeps in and wins. In formal terms we here speak of a non-monotonic logical system.[8]

Argumentation has been studied since antiquity and became a fertile research area in the middle of the twentieth century—accidentally in the same period as artificial intelligence.[9] While there the powers of mathematical, formal methods were exploited, the discussion in argumentation theory often focused on the issue that the formal methods of that period were not suitable for understanding argumentation as it appears 'in the wild' (so in politics, in courts or at home).[10] Not accidentally for the flow of this storyline these formal methods are by the way the same as the ones I discussed about the distinction between knowledge and data systems: logic and probability theory. A gap emerged between formal and informal argumentation research (Figure 11).[11]

Since the seventies informal argumentation theory is well on track, and for formal argumentation theory that is true especially since the beginning of this century. Slowly but surely formal and informal argumentation theory are again growing closer to one another—especially also influenced by research into argumentation systems in artificial intelligence.[12]

I will show you some recent developments using the mathematical foundations of argumentation, correct reasoning with forensic evidence and the connections between rules and cases in the law.

[8] Reiter 1980, Pollock 1987, 1995, Gabbay et al. 1994

[9] See van Eemeren et al. 2014

[10] Toulmin 1958. See also Hitchcock and Verheij 2006, Verheij 2009

[11] The gap is not always sharp, though. See e.g. Barth and Krabbe 1982, Freeman 1991, Walton and Krabbe 1995, Reed and Grasso 2007

[12] See Chapter 11 in van Eemeren et al. 2014 and Chesñevar et al. 2000, Reed and Norman 2004, Bench-Capon and Dunne 2007, Rahwan and Simari 2009, Atkinson et al. 2017. See also Bondarenko et al. 1997, Vreeswijk 1997, Grasso et al. 2000, Rahwan et al. 2003, García and Simari 2004, Chesñevar et al. 2006, Amgoud and Caminada 2007, Besnard and Hunter 2008, Modgil 2009, Brewka and Woltran 2010, Prakken 2010, Thimm 2012, Brewka et al. 2013, Baroni et al. 2014, Cerutti et al. 2017

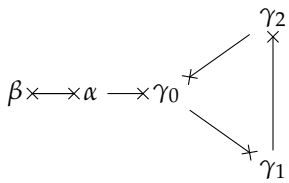


Figure 12: Abstract argumentation as a form of graph theory

Mathematical foundations of argumentation

As a start the mathematical foundations of argumentation.[13] Both in the informal and in the formal literature much attention has been paid to the structure of argumentation. We already saw an example of that when discussing the argumentation about the witness testimony.

[13] Simari and Loui 1992. See Baroni et al. 2018

It turns out that the mathematics of the evaluation of arguments that attack each another and that can be defended by counterattacks is surprisingly varied and interesting. Since the mid nineties this mathematics is studied as a form of graph theory (Figure 12).[14] In the figure five arguments are shown and the arrows indicate how the arguments attack one another. For instance the arguments α and β attack each other and argument α also attacks γ_0.

[14] Dung 1995

The evaluation of attack graphs is based on the idea that arguments not only attack but also defend one another. When argument α is not attacked, the argument is undefeated and ends up as a winner in an argumentative debate (as indicated by the color green in Figure 13). When argument β attacks α, the attacker β wins and α loses (as indicated by the color red). When β on its turn is attacked by argument γ, then β loses. Argument α no longer loses by the successful defense against β by γ.

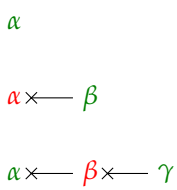


Figure 13: Evaluation of attack graphs

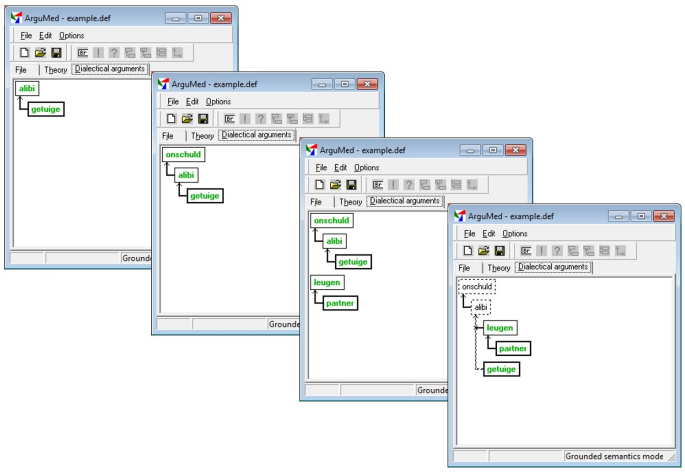


Figure 14: The argumentation example in ArguMed

One can say that the argument has been reinstated.

The thus developed mathematics can serve as the foundations for the design of computer programs for the construction and evaluation of argumentation. For instance in Figure 14 the example I just gave is constructed and evaluated in my ArguMed program.[15] A formal characteristic of this software is that the graphical structure of the diagram is isomorphic to the logical structure of the mathematics. An arrow in a diagram corresponds to a conditional sentence of the logic.

[15] Verheij 2003a, 2005a. See also Kirschner et al. 2003, van Gelder 2003, Verheij 2007a, Scheuer et al. 2010

It turns out that in general there is no unique way of evaluating attack graphs. Figure 15 (left) shows, in bold, the four original semantics and their relations as they were distinguished by the inventor of abstract argumentation Phan Minh Dung: the stable, preferred, grounded and complete semantics. During my PhD research I discovered two more: the semi-stable and stage semantics. That already gives six possibilities. When next to attacks also support by arguments is allowed there are even more possibilities, in Figure 15 (right) already 11.

The strange thing is that in real argumentation it never is an issue whether the semantics is grounded, complete, preferred or stable. The question is therefore

whether this mathematics—that is indeed beautiful and interesting—is exactly fitting for the phenomenon it covers, namely argumentation. Put otherwise: is it the right mathematics?

Search for the right mathematics

THE SEARCH FOR THE RIGHT MATHEMATICS led my colleagues and me to the study of reasoning with forensic evidence in criminal law. Something strange is at issue there. It turns out that there are three theoretical approaches to organize and evaluate reasoning with evidence (Figure 16).[16] The first approach is based on argumentation. Of that we already saw an example when discussing the witness who gave the suspect an alibi, but who turned out to be his partner. In an argumentative analysis the collection and weighing of arguments is what counts.

The second approach uses scenarios. In a scenario analysis of the evidence several scenarios are constructed and compared in connection with the evidence. The alibi scenario can for instance be compared with the guilt scenario as presented by the public prosecution. In a scenario analysis, the internal connections, the coherence of scenarios plays a role. A murder scenario without a motive or without a murder weapon is not complete.

The third approach is based on probability theory. In today's criminal law that is unavoidable by the prominent role of DNA evidence, of which the exceptionality is statistically analyzed. The probabilities reported about a good trace are so small that it almost cannot be a coin-

[16] See Anderson et al. 2005, Kaptein et al. 2009, Dawid et al. 2011, Di Bello and Verheij 2018. For specific contributions, see Wigmore 1913, Tribe 1971, Bennett and Feldman 1981, Kaye 1986, Thompson and Shumann 1987, Crombag et al. 1992, Pennington and Hastie 1993a,b, Wagenaar et al. 1993, Kadane and Schum 1996, Cook et al. 1998, Schum and Starace 2001, Thagard 2004, Zabell 2005, Keppens and Schafer 2006, Taroni et al. 2006, Hepler et al. 2007, Mortera and Dawid 2007, Pardo and Allen 2008, Tillers 2011, Fenton 2011, Keppens 2012, Fenton et al. 2013

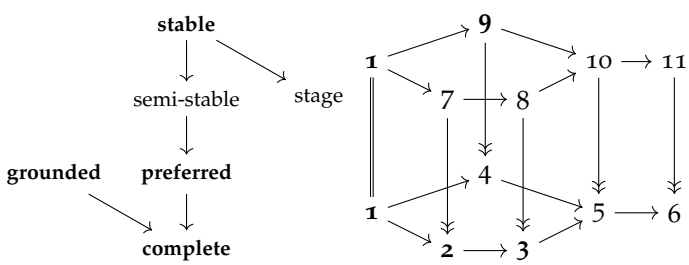


Figure 15: Left: initially **four**, later six semantics (for graphs with attack only, Dung 1995, Verheij 1996b). Right: eleven semantics, here only indicated with numbers (for graphs with both support and attack, Verheij 2003b). The numbers **1**, **2**, **3** en **9** correspond to the stable, semi-stable, preferred and stage semantics of the figure on the left.

