Contents

Foreword

Rineke Verbrugge's brilliant career can be neatly summed up in a sentence from van Benthem's contribution, "Rineke Verbrugge has blazed a conspicuous trail from theory to reality". Since I too have similarly gone from theory to reality, I must regard her as a fellow traveler.

Thus her work falls fairly neatly into three phases, loosely chronological.

a) Logic and Modal Logic

b) Computer Science and AI

c) Cognitive Science and Social Psychology.

Thus the contributions to this volume in her honor also span the three areas. The contributions from van Benthem, van Ditmarsch, van Eijck, Renardel de Lavalette, and Visser fall inside Logic although some of these authors are also sympathetic to her social side. The sole contribution to Computer Science comes from Dunin-Kęplicz and Szałas. The remaining contributions in Cognitive Science and Social Psychology come from van der Post and van der Vaart, Raijmakers, and from Scholten, Engelen and Hendriks.

In his paper, van Benthem advises us not to fixate on fixed points, but to love and respect oscillations. Imagine an infinite flower and a man plucking petals, one at a time and saying, "She loves me", "She loves me not". He would never graduate to a fixed point and if we understand van Benthem, that is just fine. But I am not sure that van Benthem's own love of oscillations also oscillates.

Van Ditmarsch discusses a variety of bisimulations to be used in a variety of situations running the gamut from epistemic logic to sabotage.

Jan van Eijck considers the relationship between probability and (qualitative) belief, an area of much difficulty as well as interest. He says, "...this is the stuff that Rineke loves."

Renardel de Lavalette discusses the issue of strong completeness in hybrid logics which are not compact. Rineke herself has contributed to this topic as recently as 2009.

The last paper in this subset is the one by Visser who proves a very interesting and technical result about weak theories of arithmetic, an area in which Rineke (and also I) have once worked.

For the second (computer science and AI) area, Dunin-Kęplicz and Szałas discuss a logic of goals treating them as first-order objects rather than formulas as such (the usual convention).

For the last area, van der Post and van der Vaart consider "kill-joy" situations which arise when a situation which appears to involve social cognition is "explained away" by computational modeling. Anthropomorphism may be a

sin, but it is a sin which we humans love to commit! Both Aesop and the Indian fables in the *Panchatantra* are a proof of this temptation.

Raijmakers considers the patterns of strategy use by both children and adults. When children make incorrect judgments, often these misjudgments are not random, but the result of simplified strategies.

And finally Scholten, Engelen and Hendriks consider the issue of irony, drawing in part from the foundational work by Wilson and Sperber. Although Grice is not referenced, I consider this discussion to be Gricean in that context plays an important role in identifying irony as distinct from a bland and sincere statement.

This is a rich collection and a tribute to Rineke's many talents. However, I feel that her career is only midway, that more contributions by her and by her colleagues will follow, and we should expect another volume of tributes in a few years.

– Rohit Parikh

Preface

Factually speaking, the concept of this book got initiated with the following chat conversation:

> Hangout between Sujata Ghosh and Jakub Szymanik
> Monday, June 2, 2014 8:03 PM
>
> **Sujata Ghosh**
> hi!
> how are you?
> **Jakub Szymanik**
> I'm very good! how're you?
> **Sujata Ghosh**
> good.
> **Jakub Szymanik**
> what are you up to?
> **Sujata Ghosh**
> i had a hunch and then got confirmation that Rineke will be 50 next march.
> i was wondering if we can confidentially arrange some celebration :-)

We are both deeply indebted to Rineke Verbrugge for being a great mentor and we quickly realized that we should take this opportunity to celebrate her many contributions to science and to the academic environment. So we set out to organize a workshop and edit a special volume devoted to the areas of Rineke's research interests, including papers developing critical work on her own contributions. After a brief discussion on whether Rineke would like it, we decided to take the risk. Obviously, we could not handle the task in such a short time by ourselves, and so we contacted close collaborators of Rineke. We were overwhelmed by the enthusiastic support we got for the project – this is quite evident in the subsequent pages of the book. The volume is a product of an incredible effort on part of Rineke's teachers, colleagues, students and friends who have all been won over by her ever-encouraging and positive presence in academia and also in daily life.

Pertaining to Rineke's research interests, the book features 9 articles on a wide range of topics – from theories of arithmetic to a study on autism. The papers on hybrid logic, formal theories of belief, probability, goals, social networks, and bisimulations enrich the logic section of the book while papers on cognitive strategizing and social cognition bringing up the cognitive perspective. The themes themselves provide a compelling perception of the vast expanse of Rineke's academic interests and endeavors. A series of personal com-

ments, stories, anecdotes, and pictures constitute the latter part of the book, adding a distinct personal touch to this volume.

We hope that the book shows our appreciation of Rineke's rich and truly interdisciplinary scientific interests spanning from mathematical logic through cognitive science to biology. Rineke is a true role model for how one can successfully and with grace move between so diverse research fields. Rineke is also an amazing group leader: working hard without glorifying the cult of 'being busy' and never forgetting the human element.

As academics we all know how hard it is to finish any project on time. We as a community have a peculiar habit of missing the deadlines and postponing our reactions to editors' requests. Not this time though! While editing this volume we had a feeling that everyone was inspired by Rineke's professionalism. We would like to thank all the people who have made this festschrift volume and the workshop possible.

We start with thanking the contributors for this volume We would also like to thank the reviewers who reacted to our request with extremely short deadline: Katja Abramova, Burcu Arslan, Torben Braüner, Mihir Chakraborty, Andrés Cordón Franco, David Gabelaia, Charlotte Hemelrijk, Thomas Icard, Leszek Kołodziejczyk, Leendert van Maanen, Alexandru Marcoci, Eric Pacuit, Guiseppe Primiero, R. Ramanujam, Jennifer Spenader, Fernando R. Velázquez Quesada, Yanjing Wang, and Marcin Zajenkowski. We express our heartfelt gratitude to one and all. We gratefully acknowledge Rohit Parikh, who kindly agreed to write a foreword for this book. We also thank Lambert Schomaker for the funding and logistics support we are getting from the Institute of Artificial Inteligence, University of Groningen.

This volume could not have happened without the help of Burcu Arslan, who has been a problem-solver for us from day one – whenever we faced any difficulty we depended on her, and Harmen de Weerd who painstakingly formatted the whole volume to its final shape, converted some of the contributions to the LaTeX format, and over all has been the ever-dependable one helping out on numerous occasions. We thank Lina Ghosh for letting us use her decorative designs, Elina Sietsema for being there for all kinds of organizational help, and we also thank Dov Gabbay, Jane Spurr, and the whole team of College Publications for the appearance of this volume.

Last but not least we would like to thank Nicole Baars for her constant support, her help in keeping this project confidential, and yet ensuring the presence of Rineke Verbrugge for the occasion.

21st of January 2015
Chennai Sujata Ghosh
Amsterdam Jakub Szymanik

Academic Section

Oscillations, Logic, and Dynamical Systems

Johan van Benthem [1, 2]

University of Amsterdam & Stanford University
http://staff.fnwi.uva.nl/j.vanbenthem

Abstract

This is a short note with small observations about big questions. We discuss how fixed-point logics, modal and first-order, can describe natural and interesting kinds of dynamic limit behavior in social networks, not just convergence to one end state. We explore what new issues arise then, and how fixed-point logics interface with other mathematical views of dynamical systems. Finally, we discuss how to relate 'blind' network dynamics to behavior of conscious agents exercising their freedom.

1 Introduction: Social agency

Rineke Verbrugge has blazed a conspicuous trail from theory to reality (some recent samples of her road are [14] and [22]), taking dynamic-epistemic logics or logics of games out of their comfort zone to psychological and computational experiments, confronting logical fine-structure and precision with the actual facts of cognition in laboratory situations. But let's get even more real.

Society itself is one great experiment, where individual rationality is rocked by the storms of public opinion, and where long-term and large-group patterns keep emerging, far beyond our individual environment. The interface of individual rationality and statistical large-scale behavior raises difficult, and sometimes disturbing questions. [3]

Now, can the tools of logic play a role in understanding this situation we find ourselves in, say, by taking a look at comprehensible global reasoning

[1] Rineke Verbrugge has already created an impressive intellectual trail, from the logical foundations of mathematics to computational and experimental studies of human agents. While her topics of research may be variable, her standards of quality are constant, winning the minds of many colleagues. However, what wins their hearts is Rineke's character and collegial behavior. Thus, having been won over twice, I am happy to congratulate Rineke, and write in this book in her honor.

[2] I thank the audience at the Workshop 'Trends in Logic' (Beijing, July 2014) for their responses, especially, Samson Abramsky, Paolo Galeazzi, and Phokion Kolaitis. I also thank Yu Junhua (Tsinghua University) for his careful reading of a draft, and Alexandru Baltag for several congenial responses. Two referees also gave helpful comments. Some further debts on specific points are acknowledged in the text.

[3] Just consider current debates about the basis of morality: is good versus bad a matter of deliberative principle, or merely a population equilibrium between predators and prey?

about long-term behavior of agents in dynamical systems, and if so, how should we go about this endeavor? After all, this area has long been the preserve of dynamical system behavior, computational simulation, and evolutionary game theory. In this brief paper, I will make some observations about ways to go – and despite their extreme simplicity, try to convince the reader that there may be something of structure and value here to pursue.

2 Dynamics in networks

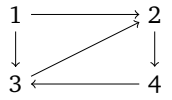
In recent work such as [15,2,10], long-term belief and behavior dynamics has been studied by logical methods in a setting of social networks, where agents' behavior is determined by that of their neighbors according to a given update rule stated in some logical language, a rule which is then applied iteratively. What will happen in the long run?

To develop more concrete intuitions, we look at a few simple cases where a finite network starts with an initial value for some predicate p of nodes, which is then updated according to a logical rule of the form

$$p := \varphi(p)$$

where $\varphi(p)$ is a formula (often taken from a simple modal language) whose universal modality quantifies over all neighbors of the current point in a network. Often p is interpreted as a belief of the agent, but it could stand for any property or short-term behavior.

Example 2.1 A network with a modal influence rule In any network, the modal formula $\Box p$ says that p is currently true at all neighboring nodes. We will see what happens with different initial predicates p in the following simple network, driven by the update rule $p := \Box p$ applied iteratively:



In this dynamics, agents follow what all their neighbors do. Here are some runs that can easily be computed from the above picture with the given rule:

Case 1: initial $p = \{1\}$. The second stage has $p = \emptyset$, and this remains the outcome ever after.

Case 2: initial $p = \{2\}$. The next successive stages are $\{3\}$, $\{4\}$, $\{2\}$, and from this stage onward, we loop.

Case 3: initial $p = \{1, 2\}$. The next stage is $\{3\}$, and we get an oscillation as before in Case 2.

Case 4: initial $p = \{1, 2, 3\}$. We get $\{1, 3, 4\}$, $\{2, 4\}$, $\{2, 3\}$, $\{1, 3, 4\}$, and an oscillation starts here.

We see how network update dynamics can stabilize in one single state (witness Case 1), but also oscillate in loops of successive predicates. These oscillations

come in different forms. Some-times, successive models in the loop are very similar, in fact isomorphic (Cases 2 and 3 have all irreflexive single points) – sometimes the loop runs through different non-isomorphic network configurations (this happened in Case 4, with predicates of different sizes).

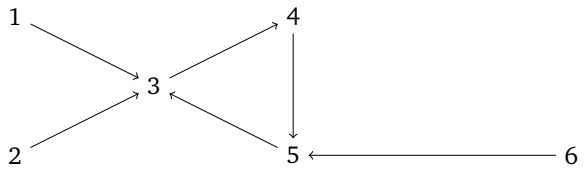
3 Oscillation and its laws

Let us now look directly at what happens in such update dynamics. For simplicity, we will consider *finite models* $\mathbf{M}$ and *ω-sequences* only. [4]

An update rule defines a function F on the power set of the domain of a model $\mathbf{M}$ like above. On finite sets, such functions all have the same pattern:

Fact 3.1 *For any function F on a finite set, there exists a finite family of disjoint loops, at each point of which there may be incoming disjoint F-sequences or F-trees arriving.*

Example 3.2 A function on a finite set Here is a simple example of a loop with incoming arrows:



We are especially interested in the structure of the loops, representing system behavior in the long run. Here 1-loops are fixed-points, a well-known form of system stability, but larger cycles, too, model natural phenomena that are stable in a more general sense, such as periodic swings in public opinion.

To describe this, we explore just one very simple notion:

Definition 3.3 Oscillation operator Given any subset (or viewed slightly differently, any unary predicate) q in a model $\mathbf{M}$, we define

$$OSCp \bullet (\varphi(p), q)$$

as the subset that is the first $F_{\varphi}^{\mathbf{M}}$ oscillation point starting from q. [5]

The oscillation operator satisfies natural fixed-point principles.

Fact 3.4 $OSCp \bullet (\varphi(p), q) \leftrightarrow OSCp \bullet (\varphi(p), OSCp \bullet (\varphi(p), q))$ *is a valid law.*

Further appealing principles of reasoning emerge when we define the following notion that is independent from the starting point:

$OSCp \bullet \varphi(p)$ for 'occurring in some predicate of an oscillation loop of $\varphi(p)$'.

---

[4] The finiteness restriction is a very serious limitation to our approach in this paper, that should be overcome eventually. Some pointers as to how can be found in later passages below.

[5] Further stages of the loop are then definable from this via successive substitutions into φ.

For instance, we have the following valid 'pre-fixed-point law':

$$\varphi(OSCp \bullet \varphi(p)) \rightarrow OSCp \bullet \varphi(p)$$

The preceding observations suggest that there may be a systematic logic to oscillation, a theme that we will explore below. Moreover, studying oscillations is not at odds with studying fixed-points.

Discussion. Fixed-points of set functions The oscillation operator relates naturally to well-known notions from the literature on fixed-points. To see this, consider maps on our models. We call a set X a 'fixed-point' for a function F if $F(X) = X$. A widely used fact in logic is that, for all inclusion-monotonic maps F on a power set, there are smallest and greatest fixed-points, as stated by the well-known Tarski-Knaster Theorem. As with oscillation, we can think of such F as defined by special predicates $\varphi(p)$, this time with p occurring only *positively*. Then we get, e.g., the following observation:

Fact 3.5 *Smallest fixed-points $\mu p.\varphi(p)$ can be defined as follows for formulas $\varphi(p)$ with p occurring only positively: $\mu p.\varphi(p) := OSCp \bullet (\varphi(p), \bot)$*

Still, this is just a start, and there is more going on here in terms of valid laws than may be obvious at a first glance. For instance, with some slight abuse of notation, smallest fixed-points satisfy the equation

$$F(\mu p.\varphi(p)) = \mu p.\varphi(p) \qquad \text{where } F(X) = \{s \in \mathbf{M} \mid \mathbf{M}[p := X], s \models \varphi(p)\}$$

Now it is interesting to see that, despite initial appearances, the earlier law $OSCp \bullet (\varphi(p), q) \leftrightarrow OSCp \bullet (\varphi(p), OSCp \bullet (\varphi(p), q))$ that we noted for oscillation is not of this kind. Its underlying approximation procedure rather refers to a binary function $F(X, Y)$ where X is the current stage, and Y the initial stage, and its format is about replacing the initial Y by some other predicate, not the running X. The final version of this paper will contain further observations about this issue of unary versus binary functions, and matching different kinds of fixed-points – but for now, it is only meant as an appetizer.

More important still is the following issue concerning a natural generalization.

Discussion. From finite to infinite models In infinite models, approximation can go on beyond the first ω steps, and the question then arises how to define the limit stages. The usual stipulations in fixed-point logics such as taking unions or intersections seem to make little sense when we allow oscillation, and we need other ideas. There are interesting analogies here with similar liftings to the infinite in philosophy, logic, and game theory. [6]

[6] Some obvious analogies are with the limit steps required in Kripke-style and Gupta-Hertzberger revision theories of truth [26,18,21], that take lim-sups or lim-infs. Related issues of generalization arise in game theory with iterative solution concepts on infinite games (e.g., iterated removal of strictly dominated strategies): cf. [27,9], and for a general analysis [1]. Also related is work on common knowledge in iterations beyond the ordinal ω: cf. [20,8]. As for a more radical logical treatment, Alexandru Baltag (p.c.) has suggested making the definition of the limit jump itself an explicit parameter in the language.

Still, when we are interested in the behavior of dynamical systems, only the first ω evolution steps matter, since there are no further stages in the behavior of real systems over time. Though this seems at odds with standard logics of the sort to be discussed now, it generates interesting issues of its own – some of which are touched upon in Section 6 below. However, we do not pretend to solve the issue of the proper infinite perspective in this paper.

4 Stability and fixed-point logics

The idea of approximation to reach a stable state of some logically defined operator on models is not at all new. It underlies well-known logics of fixed-points in the literature, of which there are two main varieties. We will take these as role models for an 'oscillation logic'.

The system *LFP*(FO) enriches first-order logic with operators for smallest and greatest fixed-points of monotonic operations, which exist in any model by the Tarski-Knaster Theorem. These operations are defined syntactically by formulas $\varphi(P)$ of the formal language in which all occurrences of the predicate P in φ are syntactically positive: see [13], while [16] provides broader background in the theory of infinite computations. Whereas *LFP*(FO) is of high computational complexity (its satisfiability problem is Π^1_1-complete), a modal version of the same idea gives rise to the well-known decidable system of the modal μ-*calculus* (cf. [31]) whose syntax works as follows.

A smallest fixed-point formula $\mu p.\varphi(p)$ (with p occurring only positively in φ) denotes the smallest fixed-point of the following operation in the lattice of all subsets of a given model $\mathbf{M}$:

$$F^{\mathbf{M}}_{\varphi}(X) = \{s \in \mathbf{M} \mid \mathbf{M}[p := X], s \models \varphi\}$$

One can view smallest fixed points of such a function as the first stage in a possibly infinite cumulating approximation sequence where applying the function F no longer changes the current set:

$\emptyset, F(\emptyset), F^2(\emptyset), \ldots, F^{\alpha}(\emptyset)$

where at limit ordinals α, we take the union of all preceding stages.

The modal μ-calculus has been axiomatized completely, with proof principles:

$\varphi(\mu p.\varphi(p)) \leftrightarrow \mu p.\varphi(p)$	Fixed-point axiom
if $\vdash \varphi(\alpha) \rightarrow \alpha$, then $\vdash \mu p.\varphi(p) \rightarrow \alpha$	Smallest fixed-point rule

Similar laws govern reasoning with dual operators $\nu p.\varphi(p)$ for greatest fixed-points, definable as $\neg\varphi(\neg\mu p.\varphi(\neg p))$. In this case, the approximation sequence starts at the whole universe of the model.

As we have noted, the emphasis in these logics is on reaching fixed-points, stable stages in the approximation process where the same set returns. However, this stability can be fragile, even with our special positive syntax. If we start the approximation sequence in an arbitrary initial predicate, there is no guarantee that even monotone transformations reach a fixed point.

Fact 4.1 *Monotone set transformations can oscillate forever when the initial input is non-trivial.*

A counterexample occurred in Section 2. Just notice that the positive modal formula $\Box p$ kept oscillating when started at non-trivial input predicates. [7]

Remark 4.2 Extended μ-calculus By the preceding fact, our oscillation perspective suggests a fresh look at existing logical systems. Alexandru Baltag (p.c.) has suggested an extended μ-calculus with operators $OSCp \bullet (\varphi(p), q)$ in which the formula $\varphi(p)$ has only positive occurrences of p. Modulo some definitional subtleties to be mentioned below, this extension makes sense, and there is some interesting structure here. For instance, the set of predicates in a loop forms an anti-chain, as is easy to see. [8]

Going still further, there have been generalizations of fixed-point logics which can deal with arbitrary formulas that need not induce monotone set transformation, just as in our network dynamics. However, such systems, such as *inflationary fixed-point logic IFP*, still enforce cumulative growth of successive approximations by means of the following stipulation:

$$F^{\mathbf{M}}_{IFP}(X) = F^{\mathbf{M}}(X) \cup X$$ [9]

Basic results about generalized fixed-point logics include the theorem that *IFP*(FO) is equal in expressive power to *LFP*(FO) (cf. [25]) – though there is still a procedural difference: recursion in the defining formulas runs over auxiliary predicates with higher arities. [10]

From the viewpoint of fixed-point logics, oscillations seem mostly like 'junk' or failure in an approximation process. What happens when we add systematic syntax for them, to get richer logical systems? In the following section, we explore this line of thought a little bit.

5 Oscillation in logical systems

The oscillation operator OSC seems a natural addition to the syntax of logical systems, and we will do so now. But caution is needed, as we have not given a general definition of OSC on arbitrary infinite models – due to problems at limit ordinals. [11] In what follows, we will stick with our earlier restriction to *finite models*. Still, many of the systems to be considered can also define loop structure in infinite models, in particular, *infinitary modal logic*. We leave it to the reader to see which of our observations generalize straightforwardly.

---

[7] Monotonicity only starts producing cumulation thanks to the starting inclusion $\emptyset \subseteq F(\emptyset)$.

[8] There may be connections here with 'partial fixed-point logics' in computer science, [24].

[9] We suppress the reference to the defining formula $\varphi(p)$ here for perspicuity of notation.

[10] However, adding inflationary fixed-points to the less expressive system of the modal μ-calculus does increase the latter system's expressive power, cf. [12].

[11] Failing a good transfer convention across limit ordinals, we couldn define a uniform 'finite-oscillation operator' $OSCp \bullet (\varphi(p), q)$ in all models, saying that the φ-approximation sequence starting from q reaches a finite loop at some finite stage. On infinite models, the latter need not always happen, as the first ω-sequence for φ might not loop.

5.1 Modal logic

Many natural cases of network dynamics work with modal update rules. Starting from this simple setting, then, we add an operator $OSCp \bullet (\varphi(p), \psi)$ to the syntax of basic modal logic, with a semantic meaning as given above. The oscillation operator fits well in a modal setting.

Fact 5.1 *Modal logic with an added oscillation operator is invariant for total bisimulations whose domain and range are the whole models.*

Proof. This can be proved by a direct argument, or by noting that the above truth definition of the oscillation operator can also be written explicitly in an infinitary modal logic with an added *universal modality*, a language which is invariant for total bisimulations. □

This modal character is reinforced by further features. In particular, using the oscillation operator as shown in Section 4, our oscillation logic extends the modal μ-calculus.

Fact 5.2 *Smallest fixed-points $\mu p.\varphi(p)$ can be defined as $OSCp \bullet (\varphi(p), \bot)$.*

Thus, the logically valid laws of oscillation immediately include the laws for fixed-points. We suspect that a converse definition is not possible, though we only have a loosely related observation.

Fact 5.3 *The finite-oscillation operator is not definable in the μ-calculus.*

Proof. The reason is that, when added, the enlarged system loses the finite model property which the modal μ-calculus possesses. Here is a concrete counter-example in the enlarged language. The formula

$$\mu p.\Box p \wedge \neg OSCp \bullet (\Box p, \bot)$$

has infinite models, where in fact it forces the 'well-founded core' is infinite, but this formula lacks finite models. □

These are just simple observations, and open problems abound. In particular,

Question. *Is the modal oscillation calculus decidable, or is it at least axiomatizable, on the class of finite models?*

Remark 5.4 Inflationary μ-calculus Next, we can also embed the inflationary μ-calculus. We can mimic inflationary approximation for arbitrary formulas $\varphi(p)$ in our network dynamics by means of disjunctive formulas

$$p := \varphi(p) \vee p$$

Formulas $OSCp \bullet (\varphi(p) \vee p, q)$ then define smallest inflationary fixed-points, reached from an initial predicate q. We suspect that a converse still fails, and that the oscillation operator is undefinable even with inflationary fixed-points.

Discussion. Fine-structure: bisimulation loops One can also pursue new kinds of issue. As we saw in Section 2, larger loops can be of different kinds. Sometimes, they are close to fixed-points as all models in the loop are isomorphic, like in all our initial examples. More relevant to the modal setting:

The successive stages in a loop can be *bisimilar* models. [12]

Here, we are not saying that identity is a bisimulation in the loop: individual points may still behave differently from one stage to another. Nevertheless, at a certain description level, the models in the loop are indeed the same, having reached a stable theory modulo bisimulation. [13]

On finite models, the models in a bisimulation loop have the same collection of 'modal types', though they may differ in which object exemplifies which type. Bisimulation loops consist of models with the same theory in the following syntax. Take the basic modal language with an added universal modality, and consider only 'global formulas', true or false throughout a model. While world-dependent formulas can still change truth values at the same world in different models of a bisimulation loop, there is no detectable change in global syntax, since the set of available modal types does not change in the loop.

Here is one interesting question out of many that arise in this perspective:

Problem 5.5 *Is there special syntax for oscillation operators that guarantees generalized fixed-points in the form of bisimulation loops?* [14]

The more general point, however, is this:

> Oscillation suggests the use of several logical languages, at different levels of detail, providing different invariants for the network dynamics. [15]

But one can also focus on the influence of the graph structure, and ask, for instance, for which graphs all modal formulas stabilize their oscillation loops when started anywhere.

Conjecture 5.6 *The graphs with guaranteed stabilization for all modal update rules are precisely the finite trees.*

Similar questions of oscillation logic arise for update formalisms for richer network update rules, such as 'graded modal logic' that counts numbers of neighbors, or modal logics of 'most' (cf. [29]). One special extension deserves separate attention here, as with fixed-point logics.

5.2 First-order logic

This time, we do not restrict attention to finite models, but take the other route mentioned in Footnote 11 above. First-order logic plus a finite-oscillation operator is of high complexity. We merely note two facts.

Fact 5.7 *The finite-oscillation operator on arbitrary models is definable in the infinitary first-order logic* $L_{\omega 1 \omega}$.

---

[12] Actually, bisimulation also makes sense in non-looping iteration sequences in infinite models, as a sort of generalized fixed-point. This, too, seems a natural concept.

[13] For this stability in higher languages, compare dynamical systems in biology where we consider a system stable when the percentages of different types of animal no longer change.

[14] We can also vary such issues, and ask which syntactic types of formulas guarantee the existence of 1-loops (i.e., fixed-points) when started at any predicate in any model.

[15] It may even be true that, at some appropriate higher level of description, in a lattice with other approximation operators, loops become ordinary monotone fixed-points again.

Proof. By direct description. For every formula $\varphi(p)$ and predicate q, one can define the n-th iteration φ^n for any finite n by successive substitution. Now one says there is some n and k where for all objects in the model, φ^n holds iff φ^{n+k} holds, and then that n is the smallest number with this property. All this can be formulated in $L_{\omega 1 \omega}$. □

Fact 5.8 *First-order logic with the oscillation operator is non-axiomatizable.*

Proof. Consider the modal μ-calculus formula $\mu p.\Box p$ that defines the upward well-founded part of the accessibility relation on any model for the modality $\Box$. Under the standard translation for modal logic, its modal part $\Box p$ is first-order. Without loss of generality, we can think of this modality as looking downward in the ordering. By an earlier observation, the formula $\mu p.\Box p$ is definable using the oscillation operator. Now consider the statement that

"all objects satisfying $\mu p.\Box p$ satisfy $OSCp \bullet (\Box p, \bot)$"

This says that every object in the well-founded part is admitted after finitely many iteration steps. But this can only happen when the upward well-founded chains are finite. And this property enforces, on models satisfying the first-order theory of 'greater than' on the natural numbers, that the model actually is a copy of the natural numbers. But then, the validities of the logic encode arithmetical truth, which is non-axiomatizable – and in fact Π^1_1-complete. □

6 Further logical perspectives

We pursued one straightforward way of adding oscillation operators to standard languages. However, there are also other natural technical perspectives on what is going on. We pursue this a little bit to show the broader circle of ideas that we have entered in this paper.

6.1 Dynamic logic of substitutions

An alternative approach would focus on the basic dynamic act itself that drives the above network dynamics, which is a *predicate substitution*

$$p_{new} := \varphi(p_{old})$$

One can study dynamic substitutions like this in a system of dynamic-epistemic logic (cf. [7]) with dynamic modal operators

$$\langle p := \varphi(p) \rangle$$

The valid laws for the basic predicate substitution modality form a simple decidable calculus *DEL(subst)* whose axioms mirror the usual recursive clauses for syntactic substitution. [16] In more complex versions, substitution actions can also be sequentially composed and even finitely iterated. The resulting system can define the notion of oscillation as defined above.

[16] We claim no originality for this system. Various dynamic-epistemic logics that deal with substitutions occur in the literature.

Fact 6.1 *The oscillation operator $OSCp \bullet (\varphi(p), q)$ is definable in DEL(subst).*

Adding step by step sequential composition still leaves this calculus simple. But adding arbitrary finite iteration of substitutions introduces complexity.

Fact 6.2 *DEL(subst*) is non-axiomatizable, and in fact Π^1_1-complete.*

Proof. The reason is that, using a simple translation, the logic *DEL(subst*)* faithfully embeds the better-known system of public announcement logic with iterations *PAL**, whose complexity is of this sort (cf. [28]). □

Even so, fragments of dynamic substitution logics with iteration might well be good tools for analyzing network limit dynamics driven by special formulas. Also relevant is the following observation made by Alexandru Baltag (p.c.): substitution logic meshes well with oscillation logic.

Fact 6.3
The equivalence $\langle q := \psi \rangle OSCp \bullet (\varphi(p), q) \leftrightarrow OSCp \bullet (\langle q := \psi \rangle \varphi(p), \psi)$ is valid.

6.2 Modal logic of dynamical systems

Fixed point logics are natural candidates for describing dynamical systems – since their laws are often simple, and yet pack quite a lot of explanatory power. [17] But there are alternatives. An earlier approach to dynamical systems is the system *DTL* [23], with a simple modal language that capture basic results on dynamical systems such as the Poincaré fixed-point theorem. The base language is more global than ours, with operators

$$O\varphi, \Box\varphi$$

These are a temporal operator $O\varphi$ for the next state of some continuous operator on the state space, plus a modality $\Box\varphi$ for topological interior. The handbook chapter [23] surveys the resulting logics on special spaces, as well as language extensions such as finite iterations of the system dynamic operator O. This modal zooming out on basic structures in dynamical systems lies at an abstraction level above our fixed-point or substitution logics in the above.