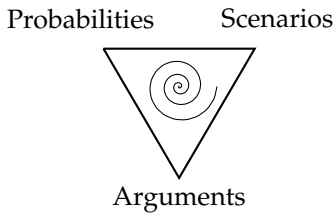


Figure 16: Three tools for organizing and evaluating evidence: arguments, scenarios and probabilities

cidence when a found DNA trace matches the suspect's DNA. Reasoning with probabilities is considered to be a daunting source of fallacies. With colleagues we have started the investigation of the connections between the three approaches.

First—supported by NWO—we have considered the relations between arguments and scenarios.[17] That led to the dissertation of Floris Bex,[18] supervised by Henry Prakken, Peter van Koppen and myself. In that work a new, formally elaborated theory of the relations between arguments and scenarios is developed. Here we see—horizontally—two scenarios about what can have happened in a crime and—vertically—the arguments for and against the elements of the scenarios (Figure 17).

[17] Bex et al. 2007, 2010, Bex and Verheij 2012, 2013

[18] Bex 2009

After that—thanks to a grant in the NWO Forensic Science program—we started to focus on probabilistic modeling. We have looked at Bayesian networks, a useful modeling technique in artificial intelligence that combines a graphical network structure with probabilities.[19]

[19] en.wikipedia.org/wiki/Bayesian_network

Figure 18 shows an example with three nodes and corresponding tables of probabilities and conditional probabilities. An important property of Bayesian networks is that they can efficiently model probability functions by the use of independencies between the variables.

In her dissertation Charlotte Vlek investigates the relations between scenarios and probabilities.[20] She has developed a method to model the scenarios about what happened in a crime case in a Bayesian network (Figure 19).[21] In the figure scenarios are represented as clusters of nodes. The evidence is visible in gray. She

[20] Vlek 2016

[21] Vlek et al. 2014

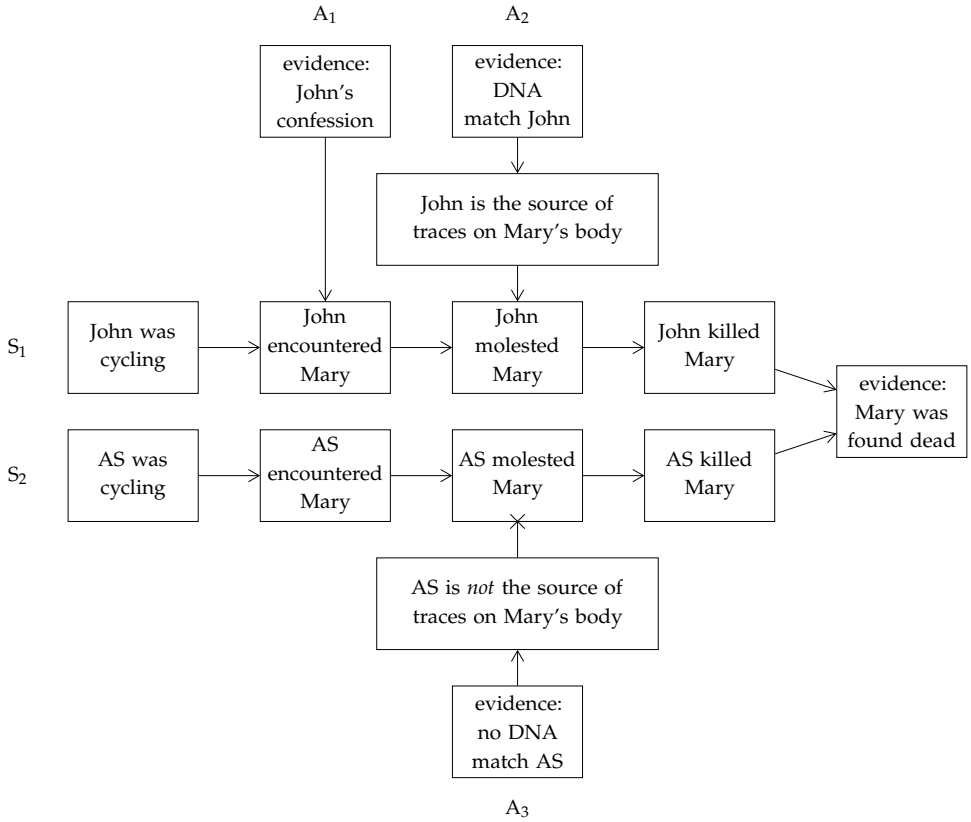
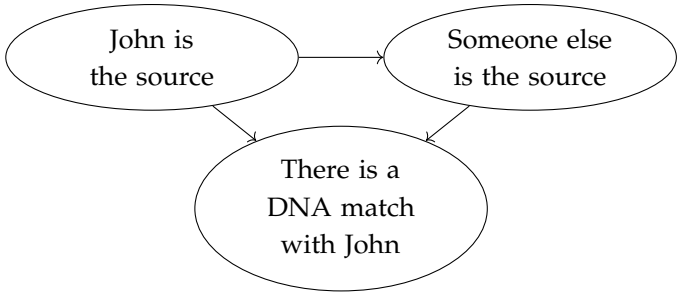


Figure 17: Arguments and scenarios (Verheij et al. 2016 after Bex 2009)



John is the source

John is the source = false	8000/8001
John is the source = true	1/8001

Someone else is the source

John is the source	false	true
Someone else is the source = false	0	1
Someone else is the source = true	1	0

There is a DNA match with John

John is the source	false		true	
Someone else	false	true	false	true
DNA match = false	0.5*	$1 - 0.66 \cdot 10^{-21}$	0	0.5*
DNA match = true	0.5*	$0.66 \cdot 10^{-21}$	1	0.5*

Figure 18: A Bayesian network (Verheij et al. 2016)

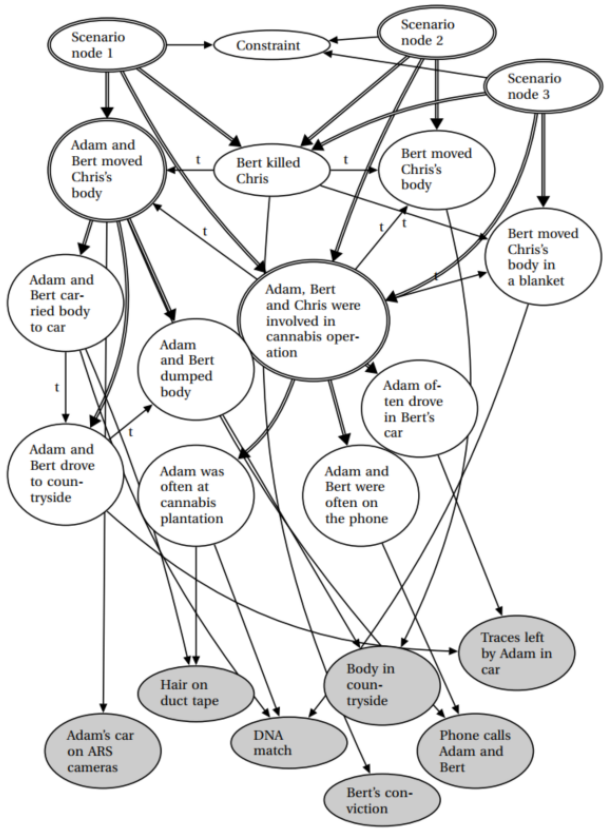


Figure 19: Scenarios and probabilities (Vlek 2016)

also showed how to explain a Bayesian network with scenarios.[22]

In his dissertation Sjoerd Timmer investigates the relations between arguments and probabilities.[23] He has developed an algorithm that can extract arguments and counterarguments from a Bayesian network (Figure 20).[24] The figure shows in the background a Bayesian network, in the middle the support graph generated from it, and in the foreground the arguments based on that. He also showed how to model argumentation schemes[25] in a Bayesian network.