There is a challenge of how to interface perspectives, since *DTL* adds important structure that we have left out. In particular, our networks with neighborhoods also support *DTL*'s topological structure, and this seems important since limit behavior is definitely influenced by two factors: (a) the logical form of the update rule, and (b) the network structure that these work on.

For more about interfacing logic and dynamical systems: see Section 7.

6.3 Temporal logic and histories of dynamical systems

Finally, while we have emphasized sparse modal languages in this paper, richer lines exist. For instance, consider the rich temporal logic of [19] for players in *iterated matrix games* responding to observed moves by others in the preceding round. There is a clear intuitive connection with social network evolution,

[17] For further examples of the surprising power of basic modal fixed-point laws in capturing essences of results in game theory or social networks, cf. [32,2].

whose precise statement goes beyond the compass of this paper. [18] Right here, our main point is just one of system choice. Temporal logics from the computational field of agency explicitly describe properties of countable histories or runs of a multi-agent process, in the format

$$\mathbf{M}, h, s \models \varphi \qquad \text{formula } \varphi \text{ is true at point } s \text{ on history } h \text{ in model } \mathbf{M}.$$

In the same manner, we could model our network evolution in temporal logic, and describe the earlier oscillation patterns in such a richer explicit formalism.

Digression. Merging perspectives What the preceding suggests is that we can use temporal logics as a sort of meta-theory for modal fixed-point logics, and represent simple notions and proofs in this richer logic. Benchmarks would be many of the simple observations in earlier sections. More ambitiously, we can also study mixtures of modal fixed-point logics and temporal logics for their computation procedures. This combination seems natural since, despite our earlier problem of defining transfer steps at limit ordinals, histories of the simplest infinite type ω fit fixed-point logics very well, witness the infinite evaluation games for the modal μ-calculus discussed in [31].

More generally, merged temporal and fixed-point logics may provide a rich reasoning style for social systems viewed at different levels.

7 Enriching the framework

Our analysis has been confined to basic logical systems that might deal with limit phenomena in social networks with update rules. We have suggested that this may be a good high-level perspective for getting qualitative insights that lie behind results obtained with the numerical models used in dynamical systems approaches to social phenomena. Of course, much more can be said about comparing qualitative logical and quantitative mathematical methods in this area, since the two methods come with different agendas. One striking difference is that numerical update rules in networks like those of the classic De Groot [17] tend to 'smoothen' values for strength of belief, whereas discrete logical approaches may create more drastic oscillations. We just note this for now, but this is obviously a point that needs much more reflection.

Next, as we said right at the start of this paper, the rules we studied are blind operations on unstructured points. What about the internal nature of the agents that make up the social network? A richer source of modeling agents than we have followed here exists in computational logic where notions from automata theory could enrich our current view (cf. [16]). This connection gets even richer when we consider the computational games associated with the logical systems that we have considered here. [19]

---

[18] This can be spelled out in precise detail, relating network update rules parametrized to individual points to strategy profiles for players, but we leave this for another occasion.

[19] In this connection, note also that oscillation patterns are also standard in automata theory, say with 'parity automata' for the modal μ-calculus, cf. [31]. Such patterns might be used, say, to obtain finer denotations and finer intensional notions of formula equivalence.

But even with automata in place, network dynamics remains austere. It would not distinguish between update rules for human agents, schools of fish, or neural networks. An obvious further focus then is actions that make us human, such as making observations, deliberating and deciding what to do on the basis of knowledge and beliefs about others, rather than just mechanically following our environment. [20] Moreover, human agents pursue goals connected to their preferences, while guided by intentions toward reaching these goals. All of this typically shows in their making *choices*, less or more rational.

To model real agency, we are not left with empty hands. Current dynamic epistemic logics are well up to extensions with informational actions, preferences, and acts of decision making (cf. [4] for a general treatment of logic of agency in this style – or for specific network examples, [15,11,3]). Moreover, one can draw on a flourishing literature connecting logic and game theory (cf. [5] and the references therein), giving agents positioned in networks choices as to what to do at each stage, with strategy profiles corresponding to update rules that can be studied for their long-term success in terms of achieving goals.

This richer view of agency is realistic, but pursuing it would take us far beyond the scope and intentions stated at the beginning of this note. Moreover, there is a risk in rushing ahead, of downplaying the virtues of blind rules and automatic updates. In the cognitive life of human agents, there is a systematic switching dynamics between conscious deliberate action and automated skills or habits – because of limited attention, or for more positive reasons of saving labor. Likewise, social life would probably be impossible without some back and forth between relegating beliefs and decisions to an 'automatic pilot', versus returning them to the realm of explicit control. [21] [22]

8 Conclusion

The point of this paper is that long-term social behavior supports reasoning patterns that invite logical analysis. To do so, we must step back from fixed-points only, and see the logical structure in oscillations: cycles are not 'junk', but regular long-term behavior in its own right. We have noted a few facts and perspectives that may help us do so – suggesting that existing fixed-point logics, suitably generalized, and supplemented with dynamic and temporal logics for system evolution, may apply to many realms of limit behavior over time.

There are several ways of taking what is proposed here, that can be pursued in tandem. One is exploring new technical views of logical systems and their connections, for which we have provided a slew of suggestions. Another line is a richer description of agency, either as logical theory about agents in social

[20] A real human agent can even decide *not* to update according to some prevalent update rule in the network, thereby exercising her freedom.

[21] I thank Erik Olsson for a stimulating discussion of this point in social agency, and beyond.

[22] The purely temporal approach in this paper also neglects another dimension of the social world, that of *size*: and in particular, the interface between the individual agents and large groups. Group size in networks poses questions that are far from being exhausted by current studies of games or group knowledge (cf. [30]).

settings, or as an account of how agents reason themselves. The way I myself would like to think about the role of logic here is as providing natural levels for identifying qualitative reasoning patterns with broad sweep and simplicity. As we just noted at the end of Section 7, there may be many such levels, from automated to deliberate. [23]

Despite the technicality of this paper, I hope that its topics still connect to the challenging interface of individual agency and social life that I started with. I feel that much can be done by logicians today in understanding, and perhaps even improving, the 'thin layer' of deliberate human thinking and acting that lies so precariously in between the blind dynamics of the social systems above us and the neural networks inside us.

References

[1] Apt, K. and J. Zvesper, *Proof-theoretic analysis of rationality for strategic games*, in: *Proceedings of the 11th International Workshop on Computational Logic in Multi-Agent Systems (CLIMA XI)*, Lecture Notes in Computer Science **6245** (2010), pp. 186–199.

[2] Baltag, A., Z. Christoff, R. Rendsvig and S. Smets, *Dynamic epistemic logic for threshold models* (2010), working paper, ILLC, University of Amsterdam. Extended version of a paper presented at the ELISIEM Workshop on Epistemic Logic for Individual, Social, and Interactive Epistemology, 26th European Summer School in Logic, Language and Information, Tübingen, Germany, August 11-15, 2014.

[3] Baltag, A., F. Liu and S. Smets, *Evidence-based belief change in networks* (2014), working paper, ILLC, University of Amsterdam and Department of Philosophy, Tsinghua University.

[4] van Benthem, J., "Logical Dynamics of Information and Interaction," Cambridge University Press, Cambridge, UK, 2011.

[5] van Benthem, J., "Logic in Games," MIT Press, Cambridge, MA, 2014.

[6] van Benthem, J., *Natural language and logic of agency*, Journal of Logic, Language and Information **23** (2014), pp. 367–382.

[7] van Benthem, J., J. van Eijck and B. Kooi, *Logics of communication and change*, Information and Computation **204** (2006), pp. 1620–1662.

[8] van Benthem, J. and D. Sarenac, *The geometry of knowledge*, in: J.-Y. Béziau, A. Costa-Leite and A. Facchini, editors, *Aspects of Universal Logic*, Centre de Recherches Sémiologiques, Université de Neuchatel, 2005 pp. 1–31.

[9] Chen, Y.-C., N. V. Long and X. Luo, *Iterated strict dominance in general games*, Games and Economic Behavior **61** (2007), pp. 299–315.

[10] Christoff, Z. and J. U. Hansen, *A logic for diffusion in social networks*, Journal of Applied Logic **13** (2015), pp. 48–77.

[11] Christoff, Z., J.-U. Hansen and C. Proietti, *Reflecting on social influence in networks* (2014), talk at the IDAS workshop on Information Dynamics in Artificial Societies (ESSLLI 2014), to appear in special issue of the Journal of Logic, Language and Information, edited by E. Lorini, L. Perrussel and R. Muehlenbernd.

[12] Dawar, A., E. Grädel and S. Kreutzer, *Inflationary fixed points in modal logic*, ACM Transactions on Computational Logic **5** (2004), pp. 282–315.

[13] Ebbinghaus, H.-D. and J. Flum, "Finite model theory," Springer Science Publishers, 2005.

[14] Ghosh, S., B. Meijering and R. Verbrugge, *Empirical reasoning in games: Logic meets cognition*, in: T. Ågotnes, N. Alechina and B. Logan, editors, *Proceedings Third Workshop Logics for Resource Bounded Agents*, 2010, pp. 15–34.

[23] This style of thinking is connected to finding 'natural logics' of agency (cf. [6] on extending linguistic monotonicity inferences) where we identify simple pervasive concepts and reasoning patterns in natural language that enable us to function in a complex world of communication.

[15] Girard, P., F. Liu and J. Seligman, *Logical dynamics of belief change in the community*, Synthese **191** (2014), pp. 2403–2431.
[16] Grädel, A., E. Thomas and T. Wilke, editors, "Automata, Logics, and Infinite Games," Lecture Notes in Computer Science, Springer Verlag, Berlin, 2002.
[17] de Groot, M., *Reaching a consensus*, Journal of the American Statistical Association **69** (1974), pp. 118–121.
[18] Gupta, A., *Truth and paradox*, Journal of Philosophical Logic **11** (1982), pp. 1–60.
[19] Gutierrez, J., P. Harrenstein and M. Wooldridge, *Reasoning about equilibria in game-like concurrent systems*, in: C. Baral, G. De Giacomo and T. Eiter, editors, *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning (KR-2014)*, 2014, Vienna, Austria.
[20] Heifetz, A. and D. Samet, *Knowledge spaces with arbitrarily high rank*, Games and Economic Behavior **22** (1998), pp. 260–273.
[21] Herzberger, H., *Notes on naive semantics*, Journal of Philosophical Logic **11** (1982), pp. 61–102.
[22] Isaac, A., J. Szymanik and R. Verbrugge, *Logic and complexity in cognitive science*, in: A. Baltag and S. Smets, editors, *Johan van Benthem on Logic and Information Dynamics*, Springer, Dordrecht, 2014 pp. 787–833.
[23] Kremer, P. and G. Mints, *Dynamic topological logic*, in: M. Aiello, I. Pratt-Harman and J. van Benthem, editors, *Handbook of Spatial Logics*, Springer Verlag, Dordrecht, 2007 pp. 565–606.
[24] Kreutzer, S., *Partial fixed-point logic on infinite structures*, in: *Proceedings of the 16th International Workshop and 11th Annual Conference of the EACSL on Computer Science Logic (CSL'02)*, Springer, 2002, pp. 337–351.
[25] Kreutzer, S., *Expressive equivalence of least and inflationary fixed-point logic*, Annals of Pure and Applied Logic **130** (2004), pp. 61–78.
[26] Kripke, S., *Outline of a theory of truth*, Journal of Philosophy **72** (1975), pp. 690–716.
[27] Lipman, B., *A note on the implications of common knowledge of rationality*, Games and Economic Behavior **6** (1994), pp. 114–129.
[28] Miller, J. and L. Moss, *The undecidability of iterated modal relativization*, Studia Logica **79** (2005), pp. 373–407.
[29] Pacuit, E. and S. Salame, *Majority logic*, in: *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR 2004)*, 2004, pp. 598–605.
[30] Ramanujam, R., *Logical player types for a theory of play*, in: A. Baltag and S. Smets, editors, *Johan van Benthem on Logic and Information Dynamics*, Springer, 2014 pp. 509–528.
[31] Venema, Y., *Lectures on the modal μ-calculus* (2007), lecture notes, ILLC, University of Amsterdam.
[32] Zvesper, J., "Playing with Information," Ph.D. thesis, ILLC, University of Amsterdam (2010).

Five Funny Bisimulations

Hans van Ditmarsch [1]

LORIA, CNRS / University of Lorraine
France

Abstract

In this survey we present various recent work proposing adjustments to the standard notion of bisimulation in order to have proper structural correspondents with epistemic, or epistemically motivated, modalities: contingency bisimulation, awareness bisimulation, plausibility bisimulation, refinement, and bisimulation for sabotage.

Keywords: Modal logic, bisimulation, multi-agent systems, epistemology

1 Introduction

In modal logics there is often a direct relation between the modality and the accessibility relation: $\Box\varphi$ is true in a state s if φ is true in all states t such that Rst, where $\Box$ is the modality and R the binary accessibility relation (where we write Rst for $(s,t) \in R$). In such logics the notion for sameness of structures with respect to the logical language is *bisimulation*. For interaction-free multi-modal modal logics with operators $\Box_a$ for $a \in A$, where A is a set of labels, this correspondence between $\Box_a$ and R_a remains the case and we require that the bisimulation clauses hold for all R_a. In such a logic we then have that bisimilarity implies modal equivalence (of pointed relational structures) and that, on image-finite (or modally saturated) structures, modal equivalence implies bisimilarity.

In logics of knowledge, $\Box_a\varphi$ stands for 'agent a knows φ', and there are many other epistemic notions with corresponding modalities, such as belief, explicit knowledge, and safe knowledge. Also there are group notions of knowledge such as common and distributed knowledge. And in those epistemic settings there may additionally be other modalities, for change of knowledge, such as public announcement, private announcement, belief revision, and ontic actions. The modalities in such epistemic logics often do not directly correspond to an accessibility relation but are somehow defined using the more primitive set of $\Box_a$ modalities or, directly as an operation on the set of R_a accessibility relations. The correspondence between modal equivalence and bisimilarity

[1] We thank the reviewers for their comments. We acknowledge support from ERC project EPS 313360. Hans is also affiliated to IMSc, Chennai, India, as research associate. Email: hans.van-ditmarsch@loria.fr.

may then break down, and in order to reestablish it we may have to adjust the clauses for bisimulation. This is well-known for group notions of knowledge, as they are infinitary modalities that are also well-studied in other modal logical settings, such as PDL. But there are many other cases of interest, even (unlike the case of infinitary modalities) involving very simple finite structures.

For example, consider the case of contingency logic, also known as the logic of knowing whether. It has primitive modalities $\Delta\varphi$, for 'φ is non-contingent', or 'the agent knows whether φ', with semantics that $\Delta\varphi$ is true in a state s if and only if φ has the same truth value in all accessible states, i.e., for all t, u such that Rst and Rsu, either φ is true in t and u or φ is false in t and u. Now consider the following structures $\mathcal{M}$ and $\mathcal{M}'$ (the names of states precede the valuation of propositional variable(s), where $\overline{p}$ means that p is false):

$$\mathcal{M} \qquad\qquad\qquad\qquad \mathcal{M}'$$

$$s : p \longrightarrow t : p \qquad\qquad s' : p \longrightarrow t' : \overline{p}$$

We now have that in $\mathcal{M}$, in all states accessible from s, p has the same value (namely true, in the unique accessible state t), whereas in $\mathcal{M}'$, in all states accessible from s', p also has the same value (namely false, in the unique accessible state t'). In an epistemic setting we would say that the agent knows whether p, in both s and s'. We also have that p is true in both s and s', and it is easy to see that $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ are modally equivalent in contingency logic. (For a pointed model, a pair $(\mathcal{M}, s)$ consisting a model $\mathcal{M}$ and a designated state s in its domain, we use notation $\mathcal{M}_s$.) But they are clearly not bisimilar in the standard sense, as the value of p is different in t and t'. We need a weaker (or, in general, different) notion of bisimulation for contingency logics, *contingency bisimulation*, such that $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ are contingency bisimilar. It says that the forth or back condition need only apply if there are non-bisimilar accessible states. We think it is a bit funny.

In this survey we present five funny bisimulations, i.e., five adjustments to the standard notion of bisimulation in order to have proper structural correspondents with epistemic, or epistemically motivated, modalities: contingency bisimulation, awareness bisimulation, plausibility bisimulation, refinement, and bisimulation for sabotage.

2 Basic multi-agent modal logic and standard bisimulation

Let a countably infinite set of propositional variables $P = \{p, q, \dots\}$ and a disjoint finite set of agents $A = \{a, b, \dots\}$ be given.

Definition 2.1 Model A *model* for A and P is a triple $M = (S, R, V)$ that consists of a *domain* S of (propositional) *states* (or 'worlds'), an *accessibility* function $R : A \to \mathcal{P}(S \times S)$, and a *valuation* function $V : P \to \mathcal{P}(S)$. For $R(a)$ we write R_a, and for $(s, t) \in R_a$ we also write $R_a st$, or $t \in R_a s$ (or $t \in R_a(s)$).

Accessibility function R can be seen as a set of *accessibility relations* R_a, and V as a set of *valuations* $V(p)$. A pointed model is a pair $(\mathcal{M}, s)$, where $s \in S$; we write this as $\mathcal{M}_s$. A model $\mathcal{M}$ is *image finite*, if for all $s \in S$ there is only a finite amount of states $t \in S$ such that $R_a st$.

Definition 2.2 Bisimulation Let models $\mathcal{M} = (S, R, V)$ and $\mathcal{M}' = (S', R', V')$ be given. A *bisimulation* is a non-empty relation Z between $\mathcal{M}$ and $\mathcal{M}'$ (i.e., $Z \subseteq S \times S'$) such that for all $s, s' \in S$ such that Zss', and for all $a \in A$:

atoms	for all $p \in P$, $s \in V(p)$ iff $s' \in V'(p)$;
forth	if $t \in S$ and $R_a st$, then there is a $t' \in S'$ such that $R'_a s't'$ and Ztt';
back	if $t' \in S'$ and $R'_a s't'$, then there is a $t \in S$ such that $R_a st$ and Ztt'.

If Zss', we say that pointed models $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ are bisimilar. If there is a bisimulation linking $\mathcal{M}$ and $\mathcal{M}'$ we write $\mathcal{M} \leftrightarrows \mathcal{M}'$, and between pointed models we write $\mathcal{M}_s \leftrightarrows \mathcal{M}'_{s'}$. If Z is that bisimulation we may also write $Z : \mathcal{M} \leftrightarrows \mathcal{M}'$ and $Z : \mathcal{M}_s \leftrightarrows \mathcal{M}'_{s'}$, respectively.

Definition 2.3 Language The language $\mathcal{L}(\Box)$ of multi-agent modal logic is inductively defined as

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box_a \varphi$$

where $p \in P$ and $a \in A$. We employ the usual abbreviations to define $\top$, $\bot$, $\vee$, $\rightarrow$, $\leftrightarrow$, and the dual modality $\Diamond_a$. Without the inductive clause for $\Box$ we get the language $\mathcal{L}$ of propositional logic.

Definition 2.4 Semantics Let $\mathcal{M} = (S, R, V)$ and $\varphi \in \mathcal{L}(\Box)$ be given. Then:

$$\begin{array}{ll} \mathcal{M}_s \models p & \text{iff } p \in V(s) \\ \mathcal{M}_s \models \neg\varphi & \text{iff } \mathcal{M}_s \not\models \varphi \\ \mathcal{M}_s \models \varphi \wedge \psi & \text{iff } \mathcal{M}_s \models \varphi \text{ and } \mathcal{M}_s \models \psi \\ \mathcal{M}_s \models \Box_a \varphi & \text{iff } \mathcal{M}_t \models \varphi \text{ for all } t \in S \text{ such that } R_a st \end{array}$$

If $\mathcal{M}_s \models \varphi$ for all $s \in S$ then we write $\mathcal{M} \models \varphi$ (φ is *valid on model* $\mathcal{M}$). If $\mathcal{M} \models \varphi$ for all $\mathcal{M}$ (of that class, given P and A) we write $\models \varphi$ (φ is *valid*). Write $[\![\varphi]\!]_{\mathcal{M}}$ for $\{s \in S \mid \mathcal{M}_s \models \varphi\}$. Two pointed models $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ are *modally equivalent* if for all $\varphi \in \mathcal{L}(\Box)$, $\mathcal{M}_s \models \varphi$ iff $\mathcal{M}'_{s'} \models \varphi$.

For more on bisimulation, see e.g. [10], [44], or [12].

3 Contingency bisimulation

A proposition is said to be *contingent* if it can be both true and false, and non-contingent if it is necessarily true or necessarily false (i.e., if it is not contingent). This notion has led to the proposal of modal logics of contingency by Montgomery and Routley [41]. We write $\Delta\varphi$ for 'φ is non-contingent' and $\nabla\varphi$ for 'φ is contingent'. In logics of knowledge, interpreted on structures with equivalence relations, $\Delta\varphi$ stands for 'the agent knows whether φ' and $\nabla\varphi$ for 'the agent is ignorant about φ'. In that setting, contingency logics are known as logics of ignorance [36].

The language and semantics are as follows.

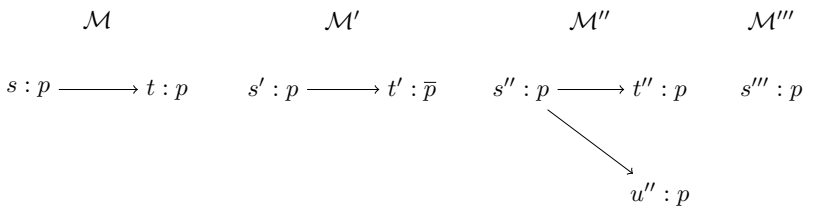
Definition 3.1 Language of contingency logic To the inductive definition of $\mathcal{L}(\Box)$ we add an inductive clause $\Delta_a\varphi$ (for any $a \in A$). The resulting language is $\mathcal{L}(\Box, \Delta)$. The language without $\Box_a$ modalities is $\mathcal{L}(\Delta)$. Contingency $\nabla_a\varphi$ is defined by abbreviation as $\neg\Delta_a\varphi$.

In this case the language $\mathcal{L}(\Box)$ is better known as the language (and logic) of *necessity*.

Definition 3.2 Semantics of non-contingency Given a model $\mathcal{M} = \langle S, R, V\rangle$, we define:

$$\mathcal{M}_s \models \Delta_a\varphi \text{ iff for all } t, u \in S \text{ such that } R_a st, R_a su : \mathcal{M}_t \models \varphi \text{ iff } \mathcal{M}_u \models \varphi.$$

For the continuation of this story we focus on the unlabeled, single-agent case. It will be clear that $\Delta\varphi \leftrightarrow \Delta\neg\varphi$, so that we also have that $\nabla\varphi \leftrightarrow \neg\Delta\neg\varphi$: contingency is not merely the negation of non-contingency but also its dual. Non-contingency is definable with necessity as $\Delta\varphi \leftrightarrow \Box\varphi \vee \Box\neg\varphi$. But necessity cannot always be defined with contingency. In [41] is it proposed to define $\Box\varphi$ as $\Delta\varphi \wedge \varphi$. However, this definition is only available in the systems containing the **T** axiom $\Box\varphi \to \varphi$ [42, page 128]. As this includes models with equivalence relations, the logics of ignorance and the logic of knowledge are interdefinable, and thus, obvious, equally expressive. No novel notion of bisimulation is needed here. But for weaker logics, lacking the **T** axiom, this is no longer the case. We can always embed $\mathcal{L}(\Delta)$ into $\mathcal{L}(\Box)$ employing $\Delta\varphi \leftrightarrow \Box\varphi \vee \Box\neg\varphi$, so necessity logic is at least as expressive as contingency logic. It is even more expressive. This we can easily demonstrate with the example from the introduction, that we recall here once more—and let us add two more also modally equivalent models $\mathcal{M}''$ and $\mathcal{M}'''$ for good measure.

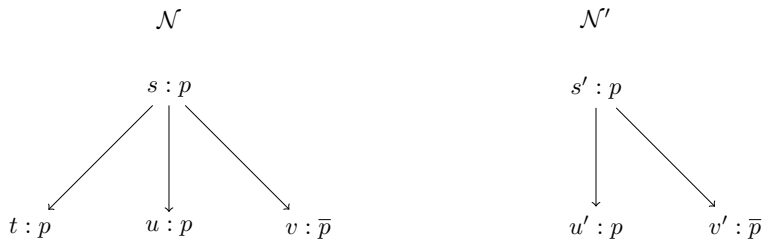


The pointed models $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ are modally equivalent in $\mathcal{L}(\Delta)$ (a state with at most one successor satisfies $\Delta\varphi$ for any φ). But they are modally different in $\mathcal{L}(\Box)$, for example $\mathcal{M}_s \models \Box p$ and $\mathcal{M}'_{s'} \not\models \Box p$. And indeed, as already mentioned, they are also not standard bisimilar. We want another notion of bisimilarity, under which $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ are bisimilar.

Now consider $\mathcal{M}''$. Pointed model $\mathcal{M}''_{s''}$ is standard bisimilar to $\mathcal{M}_s$ and it also satisfies the same $\mathcal{L}(\Delta)$ formulas. However, as s'' has more than one successor, it is less obvious that $\mathcal{M}''_{s''}$ also satisfies all $\Delta\varphi$ formulas. Finally consider $\mathcal{M}'''$. Again we have that $\mathcal{M}'''_{s'''}$ satisfies p and also all $\Delta\varphi$, so this

pointed model must be bisimilar as well to all previous three. We can swap the value of p in t and t' at will, it does not matter. But notice that we cannot swap the value of p in t'' at will: if it were false, then $\mathcal{M}''_{s''} \models \nabla p$ and we lose modal equivalence (and, presumably, bisimilarity) with the other pointed models.

For another example, we want the following models to be bisimilar (in their roots):



Both $\mathcal{N}_s$ and $\mathcal{N}'_{s'}$ satisfy ∇p. These pointed models are also standard bisimilar. If we were to swap the value of p in v or v', we lose modal equivalence.

Contingency bisimulation was proposed in [25]. In contingency bisimulation we strengthen the **forth** and **back** clauses by adding a requirement that there are at least two non-bisimilar accessible states. As those are in one of the given models, it is therefore defined as an autobisimulation.

Definition 3.3 Contingency Bisimulation Let single-agent model $\mathcal{M} = (S, R, V)$ be given. A *contingency bisimulation* is a non-empty relation Z such that for all $s, s' \in S$ with Zss':

atoms	for all $p \in P$, $s \in V(p)$ iff $s' \in V(p)$;
forth	if there are $u, v \in S$ such that Zuv does not hold, and if $t \in S$ and Rst, then there is a $t' \in S'$ such that $Rs't'$ and Ztt';
back	if there are $u, v \in S$ such that Zuv does not hold, and if $t' \in S'$ and $Rs't'$, then there is a $t \in S$ such that Rst and Ztt'.

A contingency bisimulation *between* $\mathcal{M}$ and $\mathcal{M}'$ is then a contingency (auto)bisimulation $Z \subseteq S \times S'$ on their direct sum $\mathcal{M} + \mathcal{M}'$, i.e. with domain in S and co-domain in S'.

According to this definition, all four models $\mathcal{M}$, $\mathcal{M}'$, $\mathcal{M}''$, and $\mathcal{M}'''$ are contingency bisimilar. A bisimulation Z establishing this, consists of the reflexive, symmetric and transitive closure of the set of pairs connecting their roots: $\{(s, s'), (s', s''), (s'', s''')\}$. We also have that $\mathcal{N}_s \leftrightarrows \mathcal{N}'_{s'}$, which is established by $Z = \{(s, s'), (t, u'), (u, u'), (v, v')\}$. This is more work than for the preceding case, because there are non-bisimilar accessible states from s and from s': the valuation of p in t and u, and in v, is different (and similarly for u' and v').

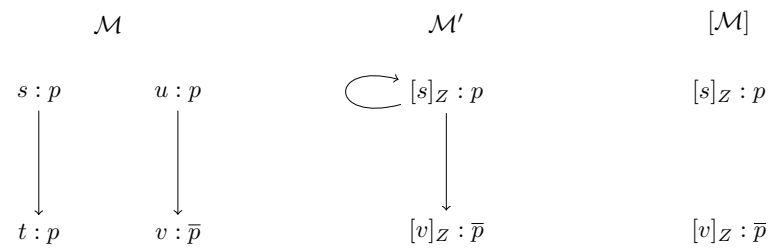
Proposition 3.4 ***[25], Prop. 3.4*** *A standard bisimulation is a contingency bisimulation.*

This will be clear, as the clauses for **forth** and **back** have been weakened in the latter case (the conditional part in the **forth** and **back** clauses is stronger). Not every contingency bisimulation (between two models) is a standard bisimulation, as the examples demonstrated.

We also have the obvious suspect

Proposition 3.5 ***[25], Prop. 3.9*** *Contingency bisimilarity implies modal equivalence, and on image-finite models, modal equivalence implies contingency bisimilarity.* [2]

A standard bisimulation contraction consists of a domain of equivalence classes of the maximal bisimulation Z, where the valuation of propositional variables on a Z equivalence class is that of any state in that Z equivalence class, and two Z equivalence classes are in the accessibility relation iff they contain states that are in the accessibility relation. Interestingly, this procedure does not work for contingency bisimulation! Consider the models $\mathcal{M}$ and $\mathcal{M}'$ below: $\mathcal{M}'$ is computed by this procedure from $\mathcal{M}$ ($[s]_Z = \{s, t, u\}$, and $[v]_Z = \{v\}$). But $\mathcal{M}_s$ and $\mathcal{M}'_{[s]}$ are not contingency bisimilar: $\mathcal{M}_s \models \Delta p$ whereas $\mathcal{M}'_{[s]_Z} \not\models \Delta p$.



In a contingency bisimulation contraction the accessibility relation is more constrained than in a standard bisimulation contraction (but the domain and the valuation are defined in the same way).