These two dissertations were created in a fruitful collaboration with Henry Prakken, Silja Renooij, John-

[22] Vlek et al. 2016

[23] Timmer 2017

[24] Timmer et al. 2017

[25] Walton et al. 2008, Garssen 2001

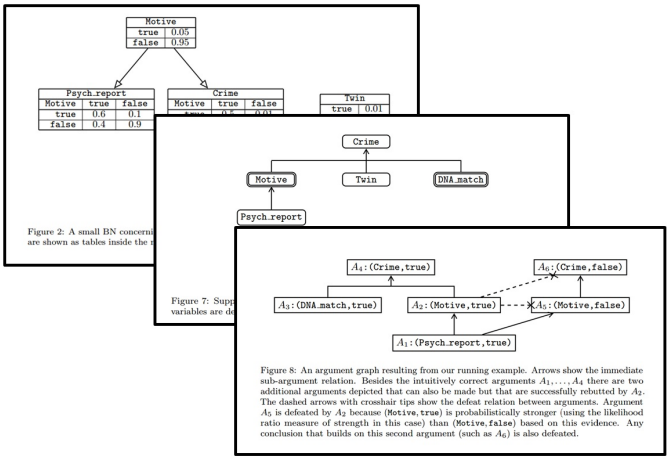


Figure 20: Arguments and probabilities (Timmer 2017)

Jules Meyer and Rineke Verbrugge. And Floris Bex whom I already mentioned always was nearby to give advice. Working in this way with the artificial intelligence technique of Bayesian networks gained us much insight about the relations between the use of arguments, scenarios and arguments as tools for the evaluation of evidential reasoning. Thereby our team has impacted the discussion about safe ways of evidential reasoning with arguments, scenarios and probabilities.

About catching a thief

THE ATTEMPTS TO COMBINE arguments, scenarios and probabilities led in the end to a new type of mathematics for argumentation. The first version of that formalism—developed under the California sun—consists of a graphical 'language' for the representation of the critical process of constructing, testing and evaluating hypotheses about what has happened in a crime case.[26] Perhaps you recognize the words I used when defining argumentation systems.

[26] Verheij 2014

As an example I discuss the crime story in Alfred Hitchock's beautiful 1955 film 'To Catch A Thief' set in the south of France. Robie—played by Cary Grant—is a former thief both famous and infamous because of his acrobatic climbing tricks.

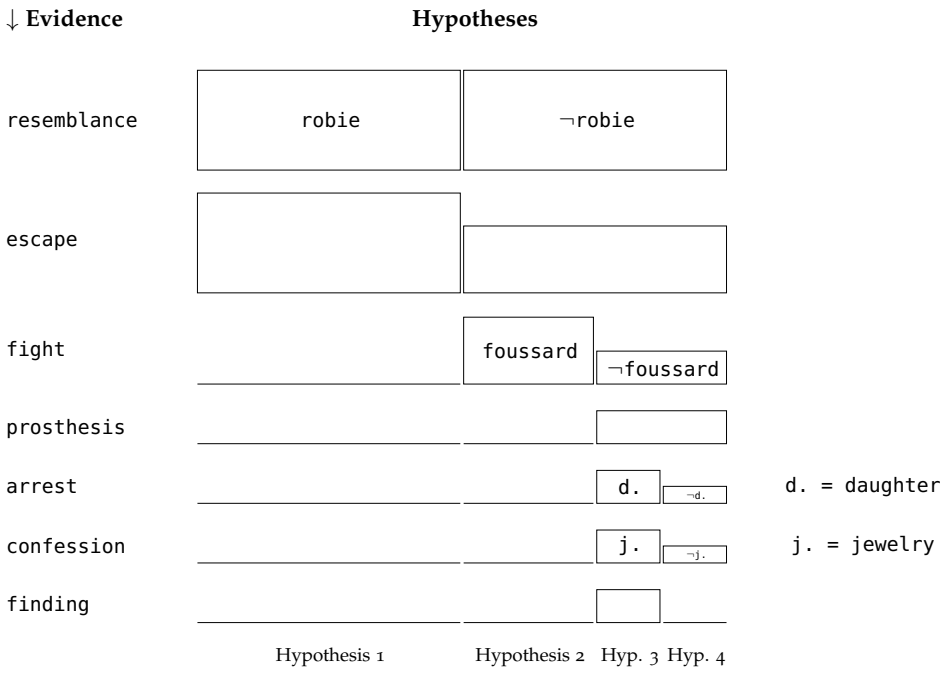


Figure 21: The crime story in Alfred Hitchcock's 'To Catch A Thief' (Verheij 2017a)

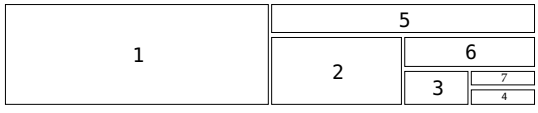


Figure 22: Case model of Alfred Hitchcock's 'To Catch A Thief' (Verheij 2017a)

On each line in Figure 21 new evidence is collected, hypotheses are constructed and tested, selected. The rectangles represent the possible hypotheses, and on each line less remains of the space of possibilities.

At the top two hypotheses are constructed: the rectangle on the left represents that Robie is committing spectacular thefts as in the old days, the rectangle on the right that he is not the thief. The evidence for that is the similarity between the new series of thefts and Robie's acrobatic style of years back. First the police has no preference for one of the hypotheses. That is why the rectangles on the first line are of equal size. On the second line the right rectangle representing Robie's innocence is smaller since after he escapes from the police his innocence has become less credible. After a nightly ambush a fight ensues, in which Foussard, a friend from the days of the *résistance*, falls from the roofs and dies.

As a consequence Robie is no longer a suspect, and therefore on the third line the rectangle representing Robie's guilt has disappeared. Now Foussard is the suspect, but with the possibility that he is innocent. Already quickly the suspicion fades as it is considered that Foussard—having a prosthetic leg—cannot have committed the required stunts. Later in the film Foussard's daughter is caught in the act and arrested. In her confession she explains where the stolen jewelry can be found. After finding the jewelry at the indicated place no one has doubts anymore that she is the thief. That is why at the bottom line only one rectangle remains, representing that Foussard's daughter is guilty.

The rectangles represent all possibilities, all cases that are considered possible. This becomes especially clear when the rectangles on all lines are placed on top of one another. That gives the model—in Figure 22—consisting of 7 rectangles, each representing a possibility of what might have happened. Rectangle 1 stands for the possibility that Robie indeed again became a thief. Rectangle 3 stands for the possibility that is believed in

the end—that the daughter of Foussard, a friend from the resistance is the thief.

The relative sizes of the rectangles represent their relative probability, their relative credibility. During the story these possibilities are gradually constructed, always with uncertainty about what should be believed. Until finally all uncertainty is removed by conclusive evidence. Then only one rectangle remains. Well: all reasonable uncertainty is removed. For at the end there is no specific reason for uncertainty left. Within the model built in a critical process no possible doubt is left; within the model it is certain that Foussard's daughter is the thief. Only by expanding the conceivable world of possibilities this can change.

A triplet of insights included in this graphical language are significant:

- First the validity of arguments can be derived from the set of possible hypotheses. As such the formalism gives a semantics for arguments and counterarguments.
- Second—and this is in contrast with Bayesian networks—few numbers are required, because the approach works primarily with the relative proportions, the ordering of the numbers.
- Third and finally there is a natural connection with logic and probability theory. Which is surprising now that as I told before both informal and formal argumentation research—including my own—is often inspired by contrasts with logic and probability theory.