Definition 3.6 Contingency Bisimulation Contraction Given is a model $\mathcal{M} = (S, R, V)$. Let Z be the maximal contingency bisimulation on $\mathcal{M}$. The *contingency bisimulation contraction* of $\mathcal{M}$ is the quotient structure $[\mathcal{M}] = ([S], [R], [V])$ defined as

- $[S] = \{[s]_Z \mid s \in S\}$ where $[s]_Z = \{t \in S \mid Zst\}$;
- $[R][s][t]$ iff there are $s' \in [s]_Z$, $t' \in [t]_Z$ such that $Rs't'$, and there are u, v such that $Rs'u$, $Rs'v$, and not Zuv;

[2] The constraint in [25, Prop. 3.9] is modal saturation, not image finiteness. That result is even stronger.

- $[V](p) = \{[s]_Z \mid s \in V(p)\}$.

Proposition 3.7 ***[25], Prop. 3.13*** *The contingency bisimulation contraction of a pointed model is contingency bisimilar to that model.*

With this definition of bisimulation contraction, we have that the examples $\mathcal{M}_s$ and $[\mathcal{M}]_{[s]_Z}$ above are contingency bisimilar. The accessibility relation in $[\mathcal{M}]_{[s]_Z}$ is empty, but the contingency bisimulation is non-empty: it contains $(s, [s]_Z)$.

The bisimulation contraction of an $S5$ model need not be an $S5$ model. For example, take a singleton model where p is true with reflexive access. We then lose the arrow in the contraction. This can be corrected by always taking the reflexivity closure of the relation $[R]$ constructed in the second clause of Def. 3.6 (add 'or $s = t$' on the right-hand side).

*The results reported in this section are based on joint work by Hans van Ditmarsch, Jie Fan, and Yanjing Wang reported in [25] and [26]. A crucial role in these works plays the axiom of 'almost definability' (***AD***) which is*

$$\nabla\psi \to (\Box\varphi \leftrightarrow \Delta\varphi \wedge \Delta(\psi \to \varphi))$$

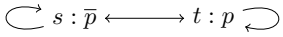
It is a validity of the logic with language $\mathcal{L}(\Box, \Delta)$. Axiom **AD** *states the precise condition under which, even in frames lacking the* **T** *axiom ($\Box\varphi \to \varphi$), necessity is definable from contingency. Axiom* **AD** *says that necessity is* almost definable *by contingency, namely when there is at least one contingent proposition ψ. Many results in [25,26] use the axiom* **AD**. *Although we restricted our discussion to the single-agent case, it seems that the bisimulation definition and results equally apply to the multi-agent case. A striking and truly multi-agent result is the axiomatization of multi-agent contingency logic on symmetric frames reported in [26]. Jie Fan is expected to defend his PhD thesis in 2015.*

4 Awareness bisimulation

Logics for knowledge and awareness are a way to model bounded rationality of agents. One approach is that agents are only aware of a subset of the set of all propositional variables. They are aware of all formulas that only contain those variables. Then, we can define that an agent knows that a formula is true in a given state, iff the agent is aware of the formula and the formula is true in all accessible states. In those accessible states the agent may in principle be aware of other propositional variables, and other agents may have other levels of awareness.

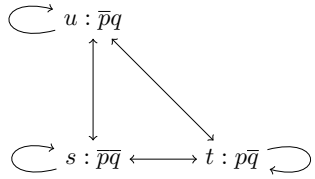
Consider this example. Hans arrives at a conference, and, waking up in the morning and rubbing stardust from his sleepy eyes, Hans realizes what is lacking: coffee. He starts to wonder if coffee would already be served in the restaurant below. It is still fairly early. Let proposition p stand for 'coffee is served' that he is uncertain about. This situation can be visualized as:

$\mathcal{M}$



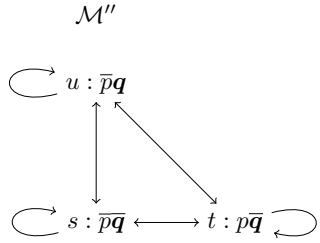
We can observe the usual things about this structure such as that coffee is actually not served but that Hans is uncertain whether coffee is served: $\mathcal{M}_s \models \neg p \wedge \neg(\Box p \vee \Box \neg p)$. Now, while on his way to the lower floor, where the restaurant is located, someone in the elevator mentions that you can't have both coffee and orange juice for breakfast. This makes Hans aware that orange juice is an issue. After this, Hans does not know whether coffee is served and also does not know whether orange juice is served. But he knows that coffee and orange juice are not both served. We now get to the following situation. Unfortunately, Hans has still not found out that the breakfast area is closed. Actually, there is no coffee and there are no oranges (i.e., neither coffee nor orange juice is served): $\mathcal{M}'_s \models \neg p \wedge \neg q \wedge \Box\neg(p \wedge q) \wedge \neg(\Box p \vee \Box \neg p)$.

$\mathcal{M}'$



What is the relation between $\mathcal{M}$ and $\mathcal{M}'$? One way to model this, is to see the former as an abstraction up to the initial level of awareness of Hans of the latter. In a picture, the model $\mathcal{M}'$ is 'really' the following structure, where the value of the variable q of which Hans is unaware is in bold font:

$\mathcal{M}''$



The action of 'Hans becomes aware of orange juice' then transforms structure $\mathcal{M}''$ into structure $\mathcal{M}'$.

Now it is common to distinguish explicit knowledge from implicit knowledge in such scenarios, where the knowledge we really want to talk about is

explicit knowledge: the thing that is true in all accessible worlds of which we are aware. Implicit knowledge is then modal accessibility, and in fact rather a technical non-intuitive notion. Let us now write $A\varphi$ for 'Hans is aware of φ' (with the suitable agent-labelled perspective $A_a\varphi$, as below), keep $\Box$ for modal accessibility, and write K^E for explicit knowledge as $K^E\varphi$ iff $\Box\varphi \wedge A\varphi$. Then we can say that Hans is explicitly uncertain about coffee

$$\mathcal{M}'', s \models \neg(K^E p \vee K^E \neg p)$$

whereas he is unaware of orange juice in that state

$$\mathcal{M}'', s \models \neg Aq$$

and that he implicitly knows that coffee and orange juice are not both served ($\Box\neg(p\wedge q)$), although in view of the example it seems to make little sense to call that implicit knowledge. It is rather something he may find out in the future of waking up even more. We can also model actions such as become aware of q, where this action transforms $\mathcal{M}''$ into $\mathcal{M}'$, but this would be beyond the scope of this survey.

In this example, Hans has the same level of awareness in all states of the model. Either he is aware of p everywhere, or he is aware of q everywhere. Because of this, the principle $A\varphi \rightarrow K^E A\varphi$ is satisfied: he knows what he is aware of. Of course he does not know what he is unaware of: then the epistemic operator binds a formula containing an unaware variable, so that the formula is false. The principle that the level of awareness is the same in all states that the agent considers possible is called *awareness introspection* (it is also contested, by economists). Even with the restriction of awareness introspection, many meaningful multi-agent scenarios involving awareness can be enacted. For example, Tim, who is aware of q, may be uncertain if Hans is aware of q or not. Tim can thus explicitly reason about Hans' awareness and knowledge.

Without awareness introspection, and without equivalence relations for agents such as in our example, very simple scenarios already illustrate the need for a different notion of bisimulation. Consider this:

$\mathcal{N}$ $\qquad\qquad\qquad\qquad$ $\mathcal{N}'$

$s : p \longrightarrow t : \boldsymbol{p} \longrightarrow u : \boldsymbol{p}$ $\qquad\qquad$ $s' : p \longrightarrow t' : \boldsymbol{p} \longrightarrow u' : \overline{\boldsymbol{p}}$

In state s, wherein the agent is aware of p, the agent considers state t possible, wherein the agent is unaware of p. In state t, as the agent is unaware of p, the agent considers u possible wherein p is true. But, as the agent is unaware of p, the value of p in u does not matter: it is below the level of visibility of the agent! Therefore, we wish to identify $\mathcal{N}_t$ and $\mathcal{N}_{t'}$ from the perspective of the agent. And therefore, we would also wish to identify $\mathcal{N}_s$ and $\mathcal{N}'_{s'}$. Differently said, we cannot distinguish $\mathcal{N}_s$ from $\mathcal{N}'_{s'}$ in the logic with as its only modality K^E.

As the pointed models $\mathcal{N}_s$ and $\mathcal{N}'_{s'}$ are obviously not bisimilar in the standard sense, we are therefore looking for a different notion of bisimilarity. There is a relation with dynamics of awareness: after becoming aware of p, the agent can after all distinguish between the two structures, for example we then have that $K^E K^E p$ is true in the former and false in the latter. So dynamics should also play a role in the expressivity of such logics.

We now continue to present the formal setup. As said, the *only* structural difference is that in every state (and for every agent) there are two types of propositional variables: aware variables and unaware variables. All the rest follows from that.

We augment standard epistemic (Kripke) models with a parameter for awareness, and subsequently introduce a proper notion of bisimilarity for these structures.

Definition 4.1 Awareness model [24] An *awareness model* for A and P is a tuple $M = (S, R, \mathcal{A}, V)$ where S, R, and V are as before and where $\mathcal{A}$ is an *awareness function* $\mathcal{A} : A \to S \to \mathcal{P}(P)$. For $\mathcal{A}(a)$ we write $\mathcal{A}_a$.

The property of *awareness introspection* [35] holds if the agents know when they are aware of a proposition: if $R_a st$, then $\mathcal{A}_a(s) = \mathcal{A}_a(t)$.

The required structural similarity is captured in the following notion, named *awareness bisimulation*. Informally, given a model and a set $Q \subseteq P$, another model is a Q awareness bisimulation if it cannot be distinguished from the first by formulas built only from propositional variables in Q, and only in the scope of modalities for agents who are aware of those propositional variables.

Definition 4.2 Awareness bisimulation [19,22] Let awareness models $\mathcal{M} = (S, R, \mathcal{A}, V)$ and $\mathcal{M}' = (S', R', \mathcal{A}', V')$ be given, and let $Q \subseteq P$. A Q *awareness bisimulation* is a function Z from the subsets of Q to the binary relations between $\mathcal{M}$ and $\mathcal{M}'$ (for $Z(Q)$ we write Z_Q), such that for all $Q' \subseteq Q$ in the domain of Z, for all $s \in S, s' \in S'$ such that $Z_{Q'} ss'$, and for all agents $a \in A$:

atoms	for all $p \in Q'$, $s \in V(p)$ iff $s' \in V'(p)$;
aware	$\mathcal{A}_a(s) \cap Q' = \mathcal{A}'_a(s') \cap Q'$;
forth	if $t \in S$ and $R_a st$ then there is a $t' \in S'$ such that $R'_a s't'$ and $Z_{Q' \cap \mathcal{A}_a(s)} tt'$;
back	if $t' \in S'$ and $R'_a s't'$ then there is a $t \in S$ such that $R_a st$ and $Z_{Q' \cap \mathcal{A}'_a(s')} tt'$.

We also call each Z_Q a Q awareness bisimulation. If there is a Q awareness bisimulation linking $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ via $Z_Q ss'$ we write $\mathcal{M}_s \leftrightarrows_Q \mathcal{M}'_{s'}$.

The **aware** clause can be considered as an additional **atoms** requirement, due to the nature of our models where states have more structure than merely propositional truth. If we were to replace $Z_{Q' \cap \mathcal{A}_a(s)}$ in the **back** and **forth** clauses with $Z_{Q'}$, we get standard (restricted) bisimulation (restricted to Q').

Thus every standard bisimulation is an awareness bisimulation. But the intersection makes all the difference: every time we travel further down a path in the structure, we can only 'see' anything all the agents along that path are aware of, in the states along that path.

If all agents are aware of all propositional variables, the awareness bisimulation is a standard bisimulation. This is what we desire: we then revert to the standard multi-agent epistemic situation, where awareness plays no role.

Awareness bisimulation is clearly more complex than standard bisimulation, however its motivation is very simple. Two states are Q awareness bisimilar if, for any observer aware only of the propositional variables in Q, the states appear identical. It gives us the "Q perspective" of an awareness model.

For an example illustrating the mechanics of the definition, consider awareness models $\mathcal{N}_s$ and $\mathcal{N}'_{s'}$ above. We show that they are $\{p\}$ awareness bisimilar. The single agent is anonymous (i.e., unlabeled relations). Constructively following the steps in the definition of awareness bisimulation, we can see this as follows:

- $\mathcal{M}_u$ and $\mathcal{M}'_{u'}$ are $\emptyset$ awareness bisimilar, because all four clauses of awareness bisimulation are trivially satisfied;
- $\mathcal{M}_t$ and $\mathcal{M}'_{t'}$ are $\{p\}$ awareness bisimilar, because t and t' coincide in p's truth value (namely, it is true) and in the agent's awareness of p (namely $\emptyset$), and because (**forth**) $\{p\} \cap \mathcal{A}_a(t) = \emptyset$ and $\mathcal{M}_u$ and $\mathcal{M}'_{u'}$ are $\emptyset$ awareness bisimilar; similarly for **back**;
- $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ are $\{p\}$ awareness bisimilar, because s and s' coincide in p's truth value (namely, it is true) and in a's awareness of p (namely $\{p\}$), and because (**forth**) $\{p\} \cap \mathcal{A}_a(s) = \{p\}$ and epistemic awareness states $\mathcal{M}_t$ and $\mathcal{M}'_{t'}$ are $\{p\}$ awareness bisimilar; similar for **back**.

Alternatively to the construction above, we could have observed that the set $\{Z_p, Z_\emptyset\}$ satisfies the clauses of awareness bisimulation, where (write Z_p for $Z_{\{p\}}$).

$$\begin{array}{rcl} Z_p & = & \{(s,s'),(t,t')\} \\ Z_\emptyset & = & \{(u,u')\} \end{array}$$

As usual, there are many bisimulations given two pointed models. Yet another awareness bisimulation is the following set $\{Z'_p, Z'_\emptyset\}$. It is maximal. It satisfies that $Z'_\emptyset \subseteq Z'_p$. (If $Q' \subseteq Q$, then, on the assumption that an awareness bisimulation exists, one can always find a $Z_{Q'}$ such that $Z_Q \subseteq Z_{Q'}$.)

$$\begin{array}{rcl} Z'_p & = & \{(s,s'),(t,t')\} \\ Z'_\emptyset & = & \{(s,s'),(t,t'),(u,u')\} \end{array}$$

Definition 4.3 Language The language $\mathcal{L}(\Box, K^E, K^S, A)$ is defined as

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_a\varphi \mid K^E_a\varphi \mid K^S_a\varphi \mid A_a\varphi$$

where $p \in P$ and $a \in A$.

We also use sublanguages of the language $\mathcal{L}(\Box, K^E, K^S, A)$ with fewer inductive constructs. The set $v(\varphi)$ of propositional variables of a formula φ in the logical language is defined in the obvious way.

Definition 4.4 Semantics Let awareness model $\mathcal{M} = (S, R, \mathcal{A}, V)$, $s \in S$, and $\varphi \in \mathcal{L}(\Box, K^E, K^S, A)$.

$\mathcal{M}_s \models A_a\varphi$ iff $v(\varphi) \subseteq \mathcal{A}_a(s)$
$\mathcal{M}_s \models K_a^E\varphi$ iff $\mathcal{M}_t \models \varphi$ for all $t \in S$ such that $R_a st$ and $v(\varphi) \subseteq \mathcal{A}_a(s)$
$\mathcal{M}_s \models K_a^S\varphi$ iff $\mathcal{M}'_{t'} \models \varphi$ for all $t \in S$ such that $R_a st$ and
for all $\mathcal{M}'_{t'}$ s.t. $\mathcal{M}_t \leftrightarrows_{\mathcal{A}_a(s)} \mathcal{M}'_{t'}$

The *logic of implicit knowledge* is the one with language $\mathcal{L}(\Box, A)$, the *logic of explicit knowledge* is the one with language $\mathcal{L}(K^E, A)$, and the *logic of speculative knowledge* is the one with language $\mathcal{L}(K^S, A)$. Speculative knowledge is the ugly duckling in this pond (she still has to grow up into a beautiful swan—or rather, lest I insult my collaborators, she already is, but not yet noticed by a crowd). In this survey we will of course neither motivate nor illustrate this concept in detail. A formula is speculatively known to an agent, if in all accessible states, in all awareness bisimilar states from the perspective of this agent, it is true. This modality has aspects of a bisimulation quantifier [37,28]. The logic of explicit knowledge has some nasty charasteristics because of the syntactic way in which the semantics of the awareness modality is given. For example we do not have necessitation. But the logic of speculative knowledge satisfies necessitation. One can speculate over variables of which one is aware, and thus observe that even for unaware variables, $p \vee \neg p$ will always be true. Formula $p \vee \neg p$ is a validity, and $K_a^S(p \vee \neg p)$ is a validity in the logic of speculative knowledge, but $K_a^E(p \vee \neg p)$ is invalid in the logic of explicit knowledge (namely, it is false if the agent is unaware of p).

On the single-agent example we have already seen, reprinted here, we indeed have that $\mathcal{N}_s \models K^E p$, but that $\mathcal{N}_s \not\models K^E K^E p$, because $\mathcal{N}_t \not\models K^E p$, which fails because $\mathcal{N}_t \not\models Ap$.

$\mathcal{N}$ $\qquad\qquad$ $\mathcal{N}'$

$s : p \longrightarrow t : \boldsymbol{p} \longrightarrow u : \boldsymbol{p}$ $\qquad\qquad$ $s' : p \longrightarrow t' : \boldsymbol{p} \longrightarrow u' : \overline{\boldsymbol{p}}$

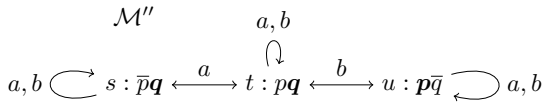
Now consider what were to happen if the agent became aware of p in the other two states as well. It is easy to model this with a dynamic modality, but we will refrain from doing so. It is sufficient to say that the result is

$\mathcal{N}^{+p}$ $\qquad\qquad$ $\mathcal{N}'^{+p}$

$s : p \longrightarrow t : p \longrightarrow u : p$ $\qquad\qquad$ $s' : p \longrightarrow t' : p \longrightarrow u' : \overline{p}$

Clearly the models $\mathcal{N}_s^{+p}$ and $\mathcal{N}'^{+p}_{s'}$ are now not bisimilar, and $K^E K^E p$ is a distinguishing formula. In the logic of explicit knowledge *with dynamics* (interpreted as model transforming operations) we can therefore distinguish $\mathcal{N}_s$ from $\mathcal{N}'_{s'}$. (The distinguishing formula is then $\langle +p \rangle K^E K^E p$, where $\langle +p \rangle$ is the diamond-form of a modality representing 'becoming aware of p', interpreted by adding p to the set $\mathcal{A}_a(s)$ for all states and all agents.)

For another example, for two agents, also illustrating the modalities, consider Hans (a) again, and Tim (b), and the availability of coffee (p) and orange juice (q). There are three states s, t, u. In states s and t, Hans is aware of p and unaware of q, whereas in state u he is unaware of p and aware of q. We let Tim be aware of p and q in all states. The valuations and accessibility relations are as indicated in the figure. The relations are equivalence relations, so it would suffice to indicate partitions only, but for good measure we have drawn all arrows of the accessibility relation again. Now in principle we cannot as before simply put the unaware variables in boldfont, as we have to indicate for each of the two agents of which variables that agent is aware. But in practice we can and we do: we let it indicate what Hans is unaware of, given that Tim is aware of both variables in all states.



This model depicts a scenario like the following—let us assume that t is the actual state. Hans is (as before) uncertain about the availability of coffee but is unaware of orange juice, Tim is aware of both and knows that there is coffee but is, instead, uncertain about the availability of orange juice. Also, he cannot distinguish a state wherein Hans is only aware of p from a state u wherein Hans is only aware of q. Frustratingly, he knows that Hans can resolve his uncertainty about q but unfortunately Hans is unaware of that! Some typical statements to evaluate are

$\mathcal{M}'', t \models \neg(K_a^E p \vee K_a^E q)$	Hans is ignorant about p ...
$\mathcal{M}'', t \models K_a^E(K_b^E p \vee K_b^E q)$	... but he knows that Tim knows whether p
$\mathcal{M}'', u \models K_a^E \neg q$	In state u Hans knows that $\neg q$ ...
$\mathcal{M}'', t \models K_b^E K_a^E \neg q$	... and in t Tim knows that Hans knows that ...
$\mathcal{M}'', t \not\models K_a^E K_b^E K_a^E \neg q$	... but Hans doesn't! He is unaware of q; ...
$\mathcal{M}'', t \models \Box_a K_b^E K_a^E \neg q$	... although he knows it 'implicitly'.

This scenario allows for some fabulous follow-up conversations, assuming that questions make the listening agent aware of all the propositional variables occurring in the question. If Hans asks Tim "Do you know if there's coffee?", Tim can truthfully respond "Thanks for asking! Yes, there is. Also, I learnt from your question that there's orange juice as well. You might not yet have thought of that."

We conclude with reporting some theoretical results for these logics.

Proposition 4.5 *Awareness bisimularity $\leftrightarrows_Q$ is an equivalence relation.*

Proposition 4.6 *Given two image-finite pointed awareness models:*

- *Awareness bisimilarity corresponds to modal equiv. in the logic of explicit knowledge;*
- *Awareness bisimilarity corresponds to modal equiv. in the logic of speculative knowledge;*
- *Standard bisimilarity corresponds to modal equivalence in the logic of implicit knowledge.*

Subject to an inductively defined translation also involving φ:

Proposition 4.7 ***[19]*** *Explicit knowledge implies speculative knowledge, and speculative knowledge implies implicit knowledge: $K_a^E\varphi \rightarrow K_a^S\varphi$ and $K_a^S\varphi \rightarrow \Box_a\varphi$.*

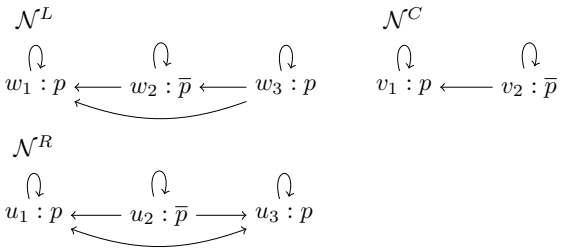
However, on the deeper level of expressivity, we regain correspondence between explicit knowledge and speculative knowledge:

Proposition 4.8 ***[22]*** *The logic of explicit knowledge and the logic of speculative knowledge are equally expressive. The logic of implicit knowledge is (strictly) more expressive than the logic of explicit knowledge and the logic of speculative knowledge.*

The logics for awareness of propositional variables, including explicit knowledge and implicit knowledge, and their axiomatizations, were proposed in [24,34]. Motivated by that and by the complete lattice of space in [35], varying combinations of authors involving Hans van Ditmarsch, Tim French, Fernando Velázquez Quesada, and Yi N. Wang developed the framework reported in [17,19,21,22]. Although this collaboration did not actually involve PhD work, it is relevant to mention that Fernando obtained his PhD degree in 2011, just before his involvement in this collaborative venture, and Yi obtained his (2nd) PhD degree in 2013, just after starting his involvement in this collaborative venture.

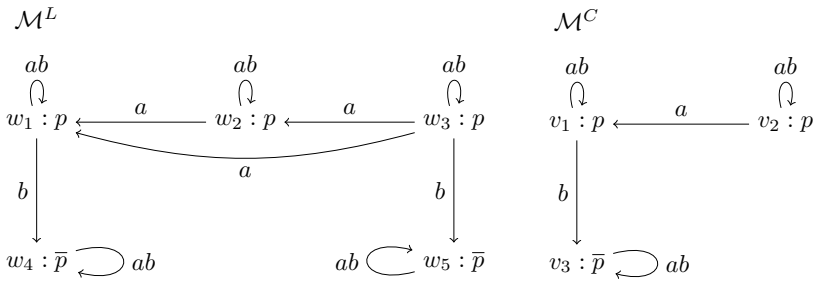
5 Plausibility bisimulation

In structures with so-called *plausibility relations* an agent knows something if it is true in all possible states and an agent believes something if it is true in the most plausible from these possible states. These structures consist of equivalence classes encoding knowledge, where in each equivalence class the states are ordered into more and less plausible states. If s is at least as plausible as t, we write $t \geq s$. Consider the following models $\mathcal{N}^L$, $\mathcal{N}^C$, and $\mathcal{N}^R$. An arrow from s to t in the figure means that $s \geq t$. (Every state is at least as plausible as itself.)



In model $\mathcal{N}^L$ we have that $w_3 > w_2 > w_1$: the agent finds it most plausible that p is true, less plausible that p is false, and even less plausible that p is true. As p is true in the most plausible state, the agent *believes* p. If we go to slightly less plausible, the agent is already uncertain about the value of p, it only *knows* trivialities such as $p \vee \neg p$. The state w_3 does not make the agent even more uncertain. We can discard it. This is the model $\mathcal{N}^C$. Another trick also works: Keep state w_3, but make it as plausible state as w_1. Then we get, modulo renaming of states, the plausibility model $\mathcal{N}^R$ in the figure above, wherein u_1 and u_3 are equally plausible.

If the arrow were to correspond directly to a modality, the models $\mathcal{N}^C$ and $\mathcal{N}^R$ would be bisimilar, but $\mathcal{N}^L$ would not be bisimilar to the other two. But in the logic of knowledge and belief interpreted on such plausibility structures, they are all three *plausibility bisimilar*. In the single-agent case, the way to achieve that is easy: identify states with the same valuation. But in the multi-agent case, this is not so easy. Let us consider such an example of multi-agent bisimilarity.



Also here, we'd like to say that the two pictured models are *plausibility bisimilar*. Clearly, in this case 'bisimilar' does not mean 'having the same valuation'. The a relations in the model $\mathcal{M}^L$ correspond to the plausibility order $w_3 >_a w_2 >_a w_1$ on the a equivalence class $\{w_1, w_2, w_3\}$, such that w_1 should be the most plausible of the three, and the singleton plausibility order on a equivalence classes $\{w_4\}$ and $\{w_5\}$. For agent b, in b class $\{w_1, w_4\}$ the more plausible world is w_4, etc. It is a model validity that a believes that b believes $\neg p$. What a believes is what is true in the most plausible worlds. From

$\{w_1, w_2, w_3\}$ this is w_1, and indeed, agent b there believes (incorrectly) that p is false. But in the classes $\{w_4\}$ and $\{w_5\}$ it is also true that b believes $\neg p$. We can repeat his exercise in $\mathcal{M}^C$ and in fact for any other formula. We would like to say that in $\mathcal{M}^L$ the states w_1 and w_3 are plausibility bisimilar, and the (for example) pointed models $\mathcal{M}^L_{w_3}$ and $\mathcal{M}^C_{v_1}$ are modally equivalent and plausibility bisimilar. However, we now leave the multi-agent case alone and in this survey we focus on the single-agent case only.

We now continue by defining the structures, language, and semantics. The only structural difference is that instead of an arbitrary relation R we now require this to be composed of so-called *well-preorders* $\geq$.[3] Because of that, we use infix and not postfix notation: we write $s \geq t$ instead of Rst.

Definition 5.1 Plausibility model A *plausibility model* is a model $\mathcal{M} = (S, \geq, V)$ such that $\geq$ is a set of mutually disjoint well-preorders covering S, called the *plausibility relation*.

If $s \geq t$ then t is *at least as plausible* as s, and the $\geq$-minimal elements are the *most plausible* worlds. For the symmetric closure of $\geq$ we write $\sim$: this is an equivalence relation on S called the *epistemic relation*. If $s \geq t$ but $t \not\geq s$ we write $s > t$ (t is *more plausible* than s).

We now proceed to define *plausibility bisimulation*. In order for an elegant definition to emerge, we allow ourselves some a further notational abbreviation. Let $X \geq Y$ stand for 'for all $x \in X$ and for all $y \in Y$, $x \geq y$'. We now write $x \succeq y$ for $Min_{\geq}\{z \mid V(z) = V(x)\} \geq Min_{\geq}\{w \mid V(w) = V(y)\}$.

Definition 5.2 Plausibility Bisimulation [4] Let $\mathcal{M} = (S, \geq, V)$ and $\mathcal{M}' = (S', \geq', V')$ be plausibility models. A *bisimulation* between $\mathcal{M}$ and $\mathcal{M}'$ is a non-empty relation $Z \subseteq S \times S'$ such that for all Zss':

atoms	for all $p \in P$, $s \in V(p)$ iff $s' \in V(p)$;
forth$_{\geq}$	if $t \in S$ and $s \succeq t$, there is a $t' \in S$ such that $s' \succeq t'$ and Ztt';
back$_{\geq}$	if $t' \in S$ and $s' \succeq t'$, there is a $t \in S$ such that $s \succeq t$ and Ztt';
forth$_{\leq}$	if $t \in S$ and $s \preceq t$, there is a $t' \in S$ such that $s' \preceq t'$ and Ztt';
back$_{\leq}$	if $t' \in S$ and $s' \preceq t'$, there is a $t' \in S$ such that $s \preceq t$ and Ztt'.

This bisimulation relation is non-standard in the **back** and **forth** clauses. That there are two of each of them is not so special. This is as in temporal logics wherein we also have to look forward and backward along the accessibility relation. The special aspect is that we use $\succeq$ instead of $\geq$ and $\preceq$ instead of $\leq$. This means that, instead of comparing s to t (i.e., instead of merely requiring $s \geq t$), we compare *the set of objects bisimilar to* s (namely the states in a given $\sim$-class that have the same valuation) to *the set of objects bisimilar to* t. The relation $\succeq$ is called the *normal plausibility relation*.

[3] A *well-preorder* is a reflexive and transitive binary relation $\geq$ such that every non-empty subset has $\geq$-minimal elements; where the set of *minimal elements* of some subset Y is $Min_{\geq} Y = \{y \in Y \mid y' \geq y \text{ for all } y' \in Y\}$. As this also holds for a two-element set $Y = \{y, z\}$, this entails that $z \geq y$ or $y \geq z$: all elements in the domain are comparable.

Consider again the models $\mathcal{N}^L$, $\mathcal{N}^C$, and $\mathcal{N}^R$. The maximal bisimulation on $\mathcal{N}^L$ is $Z = \{(w_1, w_1), (w_1, w_3), (w_3, w_1), (w_3, w_3), (w_2, w_2)\}$. Therefore, although $w_2 < w_3$, we have that $w_2 \succ w_3$: although it appeared that w_2 is more plausible than w_3, in reality w_2 is less plausible than w_3. If we replace $\geq$ by the normal plausibility relation $\succeq$ in $\mathcal{N}^L$, we get the model $\mathcal{N}^R$. The bisimulation contraction is the model $\mathcal{N}^C$.

We now continue with the language and semantics.

Definition 5.3 Logical language Language $\mathcal{L}(K, C, D, \Box)$ is inductively defined by:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid B^{\varphi}\varphi \mid B^n\varphi \mid \Box\varphi$$

where $p \in P$, and $n \in \mathbb{N}$.

As usual we also consider sublanguages. The formula $K\varphi$ stands for 'the agent knows φ' (in this section $\Box\varphi$ does not mean that the agent knows φ), $B^{\psi}\varphi$ stands for *'the agent believes φ on condition ψ'*, $B^n\varphi$ stands for *'the agent believes φ to degree n'*, and $\Box\varphi$ stands in this case for *'the agent safely believes φ'*. The logic with language $\mathcal{L}(K, C)$ is the *logic of conditional belief* (with C for 'conditional'), the logic with language $\mathcal{L}(K, D)$ is the *logic of degrees of belief* (with D for 'degrees of'), and $\mathcal{L}(K, \Box)$ defines the *logic of safe belief*.

Definition 5.4 Semantics Let now $\succeq$ be the normal plausibility relation given a plausibility model $\mathcal{M} = (S, \geq, V)$, then we define:

$$\begin{array}{ll}
\mathcal{M}_s \models K\varphi & \text{iff } \mathcal{M}_t \models \varphi \text{ for all } t \in S \text{ such that } s \sim t \\
\mathcal{M}_s \models B^{\psi}\varphi & \text{iff } \mathcal{M}_t \models \varphi \text{ for all } t \in \mathit{Min}(\llbracket\psi\rrbracket_{\mathcal{M}} \cap [s]_{\sim}) \\
\mathcal{M}_s \models B^n\varphi & \text{iff } \mathcal{M}_t \models \varphi \text{ for all } t \in \mathit{Min}^n_{\succeq}[s]_{\sim} \\
\mathcal{M}_s \models \Box\varphi & \text{iff } \mathcal{M}_t \models \varphi \text{ for all } t \text{ with } s \succeq t
\end{array}$$

where

$$\begin{array}{lll}
\mathit{Min}^0_{\succeq}[s]_{\sim} & = \mathit{Min}_{\succeq}[s]_{\sim} & \\
\mathit{Min}^{n+1}_{\succeq}[s]_{\sim} & = \begin{cases} [s]_{\sim} & \text{if } \mathit{Min}^n_{\succeq}[s]_{\sim} = [s]_{\sim} \\ \mathit{Min}^n_{\succeq}[s]_{\sim} \cup \mathit{Min}_{\succeq}([s]_{\sim} \setminus \mathit{Min}^n_{\succeq}[s]_{\sim}) & \text{otherwise.} \end{cases} &
\end{array}$$

The logics of conditional belief and safe belief go back to [43] and the logic of degrees of degrees goes back to, as far as we know, [39,30]. See [9] for an excellent review of the logics of conditional belief and safe belief interpreted on plausibility models and [16] for the logic of degrees of belief interpreted on plausibility models. Unlike the safe belief and degrees of belief logics in [9] and [16], respectively, the modalities defined above are plausibility bisimulation preserving. An alternative (and equivalent) bisimulation preserving semantics for safe belief is $\mathcal{M}_s \models \Box\varphi$ iff $\mathcal{M}_s \models B^{\psi}\varphi$ for all ψ such that $\mathcal{M}_s \models \psi$. See [4] for details.

For an example of conditional belief, consider again plausibility model $\mathcal{N}^C$. We can observe that $\mathcal{N}^C \models Bp$, $\mathcal{N}^C \models B^p p$, and $\mathcal{N}^C \models B^{\neg p}\neg p$ (any formula

of type $B^{\varphi}\psi$ that is true in any point of a model, is always a model validity in this logic). Also we have that $\mathcal{N}^C \models B^{Bp}p$ (as condition Bp on this model is satisfied in both worlds), and $\mathcal{N}^C \models B^{\neg(p\vee\neg p)}(p \vee \neg p)$, i.e., $\mathcal{N}^C \models K(p \vee \neg p)$: the agent knows (and only knows) trivialities. As all three models $\mathcal{N}^L$, $\mathcal{N}^C$, and $\mathcal{N}^R$ have the same information content, we can repeat the exercise in the other two plausibility models.

Concerning degrees of belief, we have that $\mathcal{N}^C \models B^0 p$ but not $\mathcal{N}^C \models B^1 p$, and that the maximum degree of belief is 1: $\mathcal{N}^C \models K\varphi \leftrightarrow B^1\varphi$ (where the maximum degree n of belief is the smallest $n \in \mathbb{N}$ such that for all $m \geq n$ we have $B^m\varphi \leftrightarrow B^{m+1}\varphi$). This is also true in the other two models, of course.

A typical example of safe belief is that $\mathcal{N}^C_{v_1} \models \Box p$ whereas $\mathcal{N}^C_{v_2} \not\models \Box p$. In world v_1, belief in p is safe because p is true in v_1 and the agent still believes p in the v_1 restriction of the model; whereas in v_2 belief in p is not safe: in the v_2 restriction of the model the agent believes that p is false.

Now consider $\mathcal{N}^L$, and the usual definition of safe belief as persistence in *any* model restriction (see [9,15]): with that semantics, belief in p is unsafe in w_3: $\mathcal{N}^L_{w_3} \not\models \Box p$, because in the model restriction to $\{w_2, w_3\}$ (that includes the true state of affairs w_3, as required), the agent believes $\neg p$ instead. Only after the further restriction to w_3 the agent regains belief in p. But in the above semantics of safe belief, $\mathcal{N}^L_{w_3} \models \Box p$: we first have to make the plausibility relation in the model a *normal* plausibility relation, we then get model $\mathcal{N}^R$, and in $\mathcal{N}^R$, world u_2 is not more but less plausible than u_3.

We finish this section with a bisimulation characterization result for single-agent plausibility bisimulation. Below, a relation is *preimage-finite* if the converse relation is image-finite.

Proposition 5.5 ***[4]*** *Given two image-finite and preimage-finite pointed plausibility models. Then plausibility bisimilarity corresponds to modal equivalence in the logic of conditional belief.*

Plausibility models have been used to great effect for modelling belief revision *in dynamic epistemic logic, by, among many other people, Guillaume Aucher, Alexandru Baltag, Johan van Benthem, Cedric Degrémont, Lorenz Demey, Jan van Eijck, Willem Labuschagne, Olivier Roy, Sonja Smets. This list could just as well be five times as long, and we prefer to refrain from proper references. We skip dynamics here.*

Bisimulation for plausibility models has been investigated by Mikkel Birkegaard Andersen, Thomas Bolander, Lorenz Demey, Hans van Ditmarsch, and Martin Holm Jensen. More properly said, it was initiated by Lorenz Demey in [15] and, building on his results, continued by the other four in [4]. Martin Holm Jensen obtained his PhD degree in 2014 [38] and Mikkel Birkegaard Andersen defended his PhD thesis late in 2014, and it will appear in print early in 2015 [3].

6 Refinement as quantifying over information change

In this section we do not present a different notion of bisimulation, motivated by a logical language and semantics that requires an adjustment of what 'similar' is in order to regain correspondence with logical equivalence. Instead, we present a bisimulation-inspired notion of information change, called *refinement*. It is not supposed to preserve truth, but it is supposed to model information growth modulo uncertainty among agents about the extent of the increase. We can then observe that, after all, this notion preserves some truth, namely that of the *positive formulas*.

For an example, consider the following pointed models for an unlabeled relation. We do not need propositional variables for this part of the story, so we simply let $\bullet$ and $\circ$ stand for the states, where the $\circ$ state is the designated point. We are going to juggle a bit with its arrows. We start with this model $\mathcal{M}$.

$$\circ \longrightarrow \bullet \longrightarrow \bullet \longrightarrow \bullet \qquad \mathcal{M}$$

E.g., $\Diamond\Diamond\Diamond\Box\bot$ is true in the point. From the point of view of the modal language, this structure is (standard) bisimilar to

$$\bullet \longleftarrow \bullet \longleftarrow \bullet \longleftarrow \circ \longrightarrow \bullet \longrightarrow \bullet \longrightarrow \bullet \qquad \mathcal{M}'$$

This one also satisfies $\Diamond\Diamond\Diamond\Box\bot$ and any other modal formula for that matter. A more radical structural transformation would be to consider submodels, such as

$$\circ \longrightarrow \bullet \longrightarrow \bullet \qquad \mathcal{M}''$$

A distinguishing formula between the two is $\Diamond\Diamond\Box\bot$, which is true here and false above. Can we consider other 'submodel-like' transformations that are *neither* bisimilar structures *nor* strict submodels? Yes, we can. Consider

$$\bullet \longleftarrow \circ \longrightarrow \bullet \longrightarrow \bullet \qquad \mathcal{M}'''$$

The pointed model $\mathcal{M}'''_\circ$ is neither a submodel of the initial structure $\mathcal{M}_\circ$, nor is it bisimilar. It satisfies the formula $\Diamond\Box\bot \wedge \Diamond\Diamond\Box\bot$ that is certainly false in any submodel of $\mathcal{M}_\circ$. This structure is called a *refinement* (or 'a refinement of the initial structure'), and the original structure is a *simulation* of the latter. (Such terminology is presented in works like [40,45,2,12]; simulation is a very well-studied notion in theoretical computer science.) If we consider the three requirements **atoms**, **forth**, and **back** of a bisimulation, we can see that **atoms** and **back** are satisfied but not **forth**: for example, consider the last arrow in the length-three path in the original structure $\mathcal{M}$: it has no image in $\mathcal{M}'''$, therefore, **forth** fails. There seems to be still some 'submodel-like' relation with

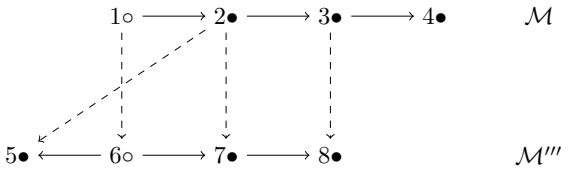
the original structure. Look at its bisimilar duplicate $\mathcal{M}'$. The last structure $\mathcal{M}'''$ is a submodel of that copy. Such a relation always holds: a refinement of a model always is a submodel of a bisimilar copy of that model. This is mirrored on the syntactic side: a *refinement quantifier* is interpreted as refinement, and it can be seen as composed of a *bisimulation quantifier* and *relativization*. We will make this formal at the end of this section.

Definition 6.1 Refinement Let two models $M = (S, R, V)$ and $M' = (S', R', V')$ be given. Consider the definition of bisimulation (Def. 2.2). Let $B \subseteq A$. A relation $Z_B \subseteq S \times S'$ that satisfies **atoms**, that satisfies **back** for all agents $a \in B$, and that satisfies **forth** and **back** for all agents $a \in A \setminus B$ is called a B *refinement* (or B *refinement relation*). If $Z_B ss'$ we say that $M'_{s'}$ *refines* M_s for group of agents B, and we write $M_s \succeq_B M'_{s'}$.

We will overload the meaning of refinement and also say that $M'_{s'}$ is a *refinement* of M_s. An A*-refinement* (of the group of *all* agents) we call a *refinement* (plain and simple) and for $\{a\}$*-refinement* we write a*-refinement*. Dually to B-refinement, we can similarly define B*-simulation*. Note that this definition of simulation then varies slightly from the one in Blackburn *et al.* [12, p.110], where only truth of propositional variables is required to be invariant but not falsity of propositional variables. The dual where only falsity of propositional variables is preserved, is obviously unsuitable for a logic of information change wherein propositional variables do not change their value (refinement modal logic is not a logic of factual change).

An a-refinement needs to satisfy **back** for that agent, but not **forth**. Consider a model and a refinement of that model. Take an arrow in that initial model. This arrow may be missing in the refined model namely when **forth** is not satisfied for that arrow. On the other hand, any arrow in the refinement should be traceable to an arrow in the initial model—the **back** condition, and there may be several arrows in the refinement that are traceable to the same arrow in the initial model. We can see the refined model as a number of copies of the initial model, knitted together, but with bits and pieces cut off so that those copies are no longer exact copies (i.e., possibly no longer bisimilar).

A simple example is as follows. Consider again the model $\mathcal{M}$ and its refinement $\mathcal{M}'''$

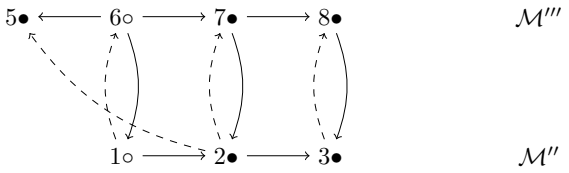


by way of refinement relation $Z = \{(1,6),(2,5),(2,7),(3,8)\}$, also depicted. The arrow $(3,4)$ has no image in the refined model. On the other hand, the arrow $(1,2)$ has two images, namely $(6,5)$ and $(6,7)$. These two arrows cannot

be identified, because 5 and 7 are non-bisimilar, and that is because there is yet another arrow from 7 but no other arrow from 5: arrow $(2,3)$ has only one image in the refined model $\mathcal{M}'''$.

The refinement relation is reflexive and transitive and also confluent (Church-Rosser), and we also have that the composition of refinements for different (groups of) agents is a refinement for their union, so that for any pointed models $\mathcal{M}_s \succeq_a \mathcal{M}'_{s'} \succeq_b \mathcal{M}''_{s''}$ iff $\mathcal{M}_s \succeq_{ab} \mathcal{M}''_{s''}$.

It is maybe curious to observe that two models can be each other's refinement but still not bisimilar. Consider again the introductory example. We have that $\mathcal{M}'''_6 \succeq \mathcal{M}''_1$ and $\mathcal{M}''_1 \succeq \mathcal{M}'''_6$, but still we do not have that $\mathcal{M}'''_6 \leftrightarrows \mathcal{M}''_1$. We obtain $\mathcal{M}'''_6 \succeq \mathcal{M}''_1$ via $\{(6,1),(7,2),(8,3)\}$ and $\mathcal{M}''_1 \succeq \mathcal{M}'''_6$ via $\{(1,6),(2,7),(3,8),(2,5)\}$.



On the other hand, given any $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$, the relation $\mathcal{M}_s \equiv \mathcal{M}'_{s'}$ defined by $\mathcal{M}_s \succeq \mathcal{M}'_{s'}$ and $\mathcal{M}'_{s'} \succeq \mathcal{M}_s$, defines an equivalence. Given that it is not a bisimulation, what is it? It characterizes invariance of the *positive fragment* ($\varphi ::= p \mid \neg p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid K_a\varphi$), or, in other words, that two structures are only different in resolvable differences in uncertainty but not in hard facts and necessary information. This result of course also applies to $\succeq_B$, for the positive formulas involving agents in B. (This observation is used to great effect in a number of ongoing follow-up studies by James Hales.)

Definition 6.2 Language We get the language $\mathcal{L}(\Box,\forall)$ of refinement modal logic by adding an inductive construct $\forall_a\varphi$ to $\mathcal{L}(\Box)$. We write $\exists_a\varphi$ for $\neg\forall_a\neg\varphi$. For $B = \{a_1,\ldots,a_n\}$ we write $\forall_B\varphi$ for $\forall_{a_1}\ldots\forall_{a_n}\varphi$. We write $\forall\varphi$ for $\forall_A\varphi$.

So, formula $\forall\varphi$ does not mean 'for all formulas φ' but it means 'after any refinement, φ (is true)'. Quantifiers usually quantify over variables, as in $\forall x$, $\forall y$, and in the bisimulation quantifiers $\forall p$. A refinement quantifier can be seen as *implicitly* quantifying over a variable, namely over a variable that does not occur in the formula that it binds. We will later present a translation into bisimulation quantified logic that makes the variable explicit.

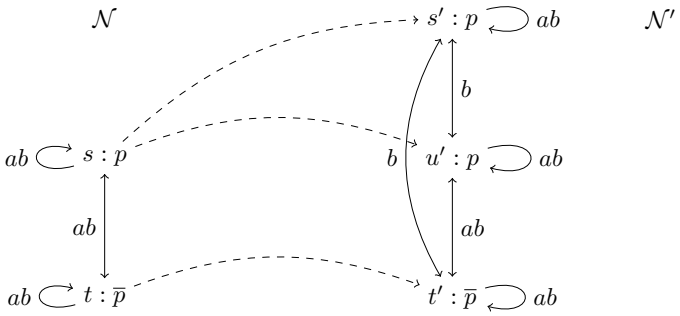
Definition 6.3 Semantics of refinement Assume a model $\mathcal{M} = (S,R,V)$.

$$\mathcal{M}_s \models \forall_a\varphi \quad \text{iff} \quad \text{for all } \mathcal{M}'_{s'} : \mathcal{M}_s \succeq_a \mathcal{M}'_{s'} \text{ implies } \mathcal{M}'_{s'} \models \varphi$$

In other words, $\forall_a \varphi$ is true in a pointed model iff φ is true in all its a-*refinements*. Typical model operations that produce an a-refinement are: blowing up the model (to a bisimilar model) such as adding copies that are indistinguishable from the current model and one another, and removing pairs of the accessibility relation for the agent a (or, alternatively worded: removing states accessible only to agent a).

We now give a multi-agent example, with a dynamic epistemic flavour. Anne and Bill are sitting in a café at a table. A messenger comes in, delivers a letter to Anne, and says to Anne and Bill that it contains the outcome of the election: we model this as the truth about a propositional variable p. In the initial situation Anne and Bill therefore commonly know their ignorance about p. While Bill is away to fetch another drink at the bar, Anne reads the letter. When Bill is back at the table, he suspects but doesn't know that Anne has read the letter (and we assume this is again background knowledge). In dynamic epistemic logic the informational transition is the one transforming model $\mathcal{N}_s$ below into model $\mathcal{N}'_{s'}$ below. For example, in s' Anne knows that p but Bill considers it possible that Anne is still uncertain about p (namely in case she had not opened the letter).

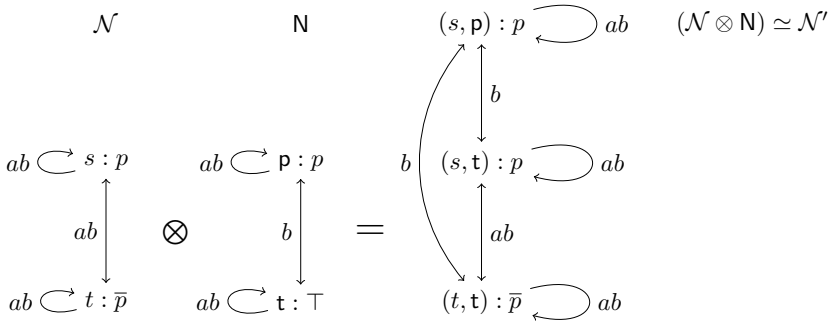
The accessibility relations in the models $\mathcal{N}$ and $\mathcal{N}'$ are all equivalence relations, it concerns knowledge, but for extra clarity we again draw all arrows. The transformation induced by the informative event in which Anne may have read the letter, as above, can be expressed as a refinement. In this case it is an a-refinement.



On the left, the formula $\mathcal{N}_s \models \exists_a(\Box_a p \wedge \neg\Box_b\Box_a p)$ is true, because $\mathcal{N}'_{s'} \models \Box_a p \wedge \neg\Box_b\Box_a p$ is true on the right. On the right, in the actual state there is no alternative for agent a (only the actual state itself is considered possible by a), so $\Box_a p$ is true, whereas agent b also considers another state possible (t' or u'), wherein agent a considers it possible that p is false (namely in state t', accessibly by a from both t' or u'). Therefore, $\Diamond_b\Diamond_a\neg p$, i.e., $\neg\Box_b\Box_a p$, is also true in the actual state s' on the right.

The model on the right is an a-refinement of the model on the left. The refinement relation is pictured by the dashed arrows.

The relation between refinement modal logic and dynamic epistemic logic is quite strong: a refinement of a (finite) pointed model can also be obtained by executing an epistemic action. Therefore, we should be able to see the refinement in this example as produced by an epistemic action. This is indeed the case. Let us not introduce action models formally, for that see e.g. [23], but merely describe the epistemic action of Anne maybe reading a letter. This epistemic action N consists of two action points t and p, they can be distinguished by agent a but not by agent b. What really happens is p; it has precondition p. Agent b cannot distinguish this from t with precondition $\top$.



The execution of this action is depicted above. The point of the structure is action p with precondition p: in fact, a is learning that p, but b is uncertain between that action and the trivial action t wherein nothing is learnt. The trivial action can be executed in both states of the initial model. The action p can only be executed in the state where p is true. Therefore, the resulting structure is the refinement with three states.

Some results for refinement modal logic are the following.

Proposition 6.4 ***[18]*** *Action model execution is a refinement, and, on finite models, every refinement is the execution of an action model.*

Proposition 6.5 ***[20]*** *Refinement modal logic has a complete axiomatization and is equally expressive as multi-agent modal logic.*

The expressivity result is rather surprising, given that arbitrary public announcement logic [8], that quantifies over all public announcements, is undecidable. The gaps filled by refinement modal logic makes the logic decidable again.

Proposition 6.6 ***[13]*** *Refinement is bisimulation plus model restriction, and refinement quantification is bisimulation quantification followed by relativization.*

In a simplifying example: $\forall\varphi$ is equivalent to $\tilde{\forall}p\varphi^p$, where $\tilde{\forall}p$ is a bisimulation quantifier, where p is a fresh atom, and where φ^p is relativization of φ to p. Note

that this is a translation of refinement modal logic into bisimulation-quantified modal logic.

Proposition 6.7 ***[31,33]*** *Refinement epistemic logic has a complete axiomatization.*

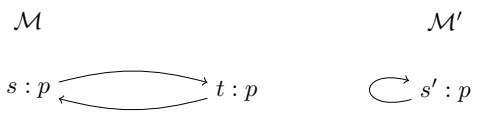
Refinement modal logic is a growing enterprise with contributions by Antonis Achilleos, Laura Bozzelli, Rowan Davies, Hans van Ditmarsch, Tim French, James Hales, Michael Lampis, Sophie Pinchinat, and Edwin Tay [18,20,13,14,31,33,32,29] . Results on complexity and succinctness are found in [13,14,1]. James Hales has results on refinement modal logics for other model classes (such as refinement epistemic logic), and on action model synthesis [31,33,32,29]. James is expected to defend his PhD thesis in 2015.

7 Bisimulation for sabotage

Sabotage logic was proposed by Johan van Benthem in [11], in the very appealing context of a traveller desperately trying to get from A to B using a railway network, while the railway operator purposefully sabotages connections in the network, thus attempting to prevent the traveller from arriving at B. It contains an operator for what is true after one removes a pair (any pair) from the accessibility relation. Let $\mathcal{M} = (S, R, V)$, $s \in S$. We present the 'diamond' version (the dual 'box' version will be obvious).

$$\mathcal{M}_s \models \langle \mathsf{sb} \rangle \varphi \quad \text{iff} \quad \text{there are } t, u \in S \text{ such that } \mathcal{M}_s^{-tu} \models \varphi$$

where $\mathcal{M}^{-tu}$ is as $\mathcal{M} = (S, R, V)$ except that (let us keep this single-agent) $R^{-tu} = R \setminus \{(t, u)\}$. In this logic we can count arrows and this has the catastrophic consequence that the sabotage operation is not bisimulation preserving. A very elementary example provide the following two models.



Observe that $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ are standard bisimilar. However, we have that $\mathcal{M}_s \models \langle \mathsf{sb} \rangle \Diamond \top$ (remove the pair (t, s)) so therefore also $\mathcal{M}_s \not\models [\mathsf{sb}] \Box \bot$; whereas $\mathcal{M}_s \models [\mathsf{sb}] \Box \bot$ (there is only a single arrow to remove, and then the accessibility relation is empty). Therefore, adding the dynamics destroys bisimilarity.

Again, this can be repaired. In this case we have to strengthen the requirements of bisimulation instead of weakening them. And instead of having a bisimulation between $\mathcal{M}_s$ and $\mathcal{M}'_{s'}$ with models $\mathcal{M} = (S, R, V)$ and $\mathcal{M}' = (S', R', V')$, as a relation between states containing the pair (s, s'), we have to define a bisimulation as a *relation between state-(accessibility)relation pairs* containing the pair $((s, R), (s', R'))$. Also, we have to add clauses for the dynamic sabotage modality. Thus, one defines *sabotage bisimilarity*. Then, the above models are not sabotage bisimilar. Sabotage bisimilarity corresponds to

modal equivalence in this language [6,27]. In this section of the survey we refrain from full details of the bisimulation.

The interest of mentioning sabotage logic, is that there is a sliding scale from this logic to logics with epistemic modalities. In [6,7] it is proposed to have various alternatives to the sabotage operators as employed by Van Benthem in [11]: instead of removing any pair from the accessibility relation, one may only be allowed to remove any pair *originating in the actual state*. Instead of removing a pair, one may *swap* a pair (i.e., replace some Rst by Rts), or one may *add* a pair. This then comes with corresponding 'swap' and (as it is called by the authors) 'bridge' modalities.

We can also imagine only being allowed to remove a pair (s, t) that satisfies some logical condition φ at s or some logical condition ψ at t. Or, beyond that, one may instead of removing a single pair, remove all pairs satisfying such logical conditions. Removing all pairs satisfying $\neg\varphi$ at the beginning or $\neg\varphi$ at the end is also known as public announcement of the formula φ (this amounts to preserving all pairs satisfying φ at the beginning and at the end): we are back into epistemic logics. Such transitions are indeed proposed in [27,5]. This seems to open up novel frontiers for epistemic logics.

Generalizations of sabotage logic to other relation-changing modalities, including generalizations to dynamic epistemic logics, have been investigated by Carlos Areces, Hans van Ditmarsch, Raul Fervari, Guillaume Hoffmann, Bastien Maubert, and François Schwarzentruber. This was mainly the PhD work of Raul Fervari. See [6,27,5,7]. Raul Fervari obtained his PhD degree in 2014.

8 Conclusions and further research

We have presented contingency bisimulation, awareness bisimulation, plausibility bisimulation, refinement, and, in lesser detail, bisimulation for sabotage. Such adjustments of standard bisimulation are required to preserve the usual correspondence between bisimulation and modal equivalence. All these variations on bisimulation are inspired by modelling knowledge and belief, and change of knowledge and belief. The work resulted from recent and ongoing PhD research in various locations over the globe, and is therefore expected to strongly develop further in the coming years.

References

[1] Achilleos, A. and M. Lampis, *Closing a gap in the complexity of refinement modal logic* (2013), `http://arxiv.org/abs/1309.5184`.

[2] Aczel, P., "Non-Well-Founded Sets," CSLI Publications, Stanford, CA, 1988, CSLI Lecture Notes 14.

[3] Andersen, M., "Towards Theory-of-Mind agents using Automated Planning and Dynamic Epistemic Logic," Ph.D. thesis, Technical University of Denmark (2015), to appear.

[4] Andersen, M., T. Bolander, H. van Ditmarsch and M. Jensen, *Bisimulation for single-agent plausibility models*, in: S. Cranefield and A. Nayak, editors, *Proc. of 26th Australasian AI*, LNCS 8272 (2013), pp. 277–288.

[5] Areces, C., H. van Ditmarsch, R. Fervari and F. Schwarzentruber, *Logics with copy and remove*, in: *Proc. of 21st WoLLIC* (2014), pp. 51–65, LNCS 8652.
[6] Areces, C., R. Fervari and G. Hoffmann, *Moving arrows and four model checking results*, in: L. Ong and R. de Queiroz, editors, *Proc. of 19th WoLLIC, Buenos Aires* (2012), pp. 142–153, LNCS 7456.
[7] Areces, C., R. Fervari and G. Hoffmann, *Swap logic*, Logic Journal of the IGPL **22(2)** (2014), pp. 309–332.
[8] Balbiani, P., A. Baltag, H. van Ditmarsch, A. Herzig, T. Hoshi and T. D. Lima, *'Knowable' as 'known after an announcement'*, Review of Symbolic Logic **1(3)** (2008), pp. 305–334.
[9] Baltag, A. and S. Smets, *A qualitative theory of dynamic interactive belief revision*, in: *Proc. of 7th LOFT*, Texts in Logic and Games 3 (2008), pp. 13–60.
[10] van Benthem, J., "Modal Logic and Classical Logic," Bibliopolis & Humanities Press, 1985.
[11] van Benthem, J., *An essay on sabotage and obstruction*, in: *Mechanizing Mathematical Reasoning*, LNCS 2605 **2605**, Springer, 2005 pp. 268–276.
[12] Blackburn, P., M. de Rijke and Y. Venema, "Modal Logic," Cambridge University Press, Cambridge, 2001, cambridge Tracts in Theoretical Computer Science 53.
[13] Bozzelli, L., H. van Ditmarsch, T. French, J. Hales and S. Pinchinat, *Refinement modal logic*, Information and Computation **239** (2014), pp. 303–339.
[14] Bozzelli, L., H. van Ditmarsch and S. Pinchinat, *The complexity of one-agent refinement modal logic*, Theoretical Computer Science (2014), to appear (expanded journal version of JELIA and IJCAI publications).
[15] Demey, L., *Some remarks on the model theory of epistemic plausibility models*, Journal of Applied Non-Classical Logics **21(3-4)** (2011), pp. 375–395.
[16] van Ditmarsch, H., *Prolegomena to dynamic logic for belief revision*, Synthese **147** (2005), pp. 229–275.
[17] van Ditmarsch, H. and T. French, *Awareness and forgetting of facts and agents*, in: P. Boldi, G. Vizzari, G. Pasi and R. Baeza-Yates, editors, *Proc. of WI-IAT Workshops 2009* (2009), pp. 478–483.
[18] van Ditmarsch, H. and T. French, *Simulation and information*, in: J. Broersen and J.-J. Meyer, editors, *Knowledge Representation for Agents and Multi-Agent Systems*, LNAI 5605 (2009), pp. 51–65.
[19] van Ditmarsch, H. and T. French, *Becoming aware of propositional variables*, in: M. Banerjee and A. Seth, editors, *Proc. of 4th ICLA*, LNCS 6521 (2011), pp. 204–218.
[20] van Ditmarsch, H., T. French and S. Pinchinat, *Future event logic - axioms and complexity*, in: L. Beklemishev, V. Goranko and V. Shehtman, editors, *Advances in Modal Logic 8* (2010), pp. 77–99.
[21] van Ditmarsch, H., T. French and F. Velázquez-Quesada, *Action models for knowledge and awareness*, in: W. van der Hoek, L. Padgham, V. Conitzer and M. Winikoff, editors, *Proc. of 11th AAMAS*, 2012, pp. 1091–1098.
[22] van Ditmarsch, H., T. French, F. Velázquez-Quesada and Y. Wang, *Knowledge, awareness, and bisimulation*, in: B. Schipper, editor, *Proc. of 14th TARK*, 2013, pp. 61–70.
[23] van Ditmarsch, H., W. van der Hoek and B. Kooi, "Dynamic Epistemic Logic," Synthese Library **337**, Springer, 2007.
[24] Fagin, R. and J. Halpern, *Belief, awareness, and limited reasoning*, Artificial Intelligence **34** (1988), pp. 39–76.
[25] Fan, J., Y. Wang and H. van Ditmarsch, *Almost necessary*, in: R. Goré, B. Kooi and A. Kurucz, editors, *Proc. of 10th AiML Groningen* (2014), pp. 178–196.
[26] Fan, J., Y. Wang and H. van Ditmarsch, *Contingency and knowing whether*, Review of Symbolic Logic (2014), to appear.
[27] Fervari, R., "Relation-Changing Modal Logics," Ph.D. thesis, Universidad Nacional de Córdoba, Argentina (2014).
[28] French, T., "Bisimulation quantifiers for modal logic," Ph.D. thesis, University of Western Australia (2006).
[29] French, T., J. Hales and E. Tay, *A composable language for action models*, in: R. Goré, B. Kooi and A. Kurucz, editors, *Advances in Modal Logic 10* (2014), pp. 197–216.
[30] Grove, A., *Two modellings for theory change*, Journal of Philosophical Logic **17** (1988), pp. 157–170.