Return to the mathematical foundations of argumentation

THE FORMAL ELABORATION of this graphical language has led to a formalism in which cases are central. The rectangles in the figures correspond to cases. Each case represents a possibility, a cluster of properties that can occur together.

The mathematical definition of case models follows. Cases are represented by logically consistent, logically

different sentences, that are pairwise incompatible. Their ordering is a total preorder, that is total and transitive, but not necessarily antisymmetric.

Definition (Verheij 2016b,a, 2017a)**.** *A* case model *is a pair* $(C, \geq)$ *with finite* $C \subseteq L$*, such that the following hold, for all* φ*,* ψ *and* $\chi \in C$*:*

1. $\not\models \neg\varphi$*;*
2. *If* $\not\models \varphi \leftrightarrow \psi$*, then* $\models \neg(\varphi \wedge \psi)$*;*
3. *If* $\models \varphi \leftrightarrow \psi$*, then* $\varphi = \psi$*;*
4. $\varphi \geq \psi$ *or* $\psi \geq \varphi$ *;*
5. *If* $\varphi \geq \psi$ *and* $\psi \geq \chi$*, then* $\varphi \geq \chi$*.*

Case models can be used to define three types of logical validity of arguments. Coherent arguments support a possible case; presumptive arguments support a maximally preferred case; and conclusive arguments moreover allow for no coherent attack.

Definition (Verheij 2016b,a, 2017a)**.** *Let* $(C, \geq)$ *be a case model. Then we define, for all* φ, ψ *and* $\chi \in L$*:*

1. $(C, \geq) \models (\varphi, \psi)$ *if and only if* $\exists\omega \in C$*:* $\omega \models \varphi \wedge \psi$*.*

   We then say that the argument from φ *to* ψ *is* coherent *with respect to the case model.*

2. $(C, \geq) \models \varphi \Rightarrow \psi$ *if and only if* $\exists\omega \in C$*:* $\omega \models \varphi \wedge \psi$ *and* $\forall\omega \in C$*: if* $\omega \models \varphi$*, then* $\omega \models \varphi \wedge \psi$*.*

   We then say that the argument from φ *to* ψ *is* conclusive *with respect to the case model.*

3. $(C, \geq) \models \varphi \rightsquigarrow \psi$ *if and only if* $\exists\omega \in C$*:*
   (a) $\omega \models \varphi \wedge \psi$*; and*
   (b) $\forall\omega' \in C$ *: if* $\omega' \models \varphi$*, then* $\omega \geq \omega'$*.*

   We then say that the argument from φ *to* ψ *is* presumptively valid *with respect to the case model. Such an argument is* properly defeasible*, when it is not conclusive.*

Circumstances χ are defeating *or* successfully attacking *when $(\varphi \wedge \chi, \psi)$ is not presumptively valid. Defeating circumstances are* rebutting *when $(\varphi \wedge \chi, \neg\psi)$ is presumptively valid; otherwise they are* undercutting. *Defeating circumstances are* excluding *when $(\varphi \wedge \chi, \psi)$ is not coherent.*

It is mathematically precise and philosophically relevant that total preorders are those orderings that can be realized numerically. Thereby these are the orderings that are simultaneously qualitative and quantitative; they are simultaneously with and without numbers (cf. the titles of Verheij 2014, 2017a). And hence also the preference ordering of case models is numerically realizable, and it turns out that that can even be done in a way that is compatible with probability theory.

Thus the three definitions of types of valid arguments can also be rewritten quantitatively. Coherent arguments correspond to a positive conditional probability of the conclusions given the premises; presumptive arguments with a probability higher than a threshold; and conclusive arguments with a probability equal to 1. In this way argumentation bridges logic and probability theory. Logic describes the properties of cases, and probability theory their preference ordering.

Arguments, cases and rules

[27] A key theme in the field AI & Law, see Bench-Capon et al. 2012 for a historical perspective and Gardner 1987, Rissland and Ashley 1987, Ashley 1990, Branting 1991, Berman and Hafner 1995, Gordon 1995, Loui and Norman 1995, Aleven and Ashley 1995, Prakken and Sartor 1996, Hage 1997, McCarty 1997, Prakken and Sartor 1998, Bench-Capon and Sartor 2003 for specific contributions

As said the mathematics of case models is developed by thinking about correct reasoning with forensic evidence, and then especially by the puzzle how arguments, scenarios and probabilities go together without hindering one another. The exciting issue is now whether the validity of a complex argument structure can also be reconstructed using case models. More precisely: given a rule structure of rules with their exceptions, is there a case model in which these rules are valid?

It turns out that this is possible. A good example can be given by considering the relations between arguments, cases and rules in the law,[27] a topic that I looked at earlier when working with the first PhD candidate that I

could advise, Bram Roth, first with Jaap Hage, later with Hans Crombag.[28] Earlier we encountered the formal structural core of Dutch law on the repair of damages in case of an unlawful act (Figure 5, p. 46) that I already used to illustrate knowledge systems. And here you see a case model—consisting of 16 cases—in which this structure is valid (Figure 23). What is special is that a collection of cases, apparently with little structure, makes a complex argument structure valid.

[28] Roth 2003, Roth and Verheij 2004

Thus the knowledge that is represented in the argument structure of Figure 5 (p. 46) is grounded in the data of the collection of cases in Figure 23.

Arguments bridge the gap between knowledge and data systems

NOW WE HAVE RETURNED to bridging the gap in artificial intelligence, the gap between knowledge and data systems. Because it is not a big step to view the cases in a case model as the data that can be used for learning—like in data systems—and the rules that are valid in it as the knowledge structure of knowledge systems. The developed mathematics can hence provide the foundations for argumentation systems that bridge the gap between knowledge and data (Figure 24).

Such argumentation systems can grow to become the good artificial intelligence that has to be developed: intelligent systems that give *good answers*, have *good reasons* and make *good choices*.

1. First the good answers. An argumentation system searches for an answer to a complex problem in a critical discussion—and if such an answer cannot be found it searches for the best answer. This goes

Figure 23: Case model for Dutch law of unlawful acts (Verheij 2017b)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	¬dut	¬dut	¬dut	dut	dut	dut	dut	dut	dut	dut	dut	dut	¬dut	¬dut	¬dut
¬dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg	dmg
	¬unl	unl	unl	unl	unl	unl	unl	unl	unl	unl	unl	unl	¬unl	¬unl	unl
		¬imp	imp	imp	imp	imp	imp	imp	imp	imp	imp	imp			imp
			¬cau	cau	cau	cau	cau	cau	cau	cau	cau	cau			cau
	¬vrt			vrt	vrt	vrt	¬vrt	¬vrt	¬vrt	¬vrt	¬vrt	¬vrt	vrt	¬vrt	
	¬vst			¬vst	¬vst	¬vst	vst	vst	vst	¬vst	¬vst	¬vst	¬vst	vst	vst
	¬vun			¬vun	¬vun	¬vun	¬vun	¬vun	¬vun	vun	vun	vun			
		¬ift		ift	¬ift	¬ift	ift	¬ift	¬ift	ift	¬ift	¬ift			
		¬ila		¬ila	ila	¬ila	¬ila	ila	¬ila	¬ila	ila	¬ila			
		¬ico		¬ico	¬ico	ico	¬ico	¬ico	ico	¬ico	¬ico	ico			
				¬jus	¬jus	¬jus	¬jus	¬jus	¬jus	¬jus	¬jus	¬jus	jus	jus	
				prp	prp	prp									¬prp

1 > 2 > 3 > 4 > 5 ~ 6 ~ 7 ~ 8 ~ 9 ~ 10 ~ 11 ~ 12 ~ 13 > 14 ~ 15 ~ 16

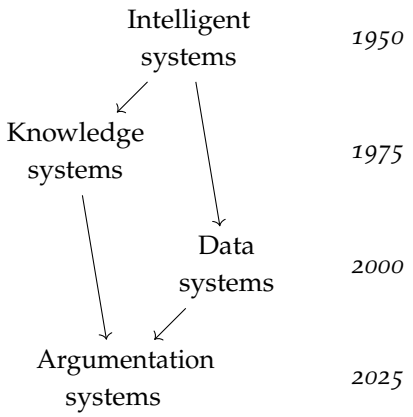


Figure 24: Beyond the gap in artificial intelligence

beyond the maximization of the number of good answers as we know it from data systems. At the same time the knowledge in an argumentation system can persistently be developed by the construction of new hypotheses and their evaluation using the available data.