[31] Hales, J., *Refinement quantifiers for logics of belief and knowledge* (2011), honours Thesis, University of Western Australia.
[32] Hales, J., *Arbitrary action model logic and action model synthesis*, in: *Proc. of 28th LICS*, IEEE, 2013, pp. 253–262.
[33] Hales, J., T. French and R. Davies, *Refinement quantified logics of knowledge and belief for multiple agents*, in: *Advances in Modal Logic 9* (2012), pp. 317–338.
[34] Halpern, J., *Alternative semantics for unawareness*, Games and Economic Behavior **37(2)** (2001), pp. 321–339.
[35] Heifetz, A., M. Meier and B. Schipper, *Interactive unawareness*, Journal of Economic Theory **130** (2006), pp. 78–94.
[36] van der Hoek, W. and A. Lomuscio, *A logic for ignorance*, Electronic Notes in Theoretical Computer Science **85(2)** (2004), pp. 117–133.
[37] Hollenberg, M., "Logic and bisimulation," Ph.D. thesis, University of Utrecht (1998).
[38] Jensen, M., "Epistemic and Doxastic Planning," Ph.D. thesis, Technical University of Denmark (2014).
[39] Kraus, S., D. Lehmann and M. Magidor, *Nonmonotonic reasoning, preferential models and cumulative logics*, Artificial Intelligence **44** (1990), pp. 167–207.
[40] Lomuscio, A. and M. Ryan, *An algorithmic approach to knowledge evolution*, Artificial Intelligence for Engineering Design, Analysis and Manufacturing **13(2)** (1998), pp. 119–132.
[41] Montgomery, H. and R. Routley, *Contingency and non-contingency bases for normal modal logics*, Logique et Analyse **9** (1966), pp. 318–328.
[42] Segerberg, K., "Classical Propositional Operators," Oxford, Clarendon Press, 1982.
[43] Stalnaker, R., *Knowledge, belief and counterfactual reasoning in games*, Economics and Philosophy **12** (1996), pp. 133–163.
[44] Stirling, C., *The joys of bisimulation*, in: L. Brim, J. Gruska and J. Zlatuska, editors, *Mathematical Foundations of Computer Science 1998 (Proc. of 23rd International Symposium)*, LNCS 1450 (1998), pp. 142–151.
[45] Woodcock, J. and J. Davies, "Using Z — Specification, Refinement and Proof," Prentice Hall, 1996.

A New Perspective on Goals

Barbara Dunin-Kęplicz [1]

Institute of Informatics
University of Warsaw, Poland

Andrzej Szałas [2]

Institute of Informatics
University of Warsaw, Poland
and
Department of Computer and Information Science
Linköping University, Sweden

Abstract

Logical characterizations of goals have been intensively studied in the context of planning and multiagent systems. However, as we show in this paper, goals require further attention. First, the current formalizations are not fully satisfactory as they lack an adequate predictive power, sometimes providing unwanted results. Second, algebras on goals do not have to be entirely compatible with algebras of the underlying logics. Third, even though agents typically reason over incomplete and/or inconsistent belief bases, usually the issues of missing or conflicting information are not directly addressed in the existing formalizations of goals. Last but not least, current approaches typically lead to intractable reasoning.
In this paper we investigate the pragmatics of everyday reasoning about goals as mental attitudes. As human reasoners, we often perceive a new goal as an abstract entity to be achieved. This mental leap helps us to reason about goals on a meta-level without immediately considering their, possibly complex, specification. In our formal approach, we view goals as a sort of abstract objects, further combined with a detailed specification of how to achieve them. The semantics of goals is specified by means of predicates *goal, achieved* and *achievable* that are embedded in user-defined belief bases. This embedding requires a new form of reification.

1 Towards Pragmatic Models of Goals

The relationship between knowledge and action has been of interest of researchers form many fields. Some of the related issues concerning goals have

[1] E-mail: keplicz@mimuw.edu.pl.
[2] E-mails: andrzej.szalas@{mimuw.edu.pl, liu.se}.

been studied mostly in (distributed) artificial intelligence and computer science. Next to it, psychologists have studied the relationship between goals and human creativity. Along this line, a well established psychological theory states that perfectionism as a virtue does not support human creativity. How is this related to goals?

In the context of goals this means that very precise and detailed specification of goals is threatened in a dynamic and, even worse, unpredictable situations. At the other end, when goals are approached flexibly, remaining underspecified until the moment of their realization, we stay open on different ways of achieving them. Accordingly, accepting the initial ignorance about the means of achieving goals, we are prepared to be creative in finding the proper, in the light of current circumstances, ones. In the presented research we will share this point of view, studying matters pertaining goals in multiagent systems. Such issues, apart from their complexity, are typically caused by identifying goals with formulas. This way a concept of a goal is combined with its logical specification. Even though such an approach is tempting by its simplicity, it often leads to undesirable effects when goals naturally inherit some, not always adequate, behaviors of logical operators and connectives. In order to solve this problem, we propose a methodology that permits to isolate and separately analyze goals and their logical specifications.

In principle, we intend to reflect the pragmatics of everyday reasoning about goals. As human reasoners, most of the time we perceive a new goal as *an abstract entity to be achieved*. This mental leap is especially useful in cases of complicated, perhaps long lasting, goals. People tend to reason about these goals of potentially complex structures without considering immediately their precise specification. For example, a person may have goals: "to be a good programmer", "to earn a lot", "to practice freediving" without keeping in mind what do they really mean. Such goals can be considered to be *first-order objects* and denoted by constants. In fact, "to be a good programmer" can mean "to be fluent in Java and UML" or, in other circumstances, "to be able to formulate the required SQL queries". Thus, goals as mental attitudes should remain distinguished from their, more or less detailed, specification. Finally, we are interested in certain properties of goals, like:

- when is a particular goal achieved?
- is a particular goal achievable?
- what are the circumstances blocking its achievement?

This kind of meta-properties may be characterized in terms of possibly complex formulas. Accordingly, the main idea of this paper amounts to separating goals as mental attitudes from different conditions specifying both their achievement as well as their meta-properties.

We need always to remember that in pragmatic applications of intelligent systems, the key requirement is efficiency of representation and reasoning. This means that tractability is our prime prerequisite. However, typical real-life environments are dynamic and unpredictable what, together with the multiplicity

of information sources, leads to gaps and/or inconsistencies in agents' beliefs. Such circumstances call for adequate tools for information completion and disambiguation. Typically, different forms of nonmonotonic/defeasible reasoning techniques can serve for these purposes [8,22,26,28]. They, however, usually violate tractability which can hardly be achieved even in classical propositional logic, without placing certain restrictions on formulas used in knowledge bases. In the case of nonmonotonic or multimodal formalisms, especially when used in their full generality [2,9,16–18,29], the situation is even worse.

Only since recently, in 4QL and 4QL$^+$ [23,25,32] such forms of reasoning became tractable. In these four-valued query languages tractability is achieved by restricting formulas to rules and introducing *modules*, *external literals* [23,25] and *multisource formulas* [32]. Both 4QL and 4QL$^+$ support a modular and layered architecture, and provide a tractable framework for many forms of rule-based reasoning, both monotonic and nonmonotonic. As the underpinning principle, openness of the world is assumed. This may lead to the lack of information. On the other hand, negation, allowed in premises and conclusions of rules, may lead to inconsistencies. To reduce the unknown/inconsistent zones, modules and multisource formulas provide means for:

- application-specific disambiguation of inconsistent information;
- closing the world locally (thus also globally, whenever needed);
- implementing of various forms of nonmonotonic and defeasible reasoning.

In this research we first indicate some issues with modeling goals. Then we propose a new approach which, however, is not meant to be a final one. Our intention is to solve some important issues while enjoying efficient implementation. Accordingly, we show how to understand goals in a tractable framework of paraconsistent belief bases, applying 4QL$^+$ [32] extended by a belief operator. Note that paraconsistency has not been a goal of this research, being rather its unavoidable consequence. However, we deal with quite simple and natural paraconsistent semantics. For discussions of other paraconsistent approaches, see [3–6,10].

The paper is structured as follows. In Section 2 we indicate some issues in modal approaches to representing goals. Then, in Section 3, we outline our approach to modeling the world. Section 4 is devoted to formalization of goals and goal structures. In Section 5 we discuss a simple example illustrating our approach. Section 6 shows how to represent goals in belief bases using the 4QL$^+$ language. Finally, Section 7 concludes the paper.

2 Issues in Modeling Goals

Contemporary intelligent and autonomous systems are realized according to different knowledge representation paradigms. This is especially visible in multiagent systems, implementing different models of agency. Regardless of the implementation tools, all these systems are created to realize their design objectives in the name of their owners. Even though a goal adoption is still

a relatively difficult matter, goals and an agent's decision making process are well developed in belief-desires-intention system (BDI-systems) or belief-goals-intention systems (BGI systems), when we intend to stress the autonomous role of goals, in contrast to (a bit unclear) desires. Among logical formalizations of BGI systems, modal approaches play the prominent role. Such approaches are summarized, e.g., in [16,27]. In the sequel we indicate some issues related to the use of modal logics in formalizing goals.

There are many types of goals' models in agency. The main distinction is between *achievement goals* and *maintenance goals*. In the case of achievement goals, an agent is required to bring about some state of affairs, typically specified by a number of so-called *goal states*. In contrast, maintenance tasks are about avoiding some situations, characterized as specific states. Typically, such states may be considered as a sort of danger for an agent well being or violating its routine activity. As achievement goals are the most commonly studied goals in AI, in the sequel we will deal with them solely.

2.1 Complexity

In order to formalize goals in BDI systems, typically (multi)modal logic **K** is used. Satisfiablity and provability in **K** are PSPACE-complete [20]. The same holds for its multiagent version $\mathbf{K}_n$ [19].

High complexity seems unavoidable in logical formalizations as even the simplest classical propositional logic is complex (with satisfiability being NPTIME-complete and tautology checking being CO-NPTIME-complete). However, we make a shift from reasoning from arbitrary theories to querying belief bases expressed in $4QL^+$. This, on the one hand, allows us to make our framework tractable (see Theorem 6.4) and, on the other hand, to express all PTIME-computable queries (see [24,25,32]). Therefore, in the framework we propose, complexity is no longer a substantial issue.

2.2 Distributing Goal Operator over Conjunction

When modeling goals in real-world applications, some modal properties of goals represented as modal necessity operators remain undesirable. In particular, the $\Box$ operator of (normal) modal logics can be distributed over conjunction. That is:

$$\mathit{goal}(\alpha \wedge \beta) \equiv \big(\mathit{goal}(\alpha) \wedge \mathit{goal}(\beta)\big) \tag{1}$$

For example:

$$\begin{aligned} \mathit{goal}(\text{`buy-a-car'} \wedge \text{`buy-spare-parts'}) \equiv \\ \mathit{goal}(\text{`buy-a-car'}) \wedge \mathit{goal}(\text{`buy-spare-parts'}). \end{aligned} \tag{2}$$

The choice of intentions as arbitrary subsets of goals is different when one has a single goal (‘buy-a-car’ $\wedge$ ‘buy-spare-parts’) and when one has two separate goals: ‘buy-a-car’ and ‘buy-spare-parts’. This is particularly visible when one has no resources to buy a car and has resources to buy spare parts. Then the single goal (‘buy-a-car’ $\wedge$ ‘buy-spare-parts’) is not achievable while the goal

'buy-spare-parts' can be achieved and chosen for realization, yet alone might not be sensible. From that point of view the equivalence (2) (so (1), too) is questionable. Therefore, in our solution the distribution law (1) is neither assumed nor is a consequence of the introduced formalism.

2.3 Inconsistent goals

When *goal* operator is modeled as a $\Box$ in a modal logic then, again by (1), we have:

$$goal(\alpha) \wedge goal(\neg\alpha) \equiv goal(\alpha \wedge \neg\alpha) \equiv goal(\mathsf{f}) \tag{3}$$

However,

- one might have inconsistent goals without having a goal f being impossible to achieve;
- $goal(\mathsf{f})$ is satisfied only in worlds with no alternative worlds, which blocks other, consistent goals;
- $goal(\underbrace{\mathsf{f} \rightarrow \beta}_{\mathsf{t}}) \rightarrow (goal(\mathsf{f}) \rightarrow goal(\beta))$; by generalization we have $goal(\mathsf{t})$, so conclude that $goal(\mathsf{f}) \rightarrow goal(\beta)$. Having $goal(\mathsf{f})$ we then derive $goal(\beta)$, where β is arbitrary. Thus, having inconsistent goals one has any other goal.

To avoid explosion of conclusions, inconsistencies deserve paraconsistent reasoning which is not addressed in (two-valued) modal logics used for formalizing goals. Paraconsistent reasoning used in our solution is summarized in Section 3.

2.4 Incompatibilities between Goal Algebras and Logics

In the tradition of logical formalizations of knowledge or agent systems, goals as mental attitudes are identified with formulas describing them. An important, though not necessarily intended side-effect is that the algebra on goals is induced by the underlying logic. In particular such natural operators on goals like their sequential or parallel composition do not have direct counterparts in logics typically used in the context of goals. For example, if one has a goal first to eat dinner and later to walk, operators of dynamic logics or temporal logics are to be used. However, such choices seriously affect their semantics and complexity. Moreover, when formalizing goals we do not force any particular algebra on goals. Algebraic operations on goals are left for further research.

3 Modeling the World

3.1 The Underlying Four-Valued Logic

We model the world using 4QL$^+$ [32], extending 4QL [23,25]. In order to construct belief bases, we deal with the classical first-order language over a given vocabulary without function symbols. Beliefs are modeled using belief structures, introduces by Dunin-Kęplicz and Szałas in [12–14]. In the following definitions we assume that *Const* is a set of constants, *Var* is a set of variables and Rel is a set of relation symbols, and denote the resulting set of formulas

by $\mathcal{L}$.

Definition 3.1 A *literal* is an expression of the form $R(\bar{\tau})$ or $\neg R(\bar{\tau})$, with $\bar{\tau}$ being a sequence of arguments, $\bar{\tau} \in (\mathit{Const} \cup \mathit{Var})^k$, where k is the arity of R. *Ground literals over Const*, denoted by $\mathbb{G}(\mathit{Const})$, are literals without variables, i.e., with all arguments in *Const*. If $\ell = \neg R(\bar{\tau})$ then $\neg\ell \stackrel{\text{def}}{=} R(\bar{\tau})$. ◁

Remark 3.2 In the rest of the paper we frequently do not specify *Const*. In such cases we always assume that *Const* consists of all and only constants appearing in considered (sets of) expressions. That is, we apply a form of domain closure axiom. ◁

Though we use the classical first-order syntax, the presented semantics substantially differs from the classical one. Namely,

- truth values $\mathbf{t}, \mathbf{i}, \mathbf{u}, \mathbf{f}$ (true, inconsistent, unknown, false) are explicitly present;
- the semantics is based on sets of ground literals rather than on relational structures.

This allows one to deal with the lack of information as well as inconsistencies. As 4QL$^+$ is based on the same principles, it can immediately be used as the implementation tool.

The semantics of propositional connectives is summarized in Table 1.

Table 1. Truth tables for $\wedge$, $\vee$ and $\neg$ (see [23,25,33]).

$\wedge$	$\mathbf{f}$	$\mathbf{u}$	$\mathbf{i}$	$\mathbf{t}$
$\mathbf{f}$	$\mathbf{f}$	$\mathbf{f}$	$\mathbf{f}$	$\mathbf{f}$
$\mathbf{u}$	$\mathbf{f}$	$\mathbf{u}$	$\mathbf{u}$	$\mathbf{u}$
$\mathbf{i}$	$\mathbf{f}$	$\mathbf{u}$	$\mathbf{i}$	$\mathbf{i}$
$\mathbf{t}$	$\mathbf{f}$	$\mathbf{u}$	$\mathbf{i}$	$\mathbf{t}$

$\vee$	$\mathbf{f}$	$\mathbf{u}$	$\mathbf{i}$	$\mathbf{t}$
$\mathbf{f}$	$\mathbf{f}$	$\mathbf{u}$	$\mathbf{i}$	$\mathbf{t}$
$\mathbf{u}$	$\mathbf{u}$	$\mathbf{u}$	$\mathbf{i}$	$\mathbf{t}$
$\mathbf{i}$	$\mathbf{i}$	$\mathbf{i}$	$\mathbf{i}$	$\mathbf{t}$
$\mathbf{t}$	$\mathbf{t}$	$\mathbf{t}$	$\mathbf{t}$	$\mathbf{t}$

	$\neg$
$\mathbf{f}$	$\mathbf{t}$
$\mathbf{u}$	$\mathbf{u}$
$\mathbf{i}$	$\mathbf{i}$
$\mathbf{t}$	$\mathbf{f}$

Observe that definitions of $\wedge$ and $\vee$ reflect minimum and maximum w.r.t. the ordering:

$$\mathbf{f} < \mathbf{u} < \mathbf{i} < \mathbf{t}, \tag{4}$$

As advocated, e.g., in [1,23,32,33], such a truth ordering appears to be natural and reflecting intuitions of the classical two-valued logic. For example, a conjunction is true if all its operands are true, etc. Negation behaves classically on truth values $\mathbf{t}$ and $\mathbf{f}$. If the truth value of a formula is unknown (inconsistent) then the truth value of its negation is unknown (inconsistent), too. This justifies the semantics of negation. Note that truth tables shown in Table 1, when restricted to truth values $\{\mathbf{t}, \mathbf{u}, \mathbf{f}\}$ or $\{\mathbf{t}, \mathbf{i}, \mathbf{f}\}$, are those of Łukasiewicz logic $\mathbf{Ł_3}$ as well as Kleene logic $\mathbf{K_3}$ (see [7,30]). For further justification of the assumed semantics, see [32].

Let $v : \mathit{Var} \longrightarrow \mathit{Const}$ be a *valuation of variables*. For a literal ℓ, by $\ell(v)$ we understand the ground literal obtained from ℓ by substituting each variable x occurring in ℓ by constant $v(x)$.

Definition 3.3 The *truth value* of a literal ℓ w.r.t. a set of ground literals L and valuation v, denoted by $\ell(L, v)$, is defined as follows:

$$\ell(L,v) \stackrel{\text{def}}{=} \begin{cases} \mathfrak{t} & \text{if } \ell(v) \in L \text{ and } (\neg\ell(v)) \notin L; \\ \mathfrak{i} & \text{if } \ell(v) \in L \text{ and } (\neg\ell(v)) \in L; \\ \mathfrak{u} & \text{if } \ell(v) \notin L \text{ and } (\neg\ell(v)) \notin L; \\ \mathfrak{f} & \text{if } \ell(v) \notin L \text{ and } (\neg\ell(v)) \in L. \end{cases}$$ ◁

For a formula $\alpha(x)$ with a free variable x and $c \in \mathit{Const}$, by $\alpha(x)^x_c$ we understand the formula obtained from α by substituting all free occurrences of x by c. Definition 3.3 is extended to all formulas in Table 2, where $\alpha, \beta \in \mathcal{L}$ denote first-order formulas, v is a valuation of variables, L is a set of ground literals, and the semantics of propositional connectives appearing at righthand sides of equivalences is given in Table 1.

Table 2. Semantics of first-order formulas, where $\min$ and $\max$ are calculated w.r.t. ordering (4).

- if α is a literal then $\alpha(L, v)$ is defined in Definition 3.3;
- $(\neg\alpha)(L, v) \stackrel{\text{def}}{=} \neg(\alpha(L, v))$;
- $(\alpha \circ \beta)(L, v) \stackrel{\text{def}}{=} \alpha(L, v) \circ \beta(L, v)$, where $\circ \in \{\vee, \wedge\}$;
- $\big(\forall x \in \mathit{Const}(\alpha(x))\big)(L, v) \stackrel{\text{def}}{=} \min\limits_{a \in \mathit{Const}} \{(\alpha^x_a)(L, v)\}$;
- $\big(\exists x \in \mathit{Const}(\alpha(x))\big)(L, v) \stackrel{\text{def}}{=} \max\limits_{a \in \mathit{Const}} \{(\alpha^x_a)(L, v)\}$.

Note that the set *Const* in $\forall x \in \mathit{Const}(\ldots)$ and $\exists x \in \mathit{Const}(\ldots)$ (Table 2) is usually given by considered belief bases, so we will rather write $\forall x(\ldots)$ and $\exists x(\ldots)$ whenever *Const* will be known from context.

3.2 Belief Bases

If S is a set then by $\textsc{Fin}(S)$ we understand the set of all finite subsets of S. We further assume that *Const* is always finite and by $\mathbb{C} \stackrel{\text{def}}{=} \textsc{Fin}(\mathbb{G}(\mathit{Const}))$ we denote the set of all finite sets of ground literals over the set of constants *Const*.

In what follows, for simplicity, we assume that updates are arbitrary mappings transforming belief bases. For requirements on updates see the reach literature on belief update and belief revision, e.g, [11,21,31].

Definition 3.4

- By a *belief base* we understand any finite set Δ of finite sets of ground literals over a set *Const*, i.e., any finite set $\Delta \subseteq \mathbb{C}$.

- By an *update* we mean a mapping transforming a belief base Δ into a belief base Δ', i.e., a mapping of the sort $\text{FIN}(\mathbb{C}) \longrightarrow \text{FIN}(\mathbb{C})$.
- By a *plan* we understand a finite sequence of updates. The *result of executing plan* p *on* Δ is denoted by $p(\Delta)$. For $p = \langle p_1, \ldots, p_k \rangle$, $p(\Delta) \stackrel{\text{def}}{=} p_k(\ldots(p_1(\Delta))\ldots)$. ◁

One may wonder why belief bases are sets of sets of literals rather than a single set of literals. This follows from two basic reasons:

- each set of ground literals in a belief base may represent a specific view on the considered reality (e.g., one view as perceived by a single sensor or a sensor platform, another as perceived by a camera, etc.);
- each set of ground literals in a belief base may represent a possible situation in a way similar to possible worlds in Kripke models. [3]

The following example is intended to illustrate these reasons by dealing with various views and situations.

Example 3.5 The belief base Δ consisting of sets:

$$\{hot(a), \neg hot(b), red(a)\}, \{\neg big(a), big(b), \neg red(a)\}$$

may represent:

- sensor's and camera's views on objects a and b;
- descriptions of possible situations: one where $hot(a), red(a)$ are true and $hot(b)$ is false and the second in which $big(b)$ is true and $big(a), red(a)$ are false (note that literals not listed in a given set are unknown).

Note also that in the light of Remark 3.2, the set of constants of Δ is *Const* $= \{a, b\}$. ◁

In the context of planning, updates typically result from effects of actions. Since plans are finite sequences of updates, one can consider plans to be single updates. We do not fix any particular syntax for updates using in examples self-explanatory "update commands". Of course, updates are arbitrary transformations that can remove/add sets of literals, etc. ◁

Example 3.6 Consider belief base Δ of Example 3.5 and a sequence of two updates: the first adding fact $\neg red(a)$ to the first set in Δ and the second replacing fact $\neg big(a)$ by $big(a)$ in the second set of Δ. The resulting database is:

$$\{hot(a), \neg hot(b), red(a), \neg red(a)\}, \{big(a), big(b), \neg red(a)\}.$$

These updates can be gathered and considered to be a single update.

By *information ordering* we understand the ordering on truth values shown in Figure 1. This ordering reflects the process of gathering and fusing information. Starting from the lack of information, in the course of belief acquisition,

[3] Note, however, that there are substantial differences between belief bases and and Kripke models.

evidences supporting or denying hypotheses are collected. This finally permits one to decide about the truth value of the hypotheses.

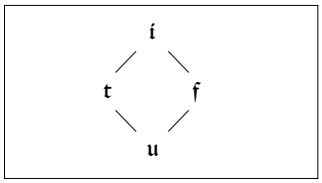


Figure 1. Information ordering.

To express beliefs we allow formulas of the form $\mathrm{Bel}(\alpha)$ with the semantics proposed in [14] and defined as follows. [4]

Definition 3.7 Let Δ be a belief base, v : *Var* $\longrightarrow$ *Const* be a valuation of variables and α be a formula. We define the semantics of the belief operator by:

$$(\mathrm{Bel}(\alpha))(\Delta, v) \stackrel{\mathrm{def}}{=} \textsc{Lub}\{\alpha(D, v) \mid D \in \Delta\},$$

where LUB denotes the least upper bound wrt the ordering shown in Figure 1. ◁

In cases when v is inapplicable, it is frequently omitted.

Example 3.8 Let Δ be the belief base considered in Example 3.5 and let $v(x) = a$ and $v(y) = b$. Then:

$$(\mathrm{Bel}(hot(x))(\Delta, v) = \mathbf{t};\quad (\mathrm{Bel}(hot(y))(\Delta, v) = \mathbf{f};$$
$$(\mathrm{Bel}(red(x))(\Delta, v) = \mathbf{i};\quad (\mathrm{Bel}(red(y))(\Delta, v) = \mathbf{u}.$$ ◁

In [12,13] deterministic belief structures and epistemic profiles have been introduced and investigated, and further developed to their indeterministic version in [14]. They can be used to model belief and goal formation. In the current paper, in order to simplify presentation, we abstract from belief/goal formation process focusing on the already established belief bases. However, belief structures can be used for a more comprehensive formalization of beliefs and goals. Their application to reasoning about group belief is investigated in [15].

4 What are the Goals?

We shall consider goals to be first-order objects, having a separate sort $\mathcal{G}$ for representing them. Further on we assume that sets of constants representing goals are given and that the set corresponding to $\mathcal{G}$ consists of respective constants. Slightly abusing notation we shall then use $\mathcal{G}$ to denote the sort of goals as well as the set of constants representing goals.

[4] However, here we consider the case of a single agent only.

We assume that there are conditions associated with goals indicating when they are achieved. Such conditions are given by mapping $\gamma : \mathcal{G} \longrightarrow \mathcal{L}$ such that $\gamma(g)$ returns as a result a first-order formula. The intuitive meaning is that the goal g is achieved in belief bases where formula $\mathrm{Bel}(\gamma(g))$ is true. When $\gamma(g)$ has free variables, we want the formula to be true for all respective tuples, that is, we universally close $\gamma(g)$.

Example 4.1 Consider the belief base Δ of Example 3.5. Assume that the goal, say g is to make objects involved not red and hot. That is:

$$\gamma(g) = \big(\neg red(x) \wedge hot(x)\big). \tag{5}$$

To verify whether goal g is achieved, we evaluate in Δ the query:

$$\mathrm{Bel}(\forall x(\neg red(x) \wedge hot(x))), \tag{6}$$

being universally closed $\gamma(g)$. The formula (6) is $\mathfrak{f}$ since the formula in the scope of $\mathrm{Bel}()$ in (6) is $\mathfrak{f}$ in the first set of Δ and $\mathfrak{u}$ in the second set of Δ.

On the other hand, the goal g' with:

$$\gamma(g') = \exists x(\neg red(x) \vee hot(x))$$

is achieved in Δ since $\gamma(g')$ is true in Δ. ◁

In intelligent systems like multiagent or robotics systems, the set of goals is typically fixed and well-defined. For example, a given robot may be able to clean surfaces and cut grass but perhaps not be designed for other activities like moving heavy objects or proving mathematical theorems. All in all, various types of agents offer a whole spectrum of actions they are able to perform. These actions are then combined into, possibly social (that is realized cooperatively by a group of agents) plans. Such plans may be predefined or generated on demand from the first principle. A fixed library of plans naturally limits applicability of the system, especially in dynamic and unpredictable environments, but is computationally more effective. Goal structures defined below reflect this situation. From another viewpoint, goal structures can be considered to be a snapshot of certain situation. That is, we assume that particular goals are specified and that there is a library of plans, predefined or generated so far. However, the existing plans do not necessarily allow to achieve *all* considered goals.

Definition 4.2 By a *goal structure* we mean a triple $\langle \mathbb{P}, \mathcal{G}, \gamma \rangle$, where:

- $\mathbb{P}$ is a finite set of plans (a *library* of plans);
- $\mathcal{G}$ is a set of constants denoting (all) goals;
- γ is a mapping from $\mathcal{G}$ to the set of first-order formulas $\mathcal{L}$; intuitively, $\gamma(g)$ is a formula specifying when goal g is achieved. ◁

We also introduce predicates:

- $goal(g)$ meaning that g is one of goals;

- *achieved*(g) meaning that the goal g is actually achieved;
- *achievable*(g) meaning that the goal g can be achieved using plans from the library.

The semantics of the introduced language is defined below.

Definition 4.3 Let $\mathbb{P}, \mathcal{G}, \gamma$ be as in Definition 4.2, Δ be a belief base and v be a valuation of variables into *Const*. Then:

$$\begin{array}{lll}
\langle \mathbb{P}, \mathcal{G}, \gamma \rangle, \Delta, v \models \textit{goal}(x) & \text{iff} & v(x) \in \mathcal{G}; \\
\langle \mathbb{P}, \mathcal{G}, \gamma \rangle, \Delta, v \models \textit{achieved}(x) & \text{iff} & \big(\mathrm{Bel}(\gamma(v(x)))\big)(\Delta, v) = \mathbf{t}; \\
\langle \mathbb{P}, \mathcal{G}, \gamma \rangle, \Delta, v \models \textit{achievable}(x) & \text{iff} & \text{there is } p \in \mathbb{P} \text{ such that } p(\Delta) = \Delta' \\
 & & \text{and } \langle \mathbb{P}, \mathcal{G}, \gamma \rangle, \Delta', v \models \gamma(v(x)).
\end{array}$$

For other first-order formulas we add the clause:

$$\langle \mathbb{P}, \mathcal{G}, \gamma \rangle, \Delta, v \models rel(\bar{x}) \text{ iff } \big(\mathrm{Bel}(rel(\bar{x}))\big)(\Delta, v) = \mathbf{t}.$$

The clause is then extended to cover the whole language by analogy with Table 2 and Definition 3.7. ◁

Note that there is a substantial difference between *achieved* and *achievable*. The former is evaluated in the "current" belief base Δ while *achievable* is evaluated in Δ' being a result of a suitable plan.

Of course, plans may lead to different outcomes, so "there is $p \in \mathbb{P}$" appearing in the third clause of Definition 4.3 is to be understood as an existential quantifier evaluated according to ordering (4).

All predicates in the language are four-valued, in particular *goal*, *achieved*, *achievable*. We find this one of the advantages of the proposed approach. Typically, inconsistency indicates a problem to be solved and, in real life, such problems can indeed frequently be solved. To illustrate this point consider inconsistency of *achievable*(g). For example, g may be a customer's goal to buy a car. Buying a car requires certain amount of money. This, in turn, may be inconsistent with financial abilities of the customer. The intelligent system may have no means to resolve this inconsistency. It then informs the customer that *achievable* ('buy_a_car') is inconsistent. However, the customer most probably knows that, in this case, achieving the goal of buying a car requires finding additional sources of money, like taking loan, etc. Once the required amount of money is collected, the inconsistency is solved. In this and similar cases, solving inconsistency of *achievable* indeed appears useful. [5]

Remark 4.4 Let us emphasize that treating goals as elements of the domain allows us to add priorities on goals and express their other properties by introducing new relations, like *preferred*(g_1, g_2), *liked*(g), $good(g)$, $risky(g)$, etc. ◁

[5] In other cases, like involving missing capabilities or resources, one might, e.g., try to find an agent or agents capable of achieving the goal and delegate the goal.

5 An Example

Consider the following simple scenario.

> I want to drive from town a to town b. I also do not want to be hungry being in b, so want to have lunch during my trip.

I therefore have goals 'be_in_b' and 'not_hungry_in_b' for which γ is specified as follows:

$$\gamma(\text{'be\_in\_b'}) = in(b)$$
$$\gamma(\text{'not\_hungry\_in\_b'}) = lunch\_between(a,b).$$

A library of plans contains plans reflecting the following sequences of actions:

1. $drive(a,c)$; $eat\_at(c)$; $drive(c,b)$;
2. $drive(a,c)$; $coffee\_at(c)$; $drive(c,b)$;
3. $coffee\_at(a)$; $drive(a,b)$.

Since we encode plans as updates, the library of plans $\mathbb{P}$ consists of three plans:
p_1: $\underbrace{\text{'replace'}\ in(a)\ \text{'by'}\ in(c)}_{drive(a,c)}$; $\underbrace{\text{'add'}lunch\_between(a,b)}_{eat\_at(c)}$; $\underbrace{\text{'replace'}\ in(c)\ \text{'by'}\ in(b)}_{drive(c,b)}$;
p_2: $\underbrace{\text{'replace'}\ in(a)\ \text{'by'}\ in(c)}_{drive(a,c)}$; $\underbrace{\text{'make'}\ thirsty(c)\ \text{'false'}}_{coffee\_at(c)}$; $\underbrace{\text{'replace'}\ in(c)\ \text{'by'}\ in(b)}_{drive(c,b)}$;
p_3: $\underbrace{\text{'make'}\ thirsty(a)\ \text{'false'}}_{coffee\_at(a)}$; $\underbrace{\text{'replace'}\ in(a)\ \text{'by'}\ in(b)}_{drive(a,b)}$. Assume that the current belief base is:

$$\Delta = \big\{\{in(a), thirsty(a)\}, \{in(a), \neg thirsty(a)\}\big\}. \tag{7}$$

Plans, when applied to Δ, result in:

$$p_1(\Delta) = \big\{\{in(b), thirsty(a), lunch\_between(a,b)\}, \{in(b), \neg thirsty(a), lunch\_between(a,b)\}\ \big\}; \tag{8}$$
$$p_2(\Delta) = \big\{\{in(b), thirsty(a), \neg thirsty(c)\}, \{in(b), \neg thirsty(a), \neg thirsty(c)\}\ \big\}; \tag{9}$$
$$p_3(\Delta) = \big\{\{in(b), \neg thirsty(a)\}, \{in(b), \neg thirsty(a)\}\ \big\}. \tag{10}$$

Now:

- goal 'be_in_b' is satisfied in $p_1(\Delta)$, $p_2(\Delta)$ and in $p_3(\Delta)$;
- goal 'not_hungry_in_b' is satisfied in $p_1(\Delta)$ but not in $p_2(\Delta)$ nor in $p_3(\Delta)$.

Therefore, both considered goals are achievable.

6 Representing Goal Structures in 4QL$^+$

To define belief bases we use 4QL$^+$, where belief bases are distributed among modules. It is then relatively easy to allocate goals into a dedicated mod-

ule, called 'goals'. As 4QL$^+$ allows for multisource first-order formulas in the premises of rules, it is almost immediate to define *goal* and *achieved* predicates. However, we need the $\mathrm{Bel}()$ operator.

Note that 4QL and 4QL$^+$ modules can be identified with their well-supported models being finite sets of ground literals [23,25,32]. [6] A belief base may then be represented by a finite set of 4QL or 4QL$^+$ modules. As the $\mathrm{Bel}()$ operator is evaluated in belief bases, we extend 4QL$^+$ by allowing $\mathrm{Bel}_M()$, where M is a set of module names, in the premises of rules. Note that this does not affect complexity of 4QL$^+$ (see also [14]).

Definition 6.1 Let Δ be a belief base. We say that the set of modules M *represents* Δ iff $\Delta = \{\textsc{Wsm}(m) \mid m \in M\}$, where $\textsc{Wsm}(m)$ denotes the well-supported model of m. ◁

Given a valuation v : *Var* $\longrightarrow$ *Const* of variables and a set of modules Δ, the semantics of $\mathrm{Bel}_\Delta()$ operator is given by:

$$\mathrm{Bel}_M(\alpha)(v) \stackrel{\mathrm{def}}{=} (\mathrm{Bel}(\alpha))(\textsc{Wsm}(M), v),$$

where $\textsc{Wsm}(M) \stackrel{\mathrm{def}}{=} \{\textsc{Wsm}(m) \mid m \in M\}$ and $\mathrm{Bel}(\alpha)$(WSM(M),v) is defined in Definition 3.7.

In what follows we often identify modules with their names. We will allow finite sets of module names to be arguments of relations. Such sets are treated as constants and do not affect data complexity of the language.

Continuing the example of Section 5, the belief base (7) is represented by two modules shown in Table 3.

Table 3. Implementation of belief base (7).

module delta_1:	module delta_2:
in(a).	in(a).
thirsty(a).	¬thirsty(a).
end.	end.

Similarly, belief base (8) can be represented by modules `p11`, `p12`, belief base (9) by modules `p21`, `p22` and belief base (10) – by modules `p31`, `p32`.

Relations *achieved* and *achievable* are evaluated in belief bases, so we have to add arguments indicating adequate sets of module names:

- *achieved*(g, M), meaning that the goal g is achieved in belief base $\textsc{Wsm}(M)$;
- *achievable*(g, M), meaning that the goal g can be achieved using a plan from the library, starting from belief base $\textsc{Wsm}(M)$.

We also have to encode plans. Therefore, we use relation:

- *plan*(P, M, M'), meaning that plan P transforms belief base $\textsc{Wsm}(M)$ into the belief base $\textsc{Wsm}(M')$.

[6] For any 4QL (4QL$^+$) module there is a unique well-supported model and it is computable in deterministic polynomial time w.r.t the size of the domain.

However, the predicate *plan* requires database updates to verify whether a given plan leads from the belief base WSM(M) to WSM(M'). Updates are not part of 4QL or 4QL$^+$ and have to be executed by an external application, verifying whether a plan transforms WSM(M) to WSM(M'). Note that, for a given initial belief base represented by modules M, we only have to compute n facts of the form `plan(P,M,M')`, where n is the number of plans in the library.

The relations *goal*, *achieved*, *achievable* and *plan* are allocated to the module `goal_struct` provided in Table 4.

Table 4. Implementation of goals considered in Section 5, where `p1,p2,p3` are names of plans considered in Section 5, and `delta={delta_1,delta_2}` is shown in Table 3.

```
module goal_struct:
  ...
  goal('be_in_b').
  goal('not_hungry_in_b').
  plan(p1,delta,{p11,p12}).
  plan(p2,delta,{p21,p22}).
  plan(p3,delta,{p31,p32}).
  achieved('be_in_b',M):- Bel_M(in('b')).
  ¬achieved('be_in_b',M):- ¬Bel_M(in('b')).
  achieved('not_hungry_in_b',M):- Bel_M(lunch_between(a,b)).
  ¬achieved('not_hungry_in_b',M):- ¬Bel_M(lunch_between(a,b)).
  achievable('be_in_b',M):- ∃ p∈ P(plan(p,M,M')∧achieved('be_in_b',M').
  ¬achievable('be_in_b',M):- ¬∃ p∈ P(plan(p,M,M')∧achieved('be_in_b',M').
  ...
end.
```

Using the obtained modules one can ask queries like:

```
goal_struct.(achievable('be_in_b',delta) ∧
             achievable('not_hungry_in_b',delta)),
```

or:

```
goal_struct.(∃p ∈ P (plan(p,delta,M') ∧ achieved('be_in_b',M') ∧
                     achieved('not_hungry_in_b',M'))).
```

The general construction of modules related to goals is defined below.

Definition 6.2 Let Δ be a belief base, $G = \langle \mathbb{P}, \mathcal{G}, \gamma \rangle$ be a goal structure and let $plans$ be a set of fresh constants denoting plans in $\mathbb{P}$. By a *G-extension* of Δ we mean a belief base obtained from Δ by adding:

- a set of 4QL$^+$ modules representing Δ with $delta$ being the set of names of these modules;
- for each $p \in plans$, module representing p(WSM$(delta)$);
- a new module containing, for each $g \in \mathcal{G}$ and $p \in plans$,
  - Fact: `goal(g)`.
  - Rules:

    `plan(p,delta,M').` for M' being a set of 4QL$^+$ modules representing

$p(\Delta)$,

achieved(g,M):- $\mathtt{Bel}_{\mathtt{M}}(\gamma(\mathtt{g}))$.

$\neg$achieved(g,M):- $\neg\mathtt{Bel}_{\mathtt{M}}(\gamma(\mathtt{g}))$.

achievable(g,M):- $\exists$P$\in$plans(plan(P,M,M')$\wedge$achieved(g,M')).

$\neg$achievable(g,M):- $\neg\exists$P$\in$plans(plan(P,M,M')$\wedge$achieved(g,M')). $\triangleleft$

When complexity of the approach is concerned, the only problematic predicate is *achievable* as its complexity depends on updates and plans. Even when updates and plans are efficiently computable, complexity is also hidden in the recursive clause for *achievable* in Definition 4.3. The following example illustrates the problem.

Example 6.3 Consider goals g, h, for which we have:

$$\gamma(g) = \textit{achievable}(h);$$
$$\gamma(h) = \textit{achievable}(g).$$

In this case we deal with a loop: to verify whether goal g can be achieved, we verify whether there is a plan leading to a belief base, where formula $\mathrm{Bel}(\textit{achievable}(h))$ is true. This, in turn calls for a similar verification concerning g. $\triangleleft$

When predicate *achievable* does not occur in the results of γ, we can replace the third clause of Definition 4.3 by:

$$\langle \mathbb{P}, \mathcal{G}, \gamma \rangle, \Delta, v \models \textit{achievable}(x) \text{ iff } \begin{array}{l} \text{there is } p \in \mathbb{P} \text{ such that } p(\Delta) = \Delta' \\ \text{and } \big(\mathrm{Bel}(\gamma(v(x)))\big)(\Delta', v) = \mathtt{t}. \end{array}$$

In this case we have the following theorem which can be proved using complexity results for 4QL and 4QL$^+$ [23,25,32].

Theorem 6.4 *Let $G = \langle \mathbb{P}, \mathcal{G}, \gamma \rangle$ be a fixed goal structure,* [7] *and let Δ be a G-extension of a belief base. Assume further that:*

- *for any $p \in \mathbb{P}$ and belief base Δ, $p(\Delta)$ is* PTIME*-computable w.r.t. the size of Δ;*
- *the predicate 'achievable' does not occur in results of γ.*

Then models for belief bases constructed for G and implemented in 4QL$^+$ are PTIME*-computable w.r.t. the size of Δ. Thus, queries expressed by (multisource) first-order formulas to such databases can also be computed in* PTIME *w.r.t. the size of Δ.* $\triangleleft$

7 Conclusions

In this paper we analyzed some matters related to the formalization of goals in contemporary intelligent systems, like multiagent systems. We have proposed an approach addressing the selected issues, by keeping a clear separation between a concept of goal (thought here as a mental object), and its actual

[7] That is, the cardinality of $\mathbb{P}$ and $\mathcal{G}$ as well as the length of formulas $\gamma(g)$, for $g \in \mathcal{G}$, are considered to be constant.

meaning (logical specification). This isolation turned up to be relatively easy to achieve with the use of reification – a classical AI technique that have been often exploited in different forms of planning.

Our proposal is not meant to be a final and exhaustive one. We have mainly investigated how to understand *achievement goals* in a tractable framework of paraconsistent belief bases, applying 4QL$^+$ with $\mathrm{Bel}()$ operators. The four-valued language used to build a realistic world model, naturally introduces new cognitive situations when compared with the two-valued world. Therefore, a natural extension of our framework, including *maintenance goals*, is postponed for future investigation.

In future research we also plan:

- to extend the framework to many agents and groups of agents;
- to add algebras on goals;
- to extend ontology by including subgoals, delegation, etc.

Acknowledgments

We would like to thank anonymous reviewers for useful comments and suggestions. The paper has been supported by the Polish National Science Centre grant 2011/01/B/ST6/02769.

References

[1] de Amo, S. and M. Pais, *A paraconsistent logic approach for querying inconsistent databases*, International Journal of Approximate Reasoning **46** (2007), pp. 366–386.
[2] Alur, R., T. Henzinger and O. Kupferman, *Alternating-time Temporal Logic*, Journal of the ACM **49** (2002), pp. 672–713.
[3] Arieli, O. and A. Avron, *The value of the four values*, Artificial Intelligence **102** (1998), pp. 97–141.
[4] Avron, A., J. Ben-Naim and B. Konikowska, *Logics of reasonable information sources*, in: F. Esteva, J. Gispert and F. Manya, editors, *ISMVL* (2010), pp. 61–66.
[5] Béziau, J.-Y., W. Carnielli and D. Gabbay, editors, "Handbook of Paraconsistency," College Publications, 2007.
[6] Blair, H. and V. Subrahmanian, *Paraconsistent logic programming*, Theor. Comput. Sci. **68** (1989), pp. 135–154.
[7] Bolc, L. and P. Borowik, "Many-Valued Logics, 1. Theoretical Foundations," Springer, 1992.
[8] Brewka, G., "Non-Monotonic Reasoning: Logical Foundations of Commonsense," Cambridge University Press, 1991.
[9] Cadoli, M. and M. Schaerf, *A survey on complexity results for non-monotonic logics*, Journal Logic Programming **17** (1993), pp. 127–160.
[10] Damásio, C. V. and L. M. Pereira, *A survey of paraconsistent semantics for logic programs*, in: D. M. Gabbay and P. Smets, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems, vol. 2* (1998), pp. 241–320.
[11] Delgrande, J. P., Y. Jin and F. J. Pelletier, *Compositional belief update*, J. Artif. Intell. Res. (JAIR) **32** (2008), pp. 757–791.
[12] Dunin-Kęplicz, B. and A. Szałas, *Epistemic profiles and belief structures*, in: *Proc. KES-AMSTA 2012: Agents and Multi-agent Systems: Technologies and Applications*, LNCS **7327** (2012), pp. 360–369.
[13] Dunin-Kęplicz, B. and A. Szałas, *Taming complex beliefs*, Transactions on Computational Collective Intelligence XI **LNCS 8065** (2013), pp. 1–21.

[14] Dunin-Kęplicz, B. and A. Szałas, *Indeterministic belief structures*, in: G. Jezic, M. Kusek, I. Lovrek, R. Howlett and L. C.J., editors, *Agent and Multi-Agent Systems: Technologies and Applications*, Advances in Intelligent Systems and Computing **296**, 2014, pp. 57–66.
[15] Dunin-Kęplicz, B., A. Szałas and R. Verbrugge, *Tractable reasoning about group beliefs*, in: F. Dalpiaz, J. Dix and M. B. van Riemsdijk, editors, *Engineering Multi-Agent Systems*, LNCS **8758** (2014), pp. 328–350.
[16] Dunin-Kęplicz, B. and R. Verbrugge, "Teamwork in Multi-Agent Systems. A Formal Approach," John Wiley & Sons, Ltd., 2010.
[17] Fagin, R., J. Halpern, Y. Moses and M. Vardi, "Reasoning About Knowledge," The MIT Press, 2003.
[18] Gottlob, G., *Complexity results for nonmonotonic logics*, J. Logic Computat. **2** (1992), pp. 397–425.
[19] Halpern, J. and Y. Moses, *A guide to completeness and complexity for modal logics of knowledge and belief*, Artif. Intell. **54** (1992), pp. 319–379.
[20] Ladner, R., *The computational complexity of provability in systems of modal propositional logic*, SIAM J. Comput. **6** (1977), pp. 467–480.
[21] Lang, J., *Belief update revisited*, in: M. Manuela M. Veloso, editor, *Proc. IJCAI 2007*, 2007, pp. 2517–2522.
[22] Łukaszewicz, W., "Non-Monotonic Reasoning - Formalization of Commonsense Reasoning," Ellis Horwood, 1990.
[23] Małuszyński, J. and A. Szałas, *Living with inconsistency and taming nonmonotonicity*, in: O. de Moor, G. Gottlob, T. Furche and A. Sellers, editors, *Datalog 2.0*, LNCS **6702** (2011), pp. 384–398.
[24] Małuszyński, J. and A. Szałas, *Logical foundations and complexity of 4QL, a query language with unrestricted negation*, Journal of Applied Non-Classical Logics **21** (2011), pp. 211–232.
[25] Małuszyński, J. and A. Szałas, *Partiality and inconsistency in agents' belief bases*, in: D. Barbucha, M. Le, R. Howlett and L. Jain, editors, *Proc. of the 7th KES AMSTA Conference*, Frontiers in Artificial Intelligence and Applications **252** (2013), pp. 3–17.
[26] Marek, V. and M. Truszczyński, "Nonmonotonic Logic," Springer-Verlag, 1993.
[27] Meyer, J.-J. C., J. Broersen and A. Herzig, *BDI logics*, in: H. van Ditmarsch, J. Halpern, W. van der Hoek and B. Kooi, editors, *Handbook of Epistemic Logic*, 2015, to appear.
[28] Nute, D., *Defeasible logic*, in: *Handbook of Logic in Artificial Intelligence and Logic Programming*, 1994, pp. 353–395.
[29] Pauly, M., *A modal logic for coalitional power in games*, Journal of Logic and Computation **12** (2002), pp. 149–166.
[30] Rescher, N., "Many-Valued Logic," McGraw Hill, 1969.
[31] Spruit, P., R. Wieringa and J.-J. C. Meyer, *Aziomatization, declarative semantics and operational semantics of passive and active updates in logic databases*, J. Log. Comput. **5** (1995), pp. 27–70.
[32] Szałas, A., *How an agent might think*, Logic Journal of the IGPL **21** (2013), pp. 515–535.
[33] Vitória, A., J. Małuszyński and A. Szałas, *Modeling and reasoning with paraconsistent rough sets*, Fundamenta Informaticae **97** (2009), pp. 405–438.

Varieties of Belief and Probability

Jan van Eijck [1]

CWI and ILLC
Amsterdam

Abstract

For reasoning about uncertain situations, we have probability theory, and we have logics of knowledge and belief. How does elementary probability theory relate to epistemic logic and the logic of belief? The paper focuses on the notion of betting belief, and interprets a language for knowledge and belief in two kinds of models: epistemic neighbourhood models and epistemic probability models. It is shown that the first class of models is more general in the sense that every probability model gives rise to a neighbourhood model, but not vice versa. The basic calculus of knowledge and betting belief is incomplete for probability models. These formal results were obtained in Van Eijck and Renne [9].

Keywords: Belief, betting, chance, foundations of subjective probability, Bayesian conditioning, neighbourhood models.

1 Introduction

Elementary probability theory, in the subjective or Bayesian style, is fascinating for cognitive scientists, for there is a marked contrast between fast error-prone assessment of chance and the slow but more accurate calculation of subjective probabilities using conditioning. Interest is added by the fact that belief about chance is an important basis of rational decision making and intelligent interaction. I know from our collaboration in the Games, Actions, and Social Software Project at NIAS that resulted in [11] and [12], that this is the stuff that Rineke loves.

Probability theorists like to view the difference between logic and probability as a difference in subject matter. Logic is the topic of reasoning about certainty, while probability theory teaches us how to reason about uncertainty. Guess which discipline has the most relevance to everyday life? Still, the probability theorists are right: epistemic or Bayesian probability can be viewed as an extension of propositional logic with hypotheses, i.e., basic propositions whose truth or falsity is uncertain. But logic has something to say, too, about reasoning under uncertainty: we have epistemic logic, doxastic logic, default

[1] Email: jve@cwi.nl

logic, logic of conditionals, and so on. So it is natural to ask how the perspectives of logic and probability theory on knowledge and belief are related. Frank Ramsey [27] considered the theory of probability as a branch of logic where arguments can be inconclusive. I wholeheartedly agree. In this paper I will argue that there is room for logics with a more general interpretation than probability measures. As an example of those, I explore neighbourhood models for a language of knowledge and belief as willingness to bet, and compare them with probabilistic models for the same language. The paper is light on formal definitions and proofs. For these, the reader is referred to Van Eijck and Renne [9].

The paper starts with some remarks, in Section 2, on the foundations of probability theory, as a comment on the views of Christiaan Huygens on probability. This is connected to the foundations of subjective probability on rational betting behaviour proposed by Ramsey, de Finetti and Savage, and, in Section 3, to the key role of probability in decision theory, which we owe to Von Neumann and Morgenstern. Section 4 introduces the notion of betting belief and compares this to some other notions of belief. In Section 5 I show that betting belief allows for a crisp analysis of the lottery puzzle, at the price of sacrificing closure of belief under conjunction. Section 6 presents a complete calculus for epistemic models with belief neighbourhoods, and Section 7 proves an incompleteness result for the calculus of betting belief with respect to probabilistic models. This shows that the logic of betting belief describes a more general kind of situation than is covered by probability models. Section 8 concludes.

2 Christiaan Huygens on the foundations of probability

Probability theory was invented by Pierre de Fermat and Blaise Pascal around 1650. The Dutch mathematician, astronomer, physicist and inventor Christiaan Huygens (1629–1695) picked up the new ideas during a visit to Paris in 1655. A digest of these was published, in Dutch, as an appendix to a textbook by a former mathematics teacher of Huygens, Frans van Schooten. This was the first treatise on probability that ever appeared, in Latin in 1657, and in Dutch in 1660. Its importance is in the game-theoretic foundation that Huygens proposes for probability, to support the technical results of Fermat and Pascal.

Huygens starts his essay on how to calculate what non-finished hazard games are worth and how to calculate winning chances in such games as follows:

> "Ick neeme tot beyder fondament, dat in het speelen de kansse, die yemant ergens toe heeft, even soo veel weerdt is als het geen, het welck hebbende hy weder tot deselfde kansse kan geraecken met rechtmatigh spel, dat is, daer in niemandt verlies geboden werdt. By exempel. So yemandt sonder mijn weeten in d'eene handt 3 schellingen verbergt, en in d'ander 7 schellingen, ende my te kiesen geeft welck van beyde ick begeere te hebben, ick segge dit my even soo veel weerdt te zijn, als of ick 5 schellingen seecker hadde.

> Om dat, als ik 5 schellingen hebbe, ick wederom daer toe kan geraecken, dat ick gelijcke kans sal hebben, om 3 of 7 schellingen te krijgen, en dat met rechtmatigh spel: gelijck hier naer sal betoont werden." [18]

Translation:

> "I take as the foundation of both [calculating what non-finished games are worth, and calculating winning chances] that in playing the chance that someone has in some matter, is worth just as much as the amount that, if he possesses it, will give him the same chances in a fair game, that is a game where no loss is offered to anyone. For instance. Suppose someone without my knowing hides in one hand 3 shillings, and in the other 7 shillings, and he offers me the choice between the two hands. Then I would say that this offer is worth the same as having 5 shillings for sure. Because, if I have 5 shillings, I can wager them in such manner that I have equal chances of getting 3 or 7 shillings, and that in a fair game, as will be explained hereafter."

Huygens explains this transformation to a symmetric game by applying it to his example:

> "Indien ick gelijcke kans heb om 3 te hebben of 7, soo is door dit Voorstel mijn kansse 5 weerdt; ende het is seecker dat ick 5 hebbende weder tot de selfde kansse kan geraecken. Want speelende om de selve tegen een ander die daer 5 tegen set, met beding dat de geene die wint den anderen 3 sal geven; soo is dit rechtmaetig spel, ende het blijckt dat ick gelijcke kans hebbe om 3 te hebben, te weeten, als ick verlies, of 7 indien ick win; want alsdan treck ick 10, daer van ick hem 3 geef."

Translation:

> "If I have equal chances to have 3 or 7, then by my Proposal this chance is worth 5; and it is sure that if I have 5, I will get to the same chance. Because putting 5 at stake against someone who stakes 5 against it, with condition that the one who wins will give the other 3, one has a fair game, and it becomes clear that I have equal chance of getting 3, namely, if I lose, or 7 if I win; because if I win I draw 10, of which I give 3 to him."

Thus, Huygens starts out from the expectation of a single individual in a lottery-like situation. He gives a reconstruction of this in terms of an n-person game, where n is the number of proposed chances, with equal stakes, and symmetric roles. Huygens argues that the value of the stakes equals the expectation. If a stake of value x buys me a ticket for a symmetric game with equal stakes that has the same outcomes as the lottery-like situation that we started out with, then it must be that the game and the lottery are worth the same. The Dutch mathematician Hans Freudenthal, in his review of Huygens' theory of probability, remarks that "Equal Chance" is validly defined as free choice for the player in a symmetric situation [15].

This is remarkably close to the famous Dutch book argument as a foundation of probability, proposed much later by Ramsey [27], de Finetti [14], and

Savage [29]. A Dutch book is a collection of bets (so it is not a book, and why it is called Dutch is unclear) that together represent either a sure win or a sure loss for the person who makes the bets, no matter how the situation turns out.

Take the case of equal chances of getting a and b again. Suppose this is offered to you as a symmetric game, at a price x that is different from $\frac{a+b}{2}$. Let G be the game where you get a if you win and b if you lose. Let G' be the game where you get b if you win and a if you lose. Then the only difference between G and G' is that the roles of the two players are reversed. So we may assume that you can enter into both games for the same price x. Now if $x < \frac{a+b}{2}$, what you should do is invest x in G and x in G', and play the games simultaneously. This costs you $2x$, and it yields $a + b$, so this is a Dutch book in your favour. If $x > \frac{a+b}{2}$, and you are willing to enter G and G' for the price x, then your investment of $2x$ will get you only $a + b$, so you are losing no matter what. There is a Dutch book against you.

Let us be a bit more precise about how Huygens would turn an individual choice situation with m possibilities $s_1, \ldots, s_m$, with revenues given by $L_1, \ldots, L_m$, into a stake distribution game G for m players. The stake x would be the same for every player. The game would match the players with the possibilities. The utility function would be given by: if player i draws s_j then i gets L_j. Obviously, the expectation for each player in this game is $\frac{\sum_{i=1}^{m} L_i}{m}$, so that should be the value of an individual stake. Also, the game is obviously symmetric, for all players have equal chances of getting each of the "prizes" $L_1, \ldots, L_m$.