2. Then the good reasons. In argumentation systems reasons are prominent anyway. An argumentation system can answer a why-question by an assorted list of reasons for and against a given answer, sometimes supplemented with possible alternatives. In this way the argument structure provides insight into the knowledge structure 'on the inside' while maintaining the connection to data.

3. Finally the good choices. Argumentation systems can stick to the rules, they can follow obtaining norms, while taking specific circumstances into account. By their ability of critical discussion they can give critical rejoinders, exactly as wanted in a serious discussion about what is the good choice.

By the development of argumentation systems the dreams of artificial intelligence come closer, and fears become more manageable. Because consider again the

proposals to regulate artificial intelligence by enforcing human control and by the prohibition of 'killer robots'. My proposal opens up an alternative route to regulate artificial intelligence because with argumentation systems one can have a critical discussion on the basis of reasonable arguments. Winning such a discussion will not always be easy. Every now and then we will even lose. In that respect things will not be different from what we are used to in science, politics and everyday life. But that is the heart of the matter: intelligent behavior does not concern the winning of a discussion but the finding of good answers to the difficult problems of life in a complex, dynamic world. As we know from science, politics and everyday life, that requires critical discussion. The development of good artificial intelligence that helps us with that will for now remain a fascinating and challenging task that can only be performed by humans.

It will be clear that before we have arrived there—and computers and robots are serious partners in a critical discussion—still much research has to be done and it is a pleasure to be able to work on that here in Groningen with our dedicated staff and more than 400 students of artificial intelligence.

Here I end my lecture and I am curious about your arguments for good artificial intelligence.

Dankwoord

Zoals uitgesproken tijdens de feestelijke avond na de oratie

Welkom allemaal! Wat een genoegen om jullie zo bij elkaar te zien—al die verschillende groepen mensen, zo belangrijk in mijn leven. Collega's, vrienden, familie, dank jullie voor jullie komst, dank jullie voor de warme woorden en liefde. Ik dank College van Bestuur van de Rijksuniversiteit Groningen en het bestuur van de Faculteit Science and Engineering voor het in mij gestelde vertrouwen door mij de leerstoel Kunstmatige Intelligentie en Argumentatie te gunnen.

Een paar mensen noem ik bij naam.

Beste Jaap, Jij hebt me het toegangskaartje tot de wetenschap gegeven. Al op de drempel, meteen na het sollicitatiegesprek, nog zonder ruggespraak met de andere commissieleden, vertelde je dat ik als promovendus was aangenomen. Je onbegrensde ambitie in de kunstmatige intelligentie, je passie voor wetenschap, de mensen daarin en rondom, zijn een blijvend voorbeeld. Dank je wel.

Beste Jaap, Door jou ben ik op het wetenschappelijk pad gekomen dat ik nog steeds bewandel. Na een jaar grasduinen in de 'multimedia information retrieval' gooiden we het roer om naar een gedeelde liefde: niet-monotone logica, en dan toegepast in het recht. Jouw diepe inzicht in recht en filosofie hebben mijn wiskundige blik op wetenschap en wereld fundamenteel veranderd—en wat was de weg vol levendige discussies een feest. Dank je wel.

Beste Lambert, We zijn het roerend eens: echte kunstmatige intelligentie bouwen, er is geen andere weg. Voordat we machines hebben die mooie verhalen kunnen begrijpen en vertellen zijn we nog niet klaar. Onder jouw leiderschap kon er in Groningen een stevig onderzoeksinstituut opgebouwd worden op de grondvesten van onze mooie opleidingen. Dank je wel.

Beste Rineke, Wat heeft het lidmaatschap van jouw multi-agentsystemengroep me veel gebracht. Je betrokken, waardecreërende stijl van leiderschap en begeleiderschap zijn een inspiratie. En het kan toch geen toeval zijn dat onze overlappende studies wiskunde in Amsterdam ons beiden op het wondermooie pad van de kunstmatige intelligentie hebben gebracht. Dank je wel.

Beste Henry, Dank je wel voor de vruchtbare en plezierige samenwerking. Ik kijk uit naar onze volgende wandelingen door wetenschap, stad en land die we vast en zeker nog gaan maken. Dank je wel.

Beste Floris, Wat geniet ik van de samenwerking met jou. Dat begon volop toen jij je mooie proefschrift schreef en is nooit meer opgehouden. Je in je sterke ontwikkeling te mogen volgen is een plezier. Dank je wel.

Beste Silja, Door jou heb ik een belangrijke draai in mijn wetenschappelijke inzicht kunnen maken, namelijk in de richting van probabilistisch modelleren. Jouw scherpe en onafhankelijke stijl zijn daarbij van grote waarde geweest—en nog steeds. Dank je wel.

Beste Charlotte en Sjoerd, Wat heb ik een plezier beleefd aan het *NWO Forensic Science* project, dat door jullie goede werk zo'n succes is geworden. Ik ben benieuwd waar jullie vele talenten je nog gaan brengen. Een genoegen om mee te maken. Dank jullie wel.

Beste Bram, Harmen en Jacky, Als begeleider heb ik jullie—in jullie heel verschillende omstandigheden en hoedanigheden—mogen volgen op je weg naar een succesvolle promotie. Juist door die variatie hebben jullie trajecten me veel gebracht. Dank jullie wel.

Beste John, Ron, Guillermo, John-Jules, Hans, Peter, Bij jullie zag ik het plezier in ambitieuze wetenschap, en het belang van onafhankelijkheid. Het vertrouwen dat ik van jullie al vroeg kreeg is van onschatbare waarde. Dank jullie wel.

Beste Roland, Mike, Anne, Jerry, Harm, Marcello, Door en met jullie heb ik onder de Californische zon kunnen genieten van wetenschap gericht op echte impact en met schaamteloze ambitie. Onvergetelijk. Dank jullie wel.

Beste Frans, Erik, Jan-Albert, Wij delen de liefde voor het interdisciplinaire onderzoek naar argumentatie en daarbij laaf ik mij aan jullie genuanceerde levensvisie, zo passend bij ons onderwerp. Dank jullie wel.

Beste Arnold, Davide, Fokie, Jacolien, Jelmer, Jennifer, Katja, Lambert, Marco, Marieke, Niels, Raffaella, Rineke, Sietse, Ton en andere collega's op de gangen van het roemruchte ALICE instituut. Prachtig om te zien hoe jullie je mooie onderzoek en onderwijs vormgeven—en daarvan te kunnen leren. Ik prijs me gelukkig met jullie en met onze studenten te werken in dit onvolprezen instituut. En dan noem ik nog niet de onontbeerlijke ondersteunende staf. Elina: wat een betrouwbare en plezierige samenwerking. Allen: Dank jullie wel.

Lieve familie, lieve vrienden, Jullie weten hoe onmisbaar jullie voor mij zijn. Dank jullie wel.

Lieve Ben, die er niet meer is, lieve Truus, Jullie liefde voor ons, jullie kinderen, heeft me gevormd op een manier waar ik jullie heel dankbaar voor ben. En de rijkdom van jullie creatieve en eigenwijze levens komt daar nog bij. Dank jullie wel.

Lieve Grietje, Wat een geluk heb ik met jou als zus. Dank je wel.

Lieve Jannes en Maarten, Jullie komst in mijn leven is de grootste omwenteling die ik heb mee gemaakt. Wat heerlijk om jullie dit leven te zien leven, en wat heerlijk om door jullie in de maling te worden genomen. Dank jullie wel.

Lieve Margreet, Liefde van mijn leven. Waar zou ik zijn zonder jouw wijsheid, zonder jouw aarding? Telkens weer is het spannend, telkens weer leerzaam, telkens weer liefdevol. Ik kijk uit naar wat nog komen gaat. Dank je wel.