Now replace the revenues by probabilities. Instead of $L_1, \ldots, L_m$ we have $p_1, \ldots, p_m$ with $\sum_{i=1}^{m} p_i = 1$. Nothing changes. The expectation in the game is $\frac{1}{m}$, so this should be the value of an individual stake. Anyone who can get a stake in the game for less that $\frac{1}{m}$ can set up a Dutch book, and anyone who is willing to enter the game for more than $\frac{1}{m}$ faces a Dutch book against him.

3 Belief and decision making

The following is a model for decision making under uncertainty that is widely used. An agent faces a choice between a finite number of possible courses of action, say $a_1, \ldots, a_n$. The agent is uncertain about the state of the world. Say she considers states $s_1, \ldots, s_m$ possible. Now suppose there is a table of consequences c, with $c(s_i, a_j)$ giving the consequences of performing action a_j in state s_i. How can the agent choose between the available actions in a rational way?

In the first place we should model the preferences of the agent. Let us suppose there is a preference ordering R on the consequences, with cRc' expressing that either the agent is indifferent between c and c', or the agent strictly prefers c to c'. Assume R is transitive and reflexive. Then define cPc' as $cRc' \wedge \neg c'Rc$, so that cPc' expresses that the agent strictly prefers c to c'. The relation P is transitive and irreflexive.

A utility function $u : C \to \mathbb{R}$ is said to *represent* R if u satisfies $u(c) \geq u(c')$ iff cRc'.

Von Neumann and Morgenstern [25] showed how to turn this into a tool for decision making if one adds a probability measure P on the state set. So assume $P(s_i) \geq 0$ and $\sum_{i=1}^{m} P(s_i) = 1$. Then a utility function u on the consequences induces a utility function U on the actions, by means of

$$U(a_j) = \sum_{i=1}^{m} P(s_i)u(s_i, a_j).$$

Now it is clear how a rational agent who disposes of (i) a utility function u representing her preferences and (ii) a probability measure on what she thinks is possible decides on what to do. Such an agent will perform the action a_j that maximizes $U(a_j)$.

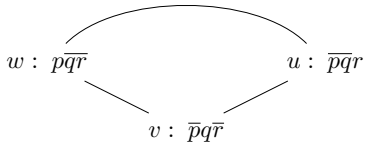
This is the reason why expositions of probability theory often make strong claims about the applicability of their subject. Blitzstein and Hwang [7] list a number of possible applications of probability, and they close off with the application to Life in general:

> "Life is uncertain, and probability is the logic of uncertainty. While it isn't practical to carry out a formal probability calculation for every decision made in life, thinking hard about probability can help us avert some common fallacies, shed light on coincidences, and make better predictions."

This cheerful attitude to decision making engenders a particularly straightforward view of belief. I believe in φ if the odds in favour of φ are larger than $1 : 1$. Odds in favour of φ are calculated by means of $\frac{P(\varphi)}{P(\neg\varphi)}$. So I believe in φ if the subjective probability I assign to the truth of φ is larger than the subjective probability I assign to the truth of $\neg\varphi$. This is in fact the straightforward view that you should only believe propositions which have a probability greater than one half. Call this notion of belief betting belief.

4 Betting belief

The notion of betting belief has a number of remarkable properties. It is not closed under conjunction: it does not follow from the facts that $P(\varphi) > P(\neg\varphi)$ and $P(\psi) > P(\neg\psi)$ that $P(\varphi \wedge \psi) > P(\neg\varphi \vee \neg\psi)$. For suppose p, q, r are three propositions that are mutually exclusive and have the same probability. Then $P(p \vee q) > P(\neg p \wedge \neg q)$ and $P(q \vee r) > P(\neg q \wedge \neg r)$. From the fact that p, q, r are mutually exclusive it follows that $(p \vee q) \wedge (q \vee r)$ is equivalent to q. On the other hand, $P((p \vee q) \wedge (q \vee r)) = P(q) < P(\neg q)$. The following model gives a picture of this situation. The propositions p, q, r are mutually exclusive and have the same probability $\frac{1}{3}$. It is left to the reader to check in the picture that $P(p \vee q) = \frac{2}{3}$, $P(\neg p \wedge \neg q) = \frac{1}{3}$, $P(q \vee r) = \frac{2}{3}$, $P(\neg q \wedge \neg r) = \frac{1}{3}$.

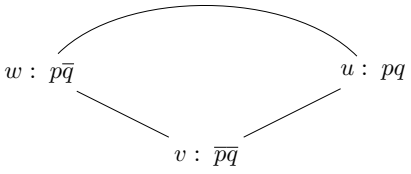


$$P(w) = P(v) = P(u) = \tfrac{1}{3}$$

This model represents probability by means of a weight function that gives each world the same weight. Note that the model also picture knowledge, which is represented by the epistemic accessibility relation.

The solid lines represent the epistemic accessibility relation of a single agent; they indicate that every world is accessible from any world. We will assume throughout this paper that knowledge accessibility is an equivalence; in other words, we are interpreting the knowledge operator K as an S5 operator. In the situation pictured above, the agent knows for instance that at least one of p, q, r is true. This is expressed by $K(p \vee q \vee r)$. The agent also knows that the propositions p, q, r are mutually exclusive. And so on.

Betting belief in φ and betting belief in $\varphi \to \psi$ does not entail betting belief in ψ. This is illustrated by the following model.



$$P(w) = P(v) = P(u) = \tfrac{1}{3}$$

Again, probability is represented by means of a weight function that gives each world the same weight. The probability of p (true in w and u) is $\frac{2}{3}$, the probability of $p \to q$ (true in v and u) is $\frac{2}{3}$, but the probability of q (true in u) is $\frac{1}{3}$. Thus, betting belief in p and $p \to q$ is justified, but betting belief in q is not.

On the other hand, betting belief in $p \wedge q$ implies betting belief in p and in q, for if the probability of $p \wedge q$ is greater than one half, then the same must hold for the probabilities of p and of q.

It is well known that people untrained in probability theory have difficulty with the notion of betting belief. Recall examples like the following. You are from a population with a statistical chance of 1 in 100 of having disease D. The initial screening test for this has a false positive rate of 0.2 and a false negative rate of 0.1. You tested positive; call this test result T. Should you believe you have the disease, with 'believe' in the sense of betting belief?

You reason: "If I test positive then, given that the test is quite reliable, the probability that I have D is quite high." So you tend to believe that you have D. But now you recall a lesson from your probability class: "True positives are

often dwarfed by false positives." You pick up pen and paper and calculate:

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\neg D)P(\neg D)}.$$

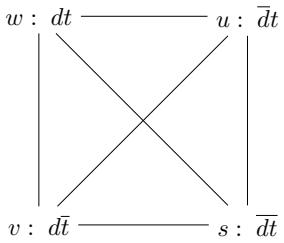
The first step uses Bayes' rule, and the second step calculates $P(T)$ by means of the rule of total probability. Filling in $P(T|D) = 0.9, P(D) = 0.01, P(\neg D) = 0.99, P(T|\neg D) = 0.2$, you arrive at the conclusion that $P(D|T) = \frac{1}{23}$.

As a result of your calculation, you don't believe anymore you have D but you agree to further testing. This step from an initial guess that the probability of D is high to a careful calculation revealing that the probability of D is low should perhaps be viewed as a switch from thinking fast to thinking slow, in the sense of Kahneman [21].

In any case, the example shows that qualitative belief judgements can be utterly misleading. Such examples made probability theorists like Richard Jeffrey urge us to give up qualitative belief altogether in favour of quantitative belief based on probability calculations [19,20]. In the rest of the paper I will show that there is room for qualitative belief linked to probability but not derived from it, after all.

The notion of betting belief introduced above can also be dubbed Bayesian belief. It is natural to interpret the uncertainties that we face in everyday life as subjective probabilities, and recalculations of betting belief based on new information can be viewed as model restrictions. The announcement φ in a model $\mathcal{M}$ leads to a new model $\mathcal{M}|\varphi$ consisting of all worlds in the old model that satisfy φ.

Consider the disease example again. Here is an epistemic probability model for it. The worlds are all connected, so this is a so-called S5 model. A weight function L gives the information about the probabilities for the four possible combinations of $d, \overline{d}$ with $t, \bar{t}$.



$$L(w) = 0.009, L(v) = 0.001, L(u) = 0.198, L(s) = 0.792$$

The weights (or probabilities, for the weight function is normalized) were computed by taking the prior probabilities for d, and multiplying with the appropriate error rates for the test. E.g., $L(u)$ is the product of $\frac{99}{100}$ (the prior probability of not having the disease) and $\frac{1}{5}$ (the false positive rate).

An update with the information t changes this model into the following restricted model, where the worlds where t is false have dropped out. This is

the public announcement update from Jan Plaza [26].

$$w:\ dt \quad\text{———}\quad u:\ \overline{dt}$$

$$L(w) = 0.009, L(u) = 0.198$$

Re-normalization of the weight function gives $L(w) = \frac{1}{23}, L(u) = \frac{22}{23}$. So after the information that the test was positive has been taken into account, the probability of d has changed from $\frac{1}{100}$ to $\frac{1}{23}$. The announcement update result agrees with the application of Bayes' rule. Bayesian conditionalisation (see [32]) and announcement update for epistemic probability models coincide. For further discussion and some qualification of this claim see [5].

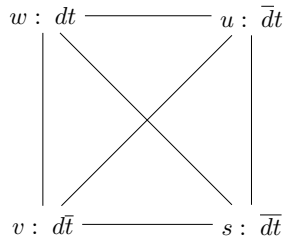
Now let us coarsen the model, and replace the weight function by a neighbourhood function that tells us which propositions are believed in the betting sense. Starting out from epistemic models (Kripke models with equivalence relations of epistemic accessibility), we add a neighbourhood function for each epistemic agent. I will assume that within each i-cell, the neighbourhoods that get assigned to different worlds are the same; this encodes the fact that if an agent believes φ then she knows that she believes φ.

Truth definition for belief in φ, in terms of neighbourhoods, is:

$$M, w \models B\varphi \text{ iff for some } X \in N(w) \text{ for all } x \in X : M, x \models \varphi.$$

Here N is a function that assigns to each world w a set of neighbourhoods for w, where each neighbourhood X is a set of worlds. See [9] for a detailed comparison of neighbourhood models and epistemic probability models. Epistemic probability models are epistemic models with a weight function that assigns positive values to all worlds, and that satisfies the condition that the sum of the weights over each epistemic partition cell is bounded (but this condition is only relevant if the number of worlds in some partition cells is infinite).

Here is a neighbourhood version of the above epistemic weight model, with the neighbourhoods defined from the probabilities by means of: $X \in N(w)$ iff $X \subseteq [w]$ and $P(X) > P([w] - X)$, where $[w]$ is the epistemic equivalence class of w.



$$N(w) = N(v) = N(u) = N(s) = \{\{s\}, \{s,u\}, \{s,v\}, \{s,w\}, \{s,u,v\}, \{s,v,w\}, \{s,w,u\}, \{s,u,v,w\}\}.$$

To understand the neighbourhood function, observe first of all that since the epistemic accessibility relation is universal, the neighbourhoods are the same for every world. Next, note that that X is a neighbourhood iff $s \in X$. This is because the probability of world s in the original probability model is higher than the probability of $W - \{s\}$. It is convenient to use $\uparrow X$ for $\{Y \subseteq U \mid X \subseteq Y\}$ (the set of all supersets of X in domain U), where the domain U is understood from context. Then the neighbourhoods in the model are given by $N(w) = N(v) = N(u) = N(s) = \uparrow \{s\}$.

Now we can see that the neighbourhood function does not give enough information to calculate a new neighbourhood after information update. After information update with t, betting belief should favour world u over world w. But no reasonable update rule on neighbourhoods will give this result, for in the original model, the neighbourhood function is symmetric between w and u: we have for all neighbourhoods X that $w \in X$ iff $u \in X$.

This indicates that instead of a neighbourhood function we need something more expressive. One option here would be to introduce plausibility relations [3,4], and no doubt there are other options. The option we will explore here is modification of the neighbourhood function.

A *conditioned neighbourhood functional* is a functional $\mathfrak{N} : W \to \mathcal{P}(W) \to \mathcal{P}\mathcal{P}(W)$ that assigns to every w a function $\mathfrak{N}_w : \mathcal{P}(W) \to \mathcal{P}\mathcal{P}(W)$, where for each $X \subseteq W$, $\mathfrak{N}_w(X)$ is a set of neighbourhoods of w conditioned by X.

A neighbourhood functional for the disease model would assign to every world w and every $X \subseteq W$ a set of neighbourhoods given by

$$\mathfrak{N}_w(X) = \{Y \subseteq X \mid P(Y) > P(X - Y)\}.$$

For the disease model, we get the following neighbourhood functional (values indicated for all sets with size > 1):

$$\begin{array}{ll}
\{s,u,v,w\} & \mapsto \uparrow \{s\} \\
\{s,u,v\} & \mapsto \uparrow \{s\} \\
\{s,u,w\} & \mapsto \uparrow \{s\} \\
\{s,v,w\} & \mapsto \uparrow \{s\} \\
\{u,v,w\} & \mapsto \uparrow \{u\} \\
\{s,u\} & \mapsto \uparrow \{s\} \\
\{s,v\} & \mapsto \uparrow \{s\} \\
\{s,w\} & \mapsto \uparrow \{s\} \\
\{u,v\} & \mapsto \uparrow \{u\} \\
\{u,w\} & \mapsto \uparrow \{u\} \\
\{v,w\} & \mapsto \uparrow \{w\}
\end{array}$$

Truth definition for belief in φ, in terms of neighbourhood functionals is (assume $[w]$ gives the partition block of w for the epistemic relation):

$$M, w \models B\varphi \text{ iff for some } X \in \mathfrak{N}_w([w]) \text{ for all } x \in X : M, x \models \varphi.$$

A reasonable update rule for neighbourhood functionals could now be: restrict the functional to the new universe U.

Using this, we see that after update of the neighbourhood functional of the neighbourhood model with t, the agent still believes that $\neg d$, as she should.

One of the properties of betting belief is *strong commitment*. To see what that means, let us first look at the dual $\widehat{B}$ of B. $\widehat{B}\varphi$ is true iff $\neg B\neg\varphi$ is true iff it is not the case that the probability of $\neg\varphi$ is higher than $\frac{1}{2}$. This is the case iff the probability of $\neg\varphi \leq \frac{1}{2}$, iff the probability of φ is $\geq \frac{1}{2}$.

Now suppose $\widehat{B}\varphi$ is true. Then $P(\varphi) \geq \frac{1}{2}$. Suppose $\widehat{K}(\neg\varphi \wedge \psi)$ is also true. Then an accessible world where φ is false and ψ true exists. Let us look at the probability of $\varphi \vee \psi$. It must be strictly larger than $\frac{1}{2}$, for the world where φ is false and ψ true has positive weight. I have just shown the soundness of the following axiom of strong commitment (SC):

$$\widehat{B}\varphi \wedge \widehat{K}(\neg\varphi \wedge \psi) \rightarrow B(\varphi \vee \psi). \tag{SC}$$

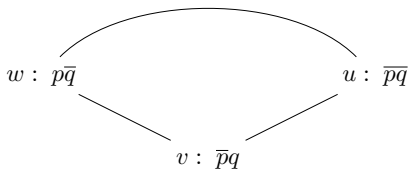
Another axiom that we get immediately from the meaning of $\widehat{B}\varphi$ is (D) for determinacy:

$$B\varphi \rightarrow \widehat{B}\varphi. \tag{D}$$

What (D) says is that it follows from that fact that I am willing to bet on φ that I am not willing to bet on $\neg\varphi$.

If we replace the notion of betting belief by that of threshold belief, by interpreting belief in φ as $P(\varphi) > t$, for some specific t with $\frac{1}{2} \leq t < 1$ (this is also known as Lockean belief), then $\widehat{B}\varphi$ gets a different meaning. Under this notion of belief, $\widehat{B}\varphi$ is true iff it is not the case that $B\neg\varphi$ is true, iff it is not the case that $P(\neg\varphi) > t$, iff $P(\neg\varphi) \leq t$, iff $P(\varphi) \geq 1 - t$. Since $\frac{1}{2} \leq t$, this certainly holds. It follows that (D) also is sound for Lockean belief.

For threshold belief with $t > \frac{1}{2}$, (SC) fails, however. This is illustrated by the following counterexample.



$$P(w) = 1 - t, P(v) = t - \tfrac{1}{2}, P(u) = \tfrac{1}{2}$$

Let $t > \frac{1}{2}$. Then $P(p) = P(w) = 1 - t$, so, as we have seen, $\widehat{B}p$ is true. Also, $\widehat{K}(\neg p \wedge q)$ is true, for there is an accessible world, v, where $\neg p \wedge q$ is true. The formula $p \vee q$ is true in worlds w and v, so $P(p \vee q) = P(w) + P(v) = 1 - t + t + \frac{1}{2} = \frac{1}{2} < t$, so $B(p \vee q)$ is false in the model.

Still another way to interpret (qualitative) belief is as follows: $S\varphi$ is true iff it holds for all consistent ψ that $P(\varphi|\psi) > P(\neg\varphi|\psi)$ (compare Leitgeb [24]). This uses $S\varphi$ for stable belief in φ. Stable belief can also be defined in terms of updates: $S\varphi$ is true in w iff it holds for all ψ that are true in w that $[!\psi]B\varphi$,

where $[!\psi]B\varphi$ expresses that $B\varphi$ is true after updating the model with ψ, and where $B\varphi$ is interpreted as betting belief.

Since neighbourhood models are not expressive enough to model betting belief update, neighbourhood models cannot provide a reasonable truth definition for $S\varphi$. But if we switch to conditioned neighbourhood models, we have a means to interpret stable belief, as follows.

$$\begin{aligned} M, w \models S\varphi \text{ iff } & \text{ for all } X \subseteq [w], X \neq \emptyset \\ & \text{it holds for all } x \in X \text{ that } M \mid X, x \models B\varphi. \end{aligned}$$

Here $M \mid X$ is model M restricted to X, with the neighbourhood functional restricted accordingly, so $B\varphi$ is interpreted with respect to the updated neighbourhood functional. The clause for $S\varphi$ expresses that (stable) belief in φ is belief that continues to hold, no matter how we restrict the model.

In fact, Leitgeb's notion is a special case of this, for Leitgeb's theory is phrased in terms of standard Kripke models instead of neighbourhood models, and standard Kripke models can be viewed as constrained neighbourhood models.

Strong belief in φ, yet another notion of qualitive belief, is a bit harder to link to probability. It is defined for plausibility models, e.g., locally connected preorders. A preorder is a reflexive and transitive relation. A relation R is weakly connected (terminology of Robert Goldblatt [16]) if the following holds:

$$\forall x, y, z((xRy \wedge xRz) \rightarrow (yRz \vee y = z \vee zRy)).$$

A relation R is locally connected if both R and $R^{\frown}$ (the converse of R) are weakly connected. A most plausible possible world is a world that is maximal in the R ordering. An agent strongly believes in φ if φ is true in all most plausible accessible worlds. This yields a KD45 notion of belief (reflexive, euclidean, and serial). See Baltag & Smets [3,4].

Finally, it is possible to interpret qualitative belief as subjective certainty. An agent i believes in φ without any doubt if $P_i(\varphi) = 1$. This is used in epistemic game theory (Aumann [1]), and can easily be expressed in epistemic models, for this notion coincides with knowledge. If one drops the requirement that weight functions assign strictly positive values to all worlds then certainty and knowledge no longer coincide.

5 The lottery puzzle

One of attractions of betting belief lies in the light it sheds on the lottery puzzle. If Alice believes of each of the tickets 000001 through 111111 that they are not winning, then this situation is described by the following formula:

$$\bigwedge_{t=000001}^{111111} B_a \neg t.$$

If her beliefs are closed under conjunction, then this follows:

$$B_a \bigwedge_{t=000001}^{111111} \neg t.$$

But actually, she believes, of course, that one of the tickets is winning:

$$B_a \bigvee_{t=000001}^{111111} t.$$

This is a contradiction. Since the lottery puzzle involves three statements, there are three possible strategies to deal with it.

(i) Deny that Alice believes that her ticket is not winning.

(ii) Block the inference from $\bigwedge_{t=000001}^{111111} B_a \neg t$ to $B_a \bigwedge_{t=000001}^{111111} \neg t$.

(iii) Deny that Alice believes that there is a winning ticket.

A notion of belief for which it holds that Alice does not believe there is a winning ticket will hardly convince anyone, so let us forget about that way out. This leaves us with two options.

The advantage of (i) is that there is no need to sacrifice closure of belief under conjunction. A disadvantage is that one has to opt for a severe restriction of what counts as belief.

An advantage of (ii): no need to artificially restrict what counts as belief. And true, one has to sacrifice closure of belief under conjunction, but this is maybe not so bad after all. As I will see below, lots of nice logical properties remain.

Proponents of (i) are many philosophers, and they are easy to recognize: they call the lottery puzzle *the lottery paradox*. But maybe this is a bit harsh on the philosophers; after all, some have taken the trouble to develop notions of stable belief where some version of (i) can be saved. Proponents of (ii) are subjective Probabilists like Jeffrey [20], and decision theorists like Kyburg [23]. As we will see in the next section, one can side with them without giving up reasonable notions of qualitative belief.

6 Neighbourhood models and completeness

To drop the closure of belief under conjunction, we need an operator B_a that does *not* satisfy (Dist).

$$B_a(\varphi \to \psi) \to B_a\varphi \to B_a\psi \qquad \text{(Dist-B)}$$

This means: B_a is not a *normal* modal operator. See also [34]. Interpreting modal operators as accessibility relations between worlds brings the distribution axiom or K axiom in its wake. In order to drop it we have to switch to (epistemic) neighbourhood models. Here is a formal definition.

An **Epistemic Doxastic Neighbourhood Model** $\mathcal{M}$ for set of agents *Ag* and set of propositions *Prop* is a tuple

$$(W, R, V, N)$$

where

- W is a non-empty set of worlds.
- R is a function that assigns to every agent $a \in Ag$ an equivalence relation $\sim_a$ on W. We use $[w]_a$ for the $\sim_a$ class of w, i.e., for the set $\{v \in W \mid w \sim_a v\}$.
- V is a valuation function that assigns to every $w \in W$ a subset of *Prop*.
- N is a function that assigns to every agent $a \in Ag$ and world $w \in W$ a collection $N_a(w)$ of sets of worlds—each such set called a *neighbourhood* of w—subject to a set of *conditions*.

The core conditions are as follows:

(c) $\forall X \in N_a(w) : X \subseteq [w]_a$. This ensures that agent a does not believe any propositions $X \subseteq W$ that she knows to be false.

(f) $\emptyset \notin N_a(w)$. This ensures that no logical falsehood is believed.

(n) $[w]_a \in N_a(w)$. This ensures that what is known is also believed.

(a) $\forall v \in [w]_a : N_A(v) = N_A(w)$. This ensures that if X is believed, then it is known that X is believed.

By dropping some of these conditions one can further weaken (or: generalize, depending on perspective) the notion of belief. But the constraints that the conditions impose on belief are quite weak, so we will not do so here.

There are three further conditions that may be imposed to further strengthen the notion of belief.

(m) $\forall X \subseteq Y \subseteq [w]_A :$ if $X \in N_A(w)$, then $Y \in N_A(w)$. This says that belief is monotonic: if an agent believes X, then she believes all propositions $Y \supseteq X$ that follow from X. This may seem entirely reasonable, but in proposals where neighbourhoods are used to model conflicting and inconclusive evidence [6] it is dropped.

(d) If $X \in N_a(w)$ then $[w]_a - X \notin N_a(w)$. This corresponds to the axiom (D) that we discussed above. This condition says that if a believes a proposition X then a does not believe the negation of that proposition. As we have seen, this holds for betting belief and threshold belief, for a threshold above $\frac{1}{2}$. For thresholds below $\frac{1}{2}$, it fails, however.

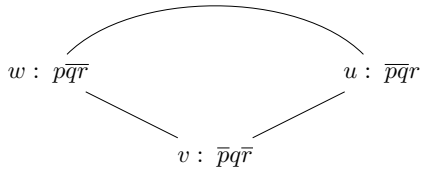
(sc) $\forall X, Y \subseteq [w]_a$: if $[w]_a - X \notin N_a(w)$ and $X \subsetneq Y$, then $Y \in N_a(w)$. If the agent does not believe the complement $[w]_a - X$, then she must believe any strictly weaker Y implied by X. We saw above that this distinguishes betting belief from threshold beliefs for thresholds above $\frac{1}{2}$.

Epistemic doxastic neighbourhood models can interpret the language of epis-

temic doxastic logic (henceforth, KB language):

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi \mid B_a\varphi.$$

The interpretation of $K_a\varphi$ uses the R relations; the interpretation of $B_a\varphi$ uses the neighbourhoods. Here is the neighbourhood version of the first example above:



$$N(w) = N(v) = N(u) = \{\{w, v\}, \{v, u\}, \{w, u\}, \{w, v, u\}\}$$

In all worlds, $K(p \vee q \vee r)$ is true. In all worlds $B\neg p$, $B\neg q$, $B\neg r$ are true. In all worlds $B(\neg p \wedge \neg q)$, $B(\neg p \wedge \neg r)$, $B(\neg q \wedge \neg r)$ are false. So the lottery puzzle is solved in neighbourhood models for belief by non-closure of belief under conjunction.

Here is a calculus for betting belief that relates belief to a standard S5 notion of knowledge.

AXIOMS

(Taut)	All instances of propositional tautologies
(Dist-K)	$K_a(\varphi \to \psi) \to K_a\varphi \to K_a\psi$
(T)	$K_a\varphi \to \varphi$
(PI-K)	$K_a\varphi \to K_aK_a\varphi$
(NI-K)	$\neg K_a\varphi \to K_a\neg K_a\varphi$
(F)	$\neg B_a\bot$.
(PI-KB)	$B_a\varphi \to K_aB_a\varphi$
(NI-KB)	$\neg B_a\varphi \to K_a\neg B_a\varphi$
(KB)	$K_a\varphi \to B_a\varphi$
(M)	$K_a(\varphi \to \psi) \to B_a\varphi \to B_a\psi$
(D)	$B_a\varphi \to \neg B_a\neg\varphi$.
(SC)	$\widehat{B}_a\varphi \wedge \widehat{K}_a(\neg\varphi \wedge \psi) \to B_a(\varphi \vee \psi)$

RULES

$$\frac{\varphi \to \psi \quad \varphi}{\psi}\ \text{(MP)} \qquad \frac{\varphi}{K_a\varphi}\ \text{(Nec-K)}$$

This calculus for betting belief is discussed in [9] and [2]. The fact that closes off this section is proved in [9].

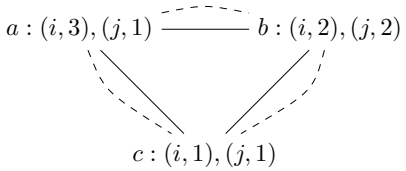
Fact 6.1 *The calculus of betting belief is complete for epistemic doxastic neighbourhood models.*

7 Incompleteness for epistemic probability models

The step from neighbourhoods to probabilities is very small, but we will see in this section that the logic of neighbourhoods and the logic of probabilities are different.

Epistemic probability models are the result of replacing the neighbourhood function of an epistemic doxastic neighbourhood model by a weight function L. A weight function L assigns to every agent a a function $L_a : W \to \mathbb{Q}^+$ (the positive rationals), subject to the constraint that the sum of the L_a values over each epistemic partition cell of a is bounded. If $X \subseteq W$ then let $L_a(X)$ be shorthand for $\sum_{x \in X} L_a(x)$. Boundedness can then be expressed as follows: for each i and w: $L_a([w]_a) < \infty$.

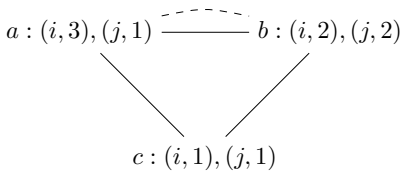
To illustrate, here is an example from investment banking. Two bankers i, j consider buying stocks in three firms a, b, c that are involved in a takeover bid. There are three possible outcomes: a for "a wins", b for " b wins", and c for "c wins." i takes the winning chances to be $3 : 2 : 1$, j takes them to be $1 : 2 : 1$. In the following picture, the knowledge of i is represented by solid lines, that of j by dashed lines.



In all worlds, i assigns probability $\frac{1}{2}$ to a, $\frac{1}{3}$ to b and $\frac{1}{6}$ to c, while j assigns probability $\frac{1}{4}$ to a and to c, and probability $\frac{1}{2}$ to b.

We see that i is willing to bet $1 : 1$ on a, while j is willing to bet $3 : 1$ against a. It follows that in this model i and j have an opportunity to gamble, for, to put it in Bayesian jargon, they do not have a common prior.

Now consider the possibility that agent j has learnt something. Suppose that, as a result of this information, agent j (dashed lines) now considers c impossible.



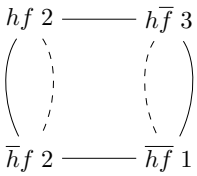
So we suppose that j has foreknowledge about what firm c will do.

The probabilities assigned by i remain as before. The probabilities assigned by j have changed, as follows. In worlds a and b, j assigns probability $\frac{1}{3}$ to a

and $\frac{2}{3}$ to b. In world c, j is sure of c.

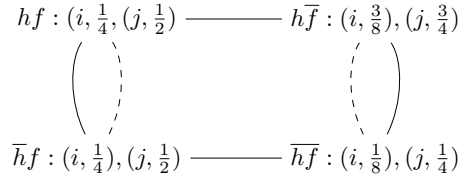
We may suppose that this new model results from j being informed about the truth value of c, while i is aware that j received this information, but without i getting the information herself. So i is aware that j's subjective probabilities have changed, and it would be unwise for i to put her beliefs to the betting test. For although i cannot distinguish the three situations, she knows that j can distinguish the c situation from the other two. Willingness of j to bet against a at any odds can be interpreted by i as an indication that c is true, thus forging an intimate link between action and information update. I leave further analysis for another occasion.