Curriculum vitae

Bart Verheij holds the chair of artificial intelligence and argumentation at the University of Groningen. He is head of the department of Artificial Intelligence in the Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence at the Faculty of Science and Engineering. He participates in the Multi-Agent Systems research program. He is president of the International Association for Artificial Intelligence and Law (IAAIL).

His research focuses on artificial intelligence and argumentation, often with the law as application domain. He is working on the theoretical, computational and empirical connections between knowledge, data and reasoning, as a contribution to explainable, responsible and social artificial intelligence.

He has an MSc degree in Mathematics (University of Amsterdam, algebraic geometry) and obtained his PhD degree at Maastricht University (Faculty of Law, Department of Metajuridica; Faculty of General Sciences, Department of Computer Science), on a dissertation about the formal modeling of argumentation, with applications in law.

He led a research project on the connections between arguments, scenarios and probabilities in forensic reasoning with evidence, funded by the NWO Forensic Science program (2012-2017). In the academic year 2013-2014, he was resident fellow at Stanford University. He participated in CodeX - the Stanford Center for Legal Informatics, a collaboration between the Stanford AI Lab and Stanford Law School, where he is now

listed as affiliated faculty. In 2013 and 2018, he was an invited visiting lecturer at the Institute of Logic and Cognition, Sun Yat-Sen University (Guangzhou, China). He was invited researcher at the Isaac Newton Institute for Mathematical Sciences (University of Cambridge, Autumn 2016).

He has published on artificial intelligence and argumentation in more than a hundred peer-reviewed publications and has a 30+ Google Scholar h-index.

He is co-editor-in-chief of the journal Argument and Computation, section editor of the journal Artificial Intelligence and Law, and board member of professional organisations (IAAIL, president; COMMA, vice-president; JURIX, vice-president/secretary; BNVKI, community builder).

He co-organized conferences at Vienna University of Technology (2012, COMMA), the National Research Council of Italy in Rome (2013, ICAIL), Stanford University (2014, Trial With and Without Mathematics), and the University of Groningen (2017, BNAIC; 2018, JURIX; 2019, ECA).

Bibliography

V. Aleven and K. D. Ashley. Doing things with factors. In *Proceedings of the Fifth International Conference on Artificial Intelligence and Law (ICAIL 1995)*, pages 31–41. ACM Press, New York (New York), 1995.

L. Amgoud and M. Caminada. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 172:286–310, 2007.

T. Anderson, D. Schum, and W. Twining. *Analysis of Evidence. 2nd Edition*. Cambridge University Press, Cambridge, 2005.

K. D. Ashley. *Modeling Legal Arguments: Reasoning with Cases and Hypotheticals*. The MIT Press, Cambridge (Massachusetts), 1990.

K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. Simari, M. Thimm, and S. Villata. Toward artificial argumentation. *AI Magazine*, 38(3):25–36, 2017.

P. Baroni, G. Boella, F. Cerutti, M. Giacomin, L. van der Torre, and S. Villata. On the input/output behavior of argumentation frameworks. *Artificial Intelligence*, 217: 144–197, 2014.

P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre, editors. *Handbook of Formal Argumentation*. College Publications, London, 2018.

E. M. Barth and E. C. W. Krabbe. *From Axiom to Dialogue. A Philosophical Study of Logics and Argumentation*. De Gruyter, New York (New York), 1982.

T. J. M. Bench-Capon and P. E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10-15):619–641, 2007.

T. J. M. Bench-Capon and G. Sartor. A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, 150(1):97–143, 2003.

T. J. M. Bench-Capon, M. Araszkiewicz, K. D. Ashley, K. Atkinson, F. J. Bex, F. Borges, D. Bourcier, D. Bourgine, J. G. Conrad, E. Francesconi, T. F. Gordon, G. Governatori, J. L. Leidner, D. D. Lewis, R. P. Loui, L. T. McCarty, H. Prakken, F. Schilder, E. Schweighofer, P. Thompson, A. Tyrrell, B. Verheij, D. N. Walton, and A. Z. Wyner. A history of AI and Law in 50 papers: 25 years of the International Conference on AI and Law. *Artificial Intelligence and Law*, 20(3):215–319, 2012.

W. L. Bennett and M. S. Feldman. *Reconstructing Reality in the Courtroom*. London: Tavistock Feldman, 1981.

D. H. Berman and C. L. Hafner. Understanding precedents in a temporal context of evolving legal doctrine. In *Proceedings of the Fifth International Conference on Artificial Intelligence and Law*, pages 42–51. ACM Press, New York (New York), 1995.

P. Besnard and A. Hunter. *Elements of Argumentation*. The MIT Press, Cambridge (Massachusetts), 2008.

F. J. Bex. *Evidence for a Good Story: A Hybrid Theory of Arguments, Stories and Criminal Evidence*. Dissertation University of Groningen, Groningen, 2009.

F. J. Bex and B. Verheij. Solving a murder case by asking critical questions: An approach to fact-finding in terms of argumentation and story schemes. *Argumentation*, 26:325–353, 2012.

F. J. Bex and B. Verheij. Legal stories and the process of proof. *Artificial Intelligence and Law*, 21(3):253–278, 2013.

F. J. Bex, S. W. van den Braak, H. van Oostendorp, H. Prakken, B. Verheij, and G. A. W. Vreeswijk. Sense-making software for crime investigation: How to combine stories and arguments? *Law, Probability and Risk*, 6:145–168, 2007.

F. J. Bex, P. J. van Koppen, H. Prakken, and B. Verheij. A hybrid formal theory of arguments, stories and criminal evidence. *Artificial Intelligence and Law*, 18:1–30, 2010.

A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93:63–101, 1997.

L. K. Branting. Building explanations from rules and structured cases. *International Journal of Man-Machine Studies*, 34(6):797–837, 1991.

G. Brewka and S. Woltran. Abstract dialectical frameworks. In *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2010)*, pages 102–111. The AAAI Press, Menlo Park (California), 2010.

G. Brewka, H. Strass, S. Ellmauthaler, J. P. Wallner, and S. Woltran. Abstract dialectical frameworks revisited. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI 2013)*, pages 803–809. The AAAI Press, Menlo Park (California), 2013.

F. Cerutti, S. A. Gaggl, M. Thimm, and J. P. Wallner. Foundations of implementations for formal argumentation. *IfCoLog Journal of Logics and their Applications*, 4(8):2623–2706, 2017.

C. I. Chesñevar, A. G. Maguitman, and R. P. Loui. Logical models of argument. *ACM Computing Surveys*, 32(4):337–383, 2000.

C. I. Chesñevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. A. W. Vreeswijk, and S. Willmott. Towards an argument interchange format. *The Knowledge Engineering Review*, 21(4):293–316, 2006.

R. Cook, I. W. Evett, G. Jackson, P. J. Jones, and J. A. Lambert. A hierarchy of propositions: deciding which level to address in casework. *Science and Justice*, 38 (4):231–239, 1998.

H. F. M. Crombag, P. J. van Koppen, and W. A. Wagenaar. *Dubieuze Zaken: De Psychologie van Strafrechtelijk Bewijs*. Uitgeverij Contact, Amsterdam, 1992.

A. P. Dawid, W. Twining, and M. Vasiliki, editors. *Evidence, Inference and Enquiry*. Oxford University Press, Oxford, 2011.

M. Di Bello and B. Verheij. Evidential reasoning. In G. Bongiovanni, G. Postema, A. Rotolo, G. Sartor, C. Valentini, and D. N. Walton, editors, *Handbook of Legal Reasoning and Argumentation*, pages 447–493. Springer, Dordrecht, 2018.

P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

N. E. Fenton. Science and law: Improve statistics in court. *Nature*, 479:36–37, 2011.

N. E. Fenton, M. D. Neil, and D. A. Lagnado. A general structure for legal arguments about evidence using Bayesian Networks. *Cognitive Science*, 37:61–102, 2013.

J. B. Freeman. *Dialectics and the Macrostructure of Arguments. A Theory of Argument Structure*. Foris, Berlin, 1991.

D. M. Gabbay, C. J. Hogger, and J. A. Robinson, editors. *Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 3. Nonmonotonic Reasoning and Uncertain Reasoning*. Clarendon Press, Oxford, 1994.

A. J. García and G. R. Simari. Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming*, 4(2):95–138, 2004.

A. Gardner. *An Artificial Intelligence Approach to Legal Reasoning*. The MIT Press, Cambridge (Massachusetts), 1987.

B. J. Garssen. Argument schemes. In *Crucial Concepts in Argumentation Theory*, pages 81–99. Amsterdam University Press, Amsterdam, 2001.

R. Girle, D.L. Hitchcock, P. McBurney, and B. Verheij. Decision support for practical reasoning: a theoretical and computational perspective. In C. Reed and T.J. Norman, editors, *Argumentation Machines. New Frontiers in Argument and Computation*, pages 55–84. Kluwer Academic Publishers, Dordrecht, 2001.

T. F. Gordon. *The Pleadings Game: An Artificial Intelligence Model of Procedural Justice*. Kluwer, Dordrecht, 1995.

F. Grasso, A. Cawsey, and R. Jones. Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *International Journal of Human-Computer Studies*, 53(6):1077–1115, 2000.

J. C. Hage. *Reasoning with Rules. An Essay on Legal Reasoning and Its Underlying Logic*. Kluwer Academic Publishers, Dordrecht, 1997.

J. C. Hage and B. Verheij. Reason-based logic: a logic for reasoning with rules and reasons. *Law, Computers and Artificial Intelligence*, 3(2–3):171–209, 1994.

J. C. Hage and B. Verheij. The law as a dynamic interconnected system of states of affairs: a legal top ontology. *International Journal of Human-Computer Studies*, 51(6): 1043–1077, 1999.

A. B. Hepler, A. P. Dawid, and V. Leucari. Object-oriented graphical representations of complex patterns of evidence. *Law, Probability and Risk*, 6(1–4):275–293, 2007.

D.L. Hitchcock and B. Verheij, editors. *Arguing on the Toulmin Model. New Essays in Argument Analysis and Evaluation (Argumentation Library, Volume 10)*. Springer, Dordrecht, 2006.

D. R. Hofstadter. *Gödel, Escher, Bach. An Eternal Golden Braid*. Basic Books, New York (New York), 1979.

D. R. Hofstadter. *Gödel, Escher, Bach. Een Eeuwige Gouden Band*. Contact, Amsterdam, 1985.

D. R. Hofstadter. On seeing A's and seeing As. *SEHR*, 4(2), 1995.

D. R. Hofstadter. *I am a strange loop*. Basic Books, New York (New York), 2007.

J. B. Kadane and D. A. Schum. *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. Wiley, Chichester, 1996.

H. Kaptein, H. Prakken, and B. Verheij, editors. *Legal Evidence and Proof: Statistics, Stories, Logic (Applied Legal Philosophy Series)*. Ashgate, Farnham, 2009.

D. H. Kaye. Do we need a calculus of weight to understand proof beyond a reasonable doubt? *Boston University Law Review*, 66:657–672, 1986.

J. Keppens. Argument diagram extraction from evidential Bayesian networks. *Artificial Intelligence and Law*, 20:109–143, 2012.

J. Keppens and B. Schafer. Knowledge based crime scenario modelling. *Expert Systems with Applications*, 30(2):203–222, 2006.

P. A. Kirschner, S. J. B. Shum, and C. S. Carr. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer, Berlin, 2003.

S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

R. P. Loui and J. Norman. Rationales and argument moves. *Artificial Intelligence and Law*, 3:159–189, 1995.

L. T. McCarty. Some arguments about legal arguments. In *Proceedings of the 6th International Conference on Artificial Intelligence and Law (ICAIL 1997)*, pages 215–224. ACM Press, New York (New York), 1997.

S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9):901–934, 2009.

J. Mortera and P. Dawid. Probability and evidence. In T. Rudas, editor, *Handbook of Probability Theory*. Sage Handbook, 2007.

M. S. Pardo and R. J. Allen. Juridical proof and the best explanation. *Law and Philosophy*, 27:223–268, 2008.

N. Pennington and R. Hastie. Reasoning in explanation-based decision making. *Cognition*, 49(1–2):123–163, 1993a.

N. Pennington and R. Hastie. The story model for juror decision making. In *Inside the Juror*, pages 192–221. Cambridge University Press, Cambridge, 1993b.

J. L. Pollock. Defeasible reasoning. *Cognitive Science*, 11(4):481–518, 1987.

J. L. Pollock. *Cognitive Carpentry: A Blueprint for How to Build a Person*. The MIT Press, Cambridge (Massachusetts), 1995.

H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

H. Prakken and G. Sartor. A dialectical model of assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law*, 4:331–368, 1996.

H. Prakken and G. Sartor. Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law*, 6:231–287, 1998.

I. Rahwan and G. R. Simari, editors. *Argumentation in Artificial Intelligence*. Springer, Dordrecht, 2009.

I. Rahwan, S. D. Ramchurn, N. R. Jenningd, P. McBurney, S. Parsons, and L. Sonenberg. Argumentation-based negotiation. *The Knowledge Engineering Review*, 18(4):343–375, 2003.

C. Reed and F. Grasso. Recent advances in computational models of natural argument. *International Journal of Intelligent Systems*, 22:1–15, 2007.

C. Reed and T.J. Norman, editors. *Argumentation Machines. New Frontiers in Argument and Computation*. Kluwer Academic Publishers, Dordrecht, 2004.

R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

E. L. Rissland and K. D. Ashley. A case-based system for trade secrets law. In *Proceedings of the First International Conference on Artificial Intelligence and Law*, pages 60–66. ACM Press, New York (New York), 1987.

B. Roth. *Case-Based Reasoning in the Law. A Formal Theory of Reasoning by Case Comparison.* Dissertation Universiteit Maastricht, Maastricht, 2003.

B. Roth and B. Verheij. Dialectical arguments and case comparison. In T. F. Gordon, editor, *Legal Knowledge and Information Systems: JURIX 2004: The Seventeenth Annual Conference*, pages 99–108. IOS Press, Amsterdam, 2004.

S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach. Third Edition*. Pearson, Boston (Massachusetts), 2010.

O. Scheuer, F. Loll, N. Pinkwart, and B. M. McLaren. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5(1):43–102, 2010.

D. A. Schum and S. Starace. *The Evidential Foundations of Probabilistic Reasoning*. Northwestern University Press, Evanston (Illinois), 2001.

G. R. Simari and R. P. Loui. A mathematical treatment of defeasible reasoning and its applications. *Artificial Intelligence*, 53:125–157, 1992.

F. Taroni, C. Aitken, P. Garbolino, and A. Biedermann. *Bayesian Networks and Probabilistic Inference in Forensic Science*. Wiley, Chichester, 2006.

P. Thagard. Causal inference in legal decision making: Explanatory coherence vs. Bayesian Networks. *Applied Artificial Intelligence*, 18:231–249, 2004.

M. Thimm. A probabilistic semantics for abstract argumentation. In *Proceedings of the European Conference on Artificial Intelligence (ECAI 2012)*, pages 750–755. IOS Press, Amsterdam, 2012.

W. C. Thompson and E. L. Shumann. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behaviour*, 11:167–187, 1987.

P. Tillers. Trial by mathematics—reconsidered. *Law, Probability and Risk*, 10:167–173, 2011.

S. T. Timmer. *Designing and Understanding Forensic Bayesian Networks using Argumentation*. Dissertation Utrecht University, Utrecht, 2017.

S. T. Timmer, J. J. Meyer, H. Prakken, S. Renooij, and B. Verheij. A two-phase method for extracting explanatory arguments from Bayesian Networks. *International Journal of Approximate Reasoning*, 80:475–494, 2017.

S. E. Toulmin. *The Uses of Argument*. Cambridge University Press, Cambridge, 1958.

L. Tribe. Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, 84:1329–1393, 1971.

J. van Benthem. Foundations of conditional logic. *Journal of Philosophical Logic*, 13: 303–349, 1984.

F. H. van Eemeren, B. Garssen, E. C. W. Krabbe, A. F. Snoeck Henkemans, B. Verheij, and J. H. M. Wagemans. *Handbook of Argumentation Theory*. Springer, Berlin, 2014.

T. van Gelder. Enhancing deliberation through computer supported argument visualization. In P. A. Kirschner, S. J. B. Shum, and C. S. Carr, editors, *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, pages 97–115. Springer, 2003.

B. Verheij. *Rules, Reasons, Arguments. Formal Studies of Argumentation and Defeat.* Dissertation Universiteit Maastricht, Maastricht, 1996a.

B. Verheij. Two approaches to dialectical argumentation: Admissible sets and argumentation stages. In J. J. Meyer and L. C. van der Gaag, editors, *Proceedings of NAIC'96*, pages 357–368. Universiteit Utrecht, Utrecht, 1996b.

B. Verheij. Automated argument assistance for lawyers. In *Proceedings of the 7th International Conference on Artificial Intelligence and Law (ICAIL 1999)*, pages 43–52. ACM Press, New York (New York), 1999a.

B. Verheij. Logic, context and valid inference. or: Can there be a logic of law? In H.J. van den Herik, M. F. Moens, J. Bing, B. van Buggenhout, J. Zeleznikow, and C. A. F. M. Grütters, editors, *Legal Knowledge Based Systems. JURIX 1999: The Twelfth Conference*, pages 109–121. Gerard Noodt Instituut, Nijmegen, 1999b.

B. Verheij. Artificial argument assistants for defeasible argumentation. *Artificial Intelligence*, 150(1–2):291–324, 2003a.

B. Verheij. DefLog: on the logical interpretation of prima facie justified assumptions. *Journal of Logic and Computation*, 13(3):319–346, 2003b.

B. Verheij. Dialectical argumentation with argumentation schemes: An approach to legal logic. *Artificial Intelligence and Law*, 11(1–2):167–195, 2003c.

B. Verheij. *Virtual Arguments. On the Design of Argument Assistants for Lawyers and Other Arguers*. T.M.C. Asser Press, The Hague, 2005a.

B. Verheij. Evaluating arguments based on Toulmin's scheme. *Argumentation*, 19(3): 347–371, 2005b.

B. Verheij. Argumentation support software: Boxes-and-arrows and beyond. *Law, Probability and Risk*, 6:187–208, 2007a.

B. Verheij. A labeling approach to the computation of credulous acceptance in argumentation. In M. Veloso, editor, *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 623–628. 2007b.

B. Verheij. The Toulmin argument model in artificial intelligence. or: How semi-formal, defeasible argumentation schemes creep into logic. In I. Rahwan and G. R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 219–238. Springer, Berlin, 2009.

B. Verheij. To catch a thief with and without numbers: Arguments, scenarios and probabilities in evidential reasoning. *Law, Probability and Risk*, 13:307–325, 2014.

B. Verheij. Formalizing value-guided argumentation for ethical systems design. *Artificial Intelligence and Law*, 24(4):387–407, 2016a.

B. Verheij. Correct grounded reasoning with presumptive arguments. In L. Michael and A. Kakas, editors, *15th European Conference on Logics in Artificial Intelligence, JELIA 2016. Larnaca, Cyprus, November 9–11, 2016. Proceedings (LNAI 10021)*, pages 481–496. Springer, Berlin, 2016b.

B. Verheij. Proof with and without probabilities. Correct evidential reasoning with presumptive arguments, coherent hypotheses and degrees of uncertainty. *Artificial Intelligence and Law*, 25(1):127–154, 2017a.

B. Verheij. Formalizing arguments, rules and cases. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*, pages 199–208. ACM Press, New York (New York), 2017b.

B. Verheij and J. C. Hage. Reasoning by analogy: a formal reconstruction. In H. Prakken, A. J. Muntjewerff, and A. Soeteman, editors, *Legal Knowledge Based Systems. The Relation with Legal Theory*, pages 65–78. Koninklijke Vermande, Lelystad, 1994.

B. Verheij, J. C. Hage, and A. R Lodder. Logical tools for legal argument: a practical assessment in the domain of tort. In *Proceedings of the 6th International Conference on Artificial Intelligence and Law (ICAIL 1997)*, pages 243–249. ACM Press, New York (New York), 1997.

B. Verheij, J. C. Hage, and H. J. van den Herik. An integrated view on rules and principles. *Artificial Intelligence and Law*, 6(1):3–26, 1998.

B. Verheij, J. C. Hage, T. van der Meer, and G. Span. *Vaardig met Recht. Over Casus Oplossen en Andere Juridische Vaardigheden (Skilful in the Law. On Case Solving and Other Legal Skills)*. Boom Juridische Uitgevers, The Hague, 2004.

B. Verheij, J. C. Hage, and G. E. van Maanen. De logica van de onrechtmatige daad. *Nederlands Tijdschrift voor Burgerlijk Recht*, 16(4):95–102, 2007.

B. Verheij, F. J. Bex, S. T. Timmer, C. S. Vlek, J. J. Meyer, S. Renooij, and H. Prakken. Arguments, scenarios and probabilities: Connections between three normative frameworks for evidential reasoning. *Law, Probability and Risk*, 15: 35–70, 2016.

C. S. Vlek. *When Stories and Numbers Meet in Court Constructing and Explaining Bayesian Networks for Criminal Cases with Scenarios*. Dissertation University of Groningen, Groningen, 2016.

C. S. Vlek, H. Prakken, S. Renooij, and B. Verheij. Building Bayesian Networks for legal evidence with narratives: a case study evaluation. *Artificial Intelligence and Law*, 22(4):375–421, 2014.

C. S. Vlek, H. Prakken, S. Renooij, and B. Verheij. A method for explaining Bayesian Networks for legal evidence with scenarios. *Artificial Intelligence and Law*, 24(3): 285–324, 2016.

G. A. W. Vreeswijk. Abstract argumentation systems. *Artificial Intelligence*, 90: 225–279, 1997.

W. A. Wagenaar, P. J. van Koppen, and H. F. M. Crombag. *Anchored Narratives. The Psychology of Criminal Evidence*. Harvester Wheatsheaf, London, 1993.

D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue. Basic Concepts of Interpersonal Reasoning*. SUNY Press, Albany (New York), 1995.

D. N. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, Cambridge, 2008.

J. H. Wigmore. *The Principles of Judicial Proof or the Process of Proof as Given by Logic, Psychology, and General Experience, and Illustrated in Judicial Trials. (Second edition 1931.)*. Little, Brown and Company, Boston (Massachusetts), 1913.

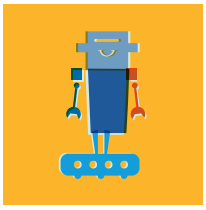
S. L. Zabell. Fingerprint evidence. *Journal of Law and Policy*, 13:143–179, 2005.

Bart Verheij

Arguments
for good artificial intelligence

How can we realize the grand dreams of Artificial Intelligence, without making our worst fears come true? Bart Verheij argues that we need to build machines that can participate in a constructive critical discussion, that tried-and-tested tool for good science, good politics and good family life. Only by developing such argumentation machines can we arrive at an artificial intelligence that provides good answers to our questions, has good reasons for its actions and makes good choices. In this text (presented in the original Dutch and in English translation), Bart Verheij leads us along the right mathematical foundations, Hitchcock's film `To Catch A Thief' and different traditions of legal reasoning. Bart Verheij predicts that by 2025 argumentation systems will have finally closed the long-standing gap between knowledge-based and data-driven artificial intelligence.

The text is the transcription of Bart Verheij's inaugural lecture, read upon accepting the chair of Artificial Intelligence and Argumentation at the University of Groningen. He currently is head of the department of Artificial Intelligence in the Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence and serves as the president of the International Assocation of Artificial Intelligence and Law.



www.ai.rug.nl/~verheij