Here is an example where two agents i (solid lines) and j (dashed lines) are uncertain about the toss of a coin. i holds it possible that the coin is fair f and that it is biased $\overline{f}$, with a bias $\frac{2}{3}$ for heads h. j can distinguish f from $\overline{f}$. The two agents share the same weight (so this is a single weight model, see [10]), and the weight values are indicated as numbers in the picture.



In world hf, i assigns probability $\frac{5}{8}$ to h and probability $\frac{1}{2}$ to f, and j assigns probability $\frac{1}{2}$ to h and probability 1 to f.

It is possible to normalize this model, but as a result of this each agent will have to get its own weight, for the weight functions are normalized within the epistemic accessibility cells.



The rules for interpretation of the KB language in epistemic probability models are obvious:

$$\mathcal{M}, w \models K_a\varphi \text{ iff } \quad \text{for all } v \in [w]_a : \mathcal{M}, v \models \varphi.$$

$$\mathcal{M}, w \models B_a\varphi \text{ iff}$$

$$\sum\{L_a(v) \mid v \in [w]_a, \mathcal{M}, v \models \varphi\} > \sum\{L_a(v) \mid v \in [w]_a, \mathcal{M}, v \models \neg\varphi\}.$$

There is also an obvious way to reduce an epistemic probability model to a neighbourhood model, while preserving betting belief. Let $\mathcal{M} = (W, R, V, N)$

be a neighbourhood model and let L be a weigth function for $\mathcal{M}$. Then L *agrees with* $\mathcal{M}$ if it holds for all agents a and all $w \in W$ that

$$X \in N_a(w) \text{ iff } L_a(X) > L_a([w]_a - X).$$

The following theorem may come as a surprise, for it shows that, in a sense, the class of epistemic doxastic neighbourhood models is more general than that of probabilistic models. In other words: the principles of betting belief given in the calculus above do not force a probabilistic interpretation of the B operator.

Theorem 7.1 *There exists an epistemic doxastic neighbourhood model $\mathcal{M}$ that has no agreeing weight function.*

Proof. The proof of this uses an adaptation of an example from [33, pp. 344-345]. Let *Prop* $:= \{a, b, c, d, e, f, g\}$. Assume a single agent 0. Define:

$$\mathcal{X} := \{efg, abg, adf, bde, ace, cdg, bcf\}.$$

$$\mathcal{X}' := \{abcd, cdef, bceg, acfg, bdfg, abef, adeg\}.$$

Notation: xyz for $\{x, y, z\}$.

$$\mathcal{Y} := \{Y \mid \exists X \in \mathcal{X} : X \leq Y \leq W\}.$$

Let $\mathcal{M} := (W, R, V, N)$ be defined by $W :=$ *Prop*, $R_0 = W \times W$, $V(w) = \{w\}$, and for all $w \in W$, $N_0(w) = \mathcal{Y}$. Check that $\mathcal{X}' \cap \mathcal{Y} = \emptyset$. So $\mathcal{M}$ is a neighbourhood model.

Toward a contradiction, suppose there exists a weight function L that agrees with $\mathcal{M}$. Since each letter $p \in W$ occurs in exactly three of the seven members of $\mathcal{X}$, we have:

$$\sum_{X \in \mathcal{X}} L_0(X) = \sum_{p \in W} 3 \cdot L_0(\{p\}).$$

Since each letter $p \in W$ occurs in exactly four of the seven members of $\mathcal{X}'$, we have:

$$\sum_{X \in \mathcal{X}'} L_0(X) = \sum_{p \in W} 4 \cdot L_0(\{p\}).$$

On the other hand, from the fact that $L_0(X) > L_o(W - X)$ for all members X of $\mathcal{X}$ we get:

$$\sum_{X \in \mathcal{X}} L_0(X) > \sum_{X \in \mathcal{X}} L_0(W - X) = \sum_{X \in \mathcal{X}'} L_0(X).$$

Contradiction. So no such L_0 exists. □

I conjecture that this is the smallest counterexample, that is, I guess that all neighbourhood models up to size 6 have an agreeing weight function, but this needs to be checked.

Fact 7.2 *The calculus of epistemic-doxastic neighbourhood logic is sound for the class of epistemic probability models. Probabilistic beliefs are neighbourhoods.*

Theorem 7.1 shows that the KB calculus is incomplete for the class of epistemic probability models. In order to get a calculus that fits this class, we have to add an infinite series of axioms. The idea behind these axioms is from Scott [30]. What the axioms say, intuitively: If agent a knows the number of true φ_i is less than or equal to the number of true ψ_i, and if a believes φ_1, and the remaining φ_i are each consistent with her beliefs, then agent a believes one of the ψ_i.

It turns out that this is expressible in the KB language; see Segerberg [31]. Let $(\varphi_1, \ldots, \varphi_m \mathbb{I}_a \psi_1, \ldots, \psi_m)$ abbreviate the KB formula expressing that agent a knows that the number of true φ_i is less than or equal to the number of true ψ_i. Put another way, $(\varphi_i \mathbb{I}_a \psi_i)_{i=1}^m$ is true if and only if every one of a's epistemically accessible worlds satisfies at least as many ψ_i as φ_i. Using this, we can express the Scott axioms:

(Scott) $\quad [(\varphi_i \mathbb{I}_a \psi_i)_{i=1}^m \wedge B_a \varphi_1 \wedge \bigwedge_{i=2}^m \widehat{B}_a \varphi_i] \rightarrow \bigvee_{i=1}^m B_a \psi_i$

Theorem 7.3 *Adding the Scott axioms to the KB calculus yields a system that is sound and complete for epistemic probability models.*

For the proof of this I refer to [9]. To say a bit more about the connection between qualitative belief and quantitative belief we need a more expressive language for interpretation in epistemic probability models.

Let i range over *Ag*, p over *Prop*, and q over $\mathbb{Q}$. Then the language of epistemic probability logic is given by:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid t_a \geq 0 \mid t_a = 0$$

$$t_a ::= q \mid q \cdot P_a\varphi \mid t_a + t_a \text{ where all indices } a \text{ are the same.}$$

This is expressive enough to compare subjective probabilities of the same agent. In particular, we can say things like $P_a(\varphi) > P_a(\psi)$. Truth for this language in epistemic probability models is defined as follows. Let $\mathcal{M} = (W, R, V, L)$ be an epistemic weight model and let $w \in W$.

$$\begin{aligned}
\mathcal{M}, w &\models \top \quad \text{always}\\
\mathcal{M}, w &\models p \text{ iff } p \in V(w)\\
\mathcal{M}, w &\models \neg\varphi \text{ iff it is not the case that } \mathcal{M}, w \models \varphi\\
\mathcal{M}, w &\models \varphi_1 \wedge \varphi_2 \text{ iff } \mathcal{M}, w \models \varphi_1 \text{ and } \mathcal{M}, w \models \varphi_2\\
\mathcal{M}, w &\models t_a \geq 0 \text{ iff } [\![t_a]\!]_w^{\mathcal{M}} \geq 0\\
\mathcal{M}, w &\models t_a = 0 \text{ iff } [\![t_a]\!]_w^{\mathcal{M}} = 0.
\end{aligned}$$

$$\begin{aligned}
[\![q]\!]_w^{\mathcal{M}} &:= q\\
[\![q \cdot P_a\varphi]\!]_w^{\mathcal{M}} &:= q \times P_{a,w}^{\mathcal{M}}(\varphi)\\
[\![t_a + t'_a]\!]_w^{\mathcal{M}} &:= [\![t_a]\!]_w^{\mathcal{M}} + [\![t'_a]\!]_w^{\mathcal{M}}
\end{aligned}$$

$$P_{a,w}^{\mathcal{M}}(\varphi) = \frac{L_a(\{u \in [w]_a \mid \mathcal{M}, u \models \varphi\})}{L_a([w]_a)}.$$

Fact 7.4 *A sound and complete calculus for the language of epistemic probability logic, interpreted in epistemic probability models, is given in [10].*

See also [13] and [22], where calculi for different epistemic probability model classes are given.

Notice that every epistemic probability model has an associated neighbourhood model. For if $\mathcal{M} = (W, R, V, L)$ is an epistemic probability model, then let $\mathcal{M}^\bullet$ be the tuple (W, R, V, N) given by replacing the weight function by a function N, where N is defined as follows, for $a \in Ag$, $w \in W$.

$$N_a(w) = \{X \subseteq [w]_a \mid L_a(X) > L_a([w]_a - X)\}.$$

Fact 7.5 *For any epistemic weight model $\mathcal{M}$ it holds that $\mathcal{M}^\bullet$ is a neighbourhood model.*

Now let us translate knowledge and belief into probability statements, by interpreting knowledge as certainty and belief as betting belief.

If φ is a KB formula, then $\varphi^\bullet$ is the formula of the language of epistemic probability logic given by the following instructions:

$$\begin{aligned}
\top^\bullet &= \top \\
p^\bullet &= p \\
(\neg\varphi)^\bullet &= \neg\varphi^\bullet \\
(\varphi_1 \wedge \varphi_2)^\bullet &= \varphi_1^\bullet \wedge \varphi_2^\bullet \\
(K_a\varphi)^\bullet &= P_a(\varphi^\bullet) = 1 \\
(B_a\varphi)^\bullet &= P_a(\varphi^\bullet) > P_a(\neg\varphi^\bullet).
\end{aligned}$$

Theorem 7.6 *For all KB formulas φ, for all epistemic probability models $\mathcal{M}$, for all worlds w of $\mathcal{M}$:*

$$\mathcal{M}^\bullet, w \models \varphi \textit{ iff } \mathcal{M}, w \models \varphi^\bullet.$$

Proof. Induction on formula structure. □

Theorem 7.7 *Let $\vdash$ denote derivability in the neighbourhood calculus for KB. Let $\vdash'$ denote derivability in the calculus of EPL. Then $\vdash \varphi$ implies $\vdash' \varphi^\bullet$.*

Proof. Induction on proof structure. □

8 Some Loose Ends

Are there applications where neighbourhoods without agreeing weight functions are natural? Is there a natural interpretation for the incompleteness example for $\{a, b, c, d, e, f, g\}$? Is the counterexample against completeness of the KB calculus for probability models the smallest counterexample?

Representation of probabibility information by means of weight functions was designed with implementation of model checking in mind. Just extend epistemic model checkers for S5 logics with a weight table for each agent. Implementations of model checkers for these logics can be found in [8] and in

[28]. The implementations can deal with Monty Hall style puzzles, urn puzzles, Bayesian updating by drawing from urns or tossing (possibly biased) coins, and 'paradoxes' such as the puzzle of the three prisoners (see, e.g., [20]). Efficiency was not a goal, but these implementations can be made quite efficient with a little effort.

Further analysis of the connection between neighbourhood logics and probabilistic logics [9] is in order. This is also connected to work of Wes Holliday and Thomas Icard [17]. Holliday and Icard investigate a language with a primitive operation $\varphi \succsim_a \psi$, for "according to a, φ is at least as probable as ψ." This is a revival of Segerberg's modal logic for comparative probability [31]. Interestingly, the qualitative probability Kripke models defined by Segerberg (and adopted by Holliday and Icard) seem better suited for defining well-behaved model restriction operations than the neighbourhood models used in the present paper. But note that the models with conditional neighbourhood functionals remedy this. Therefore, an obvious next step in the investigation of the logic of knowledge and qualitative belief is the study of the class of epistemic doxastic models with conditional neighbourhood functionals, together with operations of knowledge and belief update.

Acknowledgement

I thank the participants of the Synthese Workshop on Qualitative and Quantitative Methods in Formal Epistemology (Amsterdam, November 2014) for stimulating feedback. Two anonymous reviewers provided helpful comments on an earlier version of this paper.

References

[1] Aumann, R., *Interactive epistemology I: Knowledge*, International Journal of Game Theory **28** (1999), pp. 263–300.
[2] Baltag, A., J. van Benthem, J. van Eijck and S. Smets, *Reasoning about communication and action* (2014), book manuscript, ILLC.
[3] Baltag, A. and S. Smets, *Conditional doxastic models: A qualitative approach to dynamic belief revision*, Electronic Notes in Theoretical Computer Science (ENTCS) **165** (2006), pp. 5–21.
[4] Baltag, A. and S. Smets, *A qualitative theory of dynamic interactive belief revision*, in: G. Bonanno, W. van der Hoek and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, Texts in Logic and Games, Amsterdam University Press, 2008 pp. 11–58.
[5] van Benthem, J., *Conditional probability meets update logic*, Journal of Logic, Language and Information **12** (2003), pp. 409–421.
[6] van Benthem, J. and E. Pacuit, *Dynamic logics of evidence-based beliefs*, Studia Logica **99** (2011), pp. 61–92.
[7] Blitzstein, J. K. and J. Hwang, "Introduction to Probability," CRC Press, 2014.
[8] van Eijck, J., *Learning about probability* (2013), available from `http://homepages.cwi.nl/~jve/software/prodemo`.
[9] van Eijck, J. and B. Renne, *Belief as willingness to bet*, E-print, arXiv.org (2014), arXiv:1412.5090v1 [cs.LO].
[10] van Eijck, J. and F. Schwarzentruber, *Epistemic probability logic simplified*, in: R. Goré, B. Kooi and A. Kurucz, editors, *Advances in Modal Logic, Volume 10*, 2014, pp. 158–177.

[11] van Eijck, J. and R. Verbrugge, editors, "Discourses on Social Software," Texts in Logic and Games **5**, Amsterdam University Press, Amsterdam, 2009.
[12] van Eijck, J. and R. Verbrugge, editors, "Games, Actions, and Social Software," Texts in Logic and Games, LNAI **7010**, Springer Verlag, Berlin, 2012.
[13] Fagin, R. and J. Halpern, *Reasoning about knowledge and probability*, Journal of the ACM (1994), pp. 340–367.
[14] de Finetti, B., *La prevision: ses lois logiques, se sources subjectives*, Annales de l'Institut Henri Poincaré **7** (1937), pp. 1–68, translated into English and reprinted in Kyburg and Smokler, Studies in Subjective Probability (Huntington, NY: Krieger; 1980).
[15] Freudenthal, H., *Huygens' foundations of probability*, Historia Mathematica **7** (1980), pp. 113–117.
[16] Goldblatt, R., "Logics of Time and Computation, Second Edition, Revised and Expanded," CSLI Lecture Notes **7**, CSLI, Stanford, 1992 (first edition 1987), distributed by University of Chicago Press.
[17] Holliday, W. H. and T. F. Icard, *Measure semantics and qualitative semantics for epistemic modals*, in: *Proceedings of SALT*, SALT 23, 2013, pp. 514—-534.
[18] Huygens, C., "Van Rekeningh in Spelen van Geluck," 1660, about Calculation in Hazard Games; Latin: De ratociniis in Ludo aleae.
[19] Jeffrey, R., "The Logic of Decision," University of Chicago Press, 1983, second edition.
[20] Jeffrey, R., "Subjective Probability — The Real Thing," Cambridge University Press, 2004.
[21] Kahneman, D., "Thinking, Fast and Slow," Allen Lane, 2011.
[22] Kooi, B. P., "Knowledge, Chance, and Change," Ph.D. thesis, Groningen University (2003).
[23] Kyburg, H., "Probability and the Logic of Rational Belief," Wesleyan University Press, Middletown, CT, 1961.
[24] Leitgeb, H., *The stability theory of belief*, Philosophical Review **123** (2014), pp. 131–171.
[25] von Neumann, J. and O. Morgenstern, "Theory of Games and Economic Behavior," Princeton University Press, 1944.
[26] Plaza, J. A., *Logics of public communications*, in: M. L. Emrich, M. S. Pfeifer, M. Hadzikadic and Z. W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, 1989, pp. 201–216.
[27] Ramsey, F., *Truth and probability*, in: R. Braithwaite, editor, *The Foundations of Mathematics and Other Essays*, Humanities Press, 1931 .
[28] Santoli, T., *Haskell project epistemic logic*, Technical report, ILLC (Summer 2014).
[29] Savage, L. J., "The Foundations of Statistics — Second Revised Edition," Dover, New York, 1972.
[30] Scott, D., *Measurement structures and linear inequalities*, Journal of Mathematical Psychology **1** (1964), pp. 233–247.
[31] Segerberg, K., *Qualitative probability in a modal setting*, in: J. Fenstad, editor, *Proceedings of the 2nd Scandinavian Logic Symposium* (1971), pp. 341–352.
[32] Talbott, W., *Bayesian epistemology*, in: E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Stanford University, Fall 2013 Edition `http://plato.stanford.edu/archives/fall2013/entries/epistemology-bayesian/`.
[33] Walley, P. and T. Fine, *Varieties of modal (classificatory) and comparative probability*, Synthese **41** (1979), pp. 321–374.
[34] Zvesper, J. A., "Playing with Information," Ph.D. thesis, ILLC, University of Amsterdam (2010).

Evolving Models of Social Cognition

Daniel J. van der Post [1]

Centre for Social Learning and Cognitive Evolution
School of Biology
University of St. Andrews

Elske van der Vaart

School of Biological Sciences
Harborne Building
University of Reading

Abstract

In 2009, Rineke Verbrugge set out a research agenda which proposed to study social cognition using logic and computational modeling. The hope was that two different types of simulation might prove useful, namely cognitive models and agent-based models. Verbrugge's proposal was to use cognitive models to investigate the limits of 'higher order theory of mind' in humans, and to test different 'scenarios' for its evolution using agent-based models, essentially simulating different hypothesized 'evolutionary pressures' and studying the extent to which they select for more complex social cognition. However, many cognitive and agent-based models actually deliver an opposite message, that is often experienced as 'killjoy' – that what looks like complex social cognition might not be either social or complex at all! In this paper, we discuss *why* computational models tend to deliver such 'killjoy' explanations, and what this means for studying the evolution of social cognition in animals. We conclude that such models incorporate embodiment and embeddedness due to various dynamical feedbacks, and so give rise to counter-intuitive dynamics. Through these dynamics, seemingly simple behavior rules can generate seemingly complex behavioral patterns. In principle, such 'killjoy' results enable us to identify 'false-positives', i.e. those cases where we have identified (selection for) social cognition where there is none. However, there is a danger that computational models oversimplify matters leading to false-negatives. In the latter case, we erroneously discard cognitive explanations. We propose that future research should address whether oversimplification and false-negatives are a problem, because this will affect whether we are identifying false-positives appropriately.

Keywords: Social cognition, Computational model, 'Killjoy', Embodied, Embedded

[1] Email: d.j.vanderpost@gmail.com; e.e.vandervaart@reading.ac.uk (Note: authors are included in alphabetical order.)

1 Introduction

Broadly speaking, social cognition is the suite of skills animals of all kinds use to navigate life with their conspecifics. It is a widely studied topic; comparative researchers study social cognition in other animals [25]; developmental psychologists investigate its emergence in children [60]; neuro-scientists image its workings in the brain [46].

In 2009, Rineke Verbrugge set out a research agenda which proposed to study social cognition from another angle [64]: That of logic and computational modeling. In particular, she argued that such methods might elucidate the workings of 'higher order social cognition', which she uses as a synonym for 'higher order theory of mind'. Even without the 'higher order' label, 'theory of mind' is often considered the most complex form of social cognition [14]; it is the ability to appreciate what others see, know and want; the ability to put oneself 'in another's shoes', figuratively speaking [52], and to realize that another's perspective on the world may be different from one's own. 'Higher order theory of mind', then, is the ability to not just think about another's mental states, but to think about mental states that are themselves about mental states [21]. In this framework, Sujata having the thought 'Rineke wants the banana' would be an example of first-order 'theory of mind', while Jakub thinking 'Sujata knows that Rineke wants the banana' would be second-order (see, e.g., [43]).

With respect to computational models of 'higher order theory of mind', Verbrugge's hope was that two different types of simulation might prove useful [64]. On the one hand, she argued that cognitive models might shed light on why both children and adults find 'higher-order theory of mind' relatively difficult. In a cognitive model, the mental operations that underlie behavior are modeled explicitly, ranging from the processing of visual information to the learning of new strategies and the retrieval of facts from declarative memory [57]. Thus, in a cognitive model, it is possible to implement different theories of how humans solve 'higher order theory of mind' tasks and see which best predict the errors and reaction times of experimental participants.

Verbrugge's other suggestion was to study the evolution of 'higher order theory of mind' with agent-based models [64]. Agent-based models are like cognitive models in that they explicitly simulate individuals, but in agent-based models, the focus is much more on the interaction between different agents and their environment; the agents themselves, and certainly their cognitive processes, are usually represented in a more simplified manner than in cognitive models. Often, if an agent-based model makes testable predictions, it is at the group level, showing how a population might change across time or space (see, e.g., [20] for a review). However, in addition to aiding such extrapolation, an agent-based model can also produce new insights by generating unexpected, emergent patterns of behavior [35], as we discuss later. Verbrugge's proposal was to use agent-based models to test different 'scenarios' for the evolution of 'higher order theory of mind', essentially simulating different hypothesized 'evolutionary pressures' and investigating the extent to which they select for

more complex social cognition.

Existing theories posit that perhaps a need for cooperation [45], or 'Machiavellian manipulation' [9], or larger group size more generally [22], was the driving force behind the advanced cognitive abilities of primates and other large brained creatures, like corvids [24]; Verbrugge [64] adds to this the hypothesis that a requirement for mixed-motive negotiation might have created the final push towards 'higher order theory of mind' in humans, where 'mixed motive negotiation' refers to scenarios where individuals want to work together while still trying to maximise their own interests. In an agent-based model, these different scenarios can be 'played out' in different artificial worlds, to assess the plausibility behind each hypothesis.

Since 2009, Verbrugge and co-authors have published a number of papers simulating complex social cognition in humans [66,65,2,42,41]. However, many cognitive and agent-based models actually deliver the opposite message, that is often experienced as 'killjoy' [56] - that what looks like complex social cognition might not be either social or complex at all! In this paper, we will argue that the insights from such models are highly relevant to Verbrugge's proposed 'scenario testing'. In the rest of this article, we will first discuss why many computational models often have such 'killjoy' messages; then we will discuss what this means for 'scenario' testing more generally, and we conclude with a number of our own recommendations for future research directions.

2 Why are computational models often 'killjoy'?

First, let us illustrate our claim that computational models often take the 'complexity' out of complex social cognition with a number of examples; explanations for *how* these models generate the appropriate patterns follow in later sections. Perhaps most notably, there is the 'DomWorld' model, which focuses on the dominance hierarchies and social relationships of primates. Originally used to parsimoniously explain the differences between 'despotic' and 'egalitarian' species of macaque [34], its basic setup has now been adapted to investigate many different primate phenomena, ranging from female dominance over males [38] to the effect of fleeing on socio-spatial group structure [27]. However, with respect to simulating social cognition, DomWorld's most relevant iteration is 'GrooFiWorld' [37,53], a DomWorld extension that includes grooming in addition to fighting and fleeing.

In GrooFiWorld, simulated primate-like entities reconcile more often with valuable partners than with other individuals [37,53], they exchange grooming for support in fights [37], and they predominantly support those who are of higher rank than their opponents [37]. These are all social behaviors which have been previously given 'cognitively complex' interpretations, such as contingent reciprocity [28,29], emotional book-keeping [55] and 'triadic awareness', i.e., the ability to compare the relative strengths of others' social relationships [30,12]. However, the agents in GrooFiWorld have no such abilities; they just group, attack when they think they can win, and groom when they are anxious [37,53].

Two other examples come from our own work. The 'corvid model' [63,62] is a cognitive model based on experiments done with Western scrub jays, a member of the crow family that caches its food for later retrieval, but will also steal the caches of others when it knows where they are [26]. The scrub jay has been ascribed 'theory-of-mind-like' abilities because it displays a fascinating array of behavioral patterns: (i) it prefers to cache far away [17,16], behind barriers [16], and in dark areas [15] when it is watched; (ii) it repeatedly re-caches items while in the presence of onlookers [17,16]; and (iii) after onlookers have left, it moves its items to new locations [26,17,16,23]. These behaviors seem to suggest that the scrub jay is trying to manipulate what others see and know; however, the 'corvid model' explains most of these patterns as side-effects of stress and memory errors [62], as we explain below.

Finally, the 'diet traditions model' [48,49] investigates the emergence of diet traditions and cumulative cultural learning. Initially, this model was used to study why neighboring primate groups might eat different kinds of foods, despite many of the same resources being available in their home ranges, as is found, for instance, in white-faced capuchins [11]. Then, the same model was used to study 'cumulative culture', the idea that successive generations might improve on the innovations of previous generations, with the effect of achieving behavioral complexity that would otherwise be beyond the learning abilities of individuals. Such 'cumulative culture' is typically thought to be the explanation for humanity's technological prowess [54], and is thought to rely on sophisticated forms of social learning, such as imitation and teaching [58,19]; however, in the 'diet traditions model', successive generations of agents selectively eat ever more nutritious foods, beyond their ability to learn individually, and this occurs as a side-effect of grouping without any conscious copying or teaching at all [49].

Now, the question is, why do agent-based models, like DomWorld and the 'diet traditions' model, and cognitive models, like the 'corvid model', often suggest that behaviors might be 'less socially complex' than they at first appear to be? We suspect that part of the answer has to do with practical considerations: there may be a greater motivation to produce models that challenge our preconceptions. There is an inherent methodological argument for searching for the simplest explanation of observed patterns of behavior, if only to be able to exclude it later. However, an explanation involving relatively sophisticated cognition can turn out to be quite an intuitive explanation (i.e. one that we easily understand and already expect). In contrast, an explanation involving relatively unsophisticated cognition can turn out to be quite convoluted and/or counter-intuitive (i.e. explanations that are surprising, and/or difficult to understand). Therefore, cognitively complex explanations are typically the default, and the burden of proof is on those attempting to illustrate that simpler cognitive mechanisms suffice. As a result there may be particular motivation (and bias) to use agent-based models to show that 'smart looking' behavior might in fact be generated by simpler cognitive mechanisms, rather than using agent-based models to show what everyone already thinks anyway.

For a model to be truly valuable, it must produce some new insight, some fresh perspective on existing theories or preconceptions.

However, there is more to it than just practical considerations; the main reason why computational models often present alternative, cognitively simple hypotheses is because they are uniquely powerful tools for discovering such hypotheses. In our view, this is because computational models have two very important properties: They promote embodied, embedded thinking, and they naturally give rise to feedbacks which are difficult to think through otherwise.

The concept of embodied, embedded thinking has its roots in robotics research [47,6], philosophy of mind [13] and studies of self-organisation [35,10]. A full discussion of its intricacies is outside the scope of this paper, but for our purposes, it is sufficient to summarize it like this: *It is important to realise that individuals do not have to solve all their problems by explicit cognitive processing. Instead, the solutions to some problems arise naturally from other aspects of their own physical instantiation – they are embodied – or their interaction with the outside world – they are embedded (or 'situated')*. This kind of thinking is naturally fostered by agent-based models, and to a lesser degree by cognitive models, if only because they often try to simulate a fully functioning organism, no matter how simplified, in some kind of virtual environment, no matter how abstract.

Further, because both agent-based and cognitive models are relatively richly implemented, either in terms of their interactions or in terms of their internal mechanisms, feedback processes can occur, causing self-reinforcing loops and stabilising effects. The patterns arising in these models are therefore not end-products, but have a further impact on dynamics. Feedback is one of the central principles of self-organisation [10], and it is as important to models of social cognition as it is to subjects with which it is more directly associated, such as task division in insects [3] and collective movement in fish and birds [36].

In the rest of this section, we illustrate how embodied, embedded thinking and feedbacks explain the alternative hypotheses generated by DomWorld, the 'corvid model', and the 'diet traditions' model. Then, we will return to the question of what such 'killjoy' insights mean for Verbrugge's 'scenario-testing' research agenda with respect to the evolution of complex social cognition.

2.1 Computational models are embodied and embedded

The agents in agent-based models, in particular, are naturally embedded in the sense of 'physically located in a social or environmental context'; in fact, one could argue that this 'embeddedness' is what defines an agent-based model in the first place. Both DomWorld and 'diet tradition' agents are embedded in their groups; in addition, 'diet tradition' agents are embedded in the distribution of available food. This allows them to use the outside world as a kind of external 'memory'. In particular, if an individual's location is not independent of preceding behavioral processes, then its position relative to other individuals, or to other features in the environment, can encode a lot of relevant information.

One example of this occurs in GrooFiWorld. In GrooFiWorld, individuals engage in dominance interactions, and flee if they lose a fight. Individuals are risk-averse, and only attack if they think they can win; this is why dominants often jointly attack subordinates. But the result of fleeing is that subordinate individuals end up on the periphery of the group, while dominant ones end up in the center. Thus, the spatial positioning of individuals in GrooFiWorld causes individuals of similar dominance to interact relatively often. This is true for both fighting and grooming, explaining why individuals appear to exchange grooming for support in fights [37].

The fact of being embedded is also at the core of how information can be transmitted in groups indirectly [48,50], via the grouping process itself, rather than via direct forms of social learning, such as stimulus enhancement, where individuals observe what others are doing and how they perform the behavior [39]. This is what explains the diet traditions and cumulative culture in the 'diet traditions model' [49]: Young, inexperienced individuals learn what to eat from their group mates, not because they are actively imitating their group mates' behavior, but because their group mates are stopping to eat in areas that contain familiar foods. Then, because the inexperienced individuals want to stay close to their group mates, they end up eating and learning about the same foods. As a result preferences for certain types of food are inherited, leading to 'diet traditions' via trial-and-error learning in groups.

Trial-and-error learning in groups can even be sufficient for cumulative cultural processes, even though the latter are typically considered to rely on cognitively sophisticated forms of social learning. In the diet model, experienced foragers do not have perfect information: They may be eating suboptimal foods simply because they are unaware of more profitable foods. A young, inexperienced forager is less prone to these biases; if a food is too low quality compared to other foods it already knows about, an inexperienced forager will not stop unless its group mates do, and instead looks for new, better foods. Moreover, it will only stop to eat food its group does not eat if those foods are better than food types it is already eating. This is how diet quality can improve cumulatively across generations. Thus, indirect group-level information processing can help cognitively simplistic individuals achieve more than they could on their own [49,50], and this can even be a reason for grouping to evolve [50].

Embodiment arises less naturally in agent-based models; because individual agents are often represented quite simply, they do not usually have much of a 'body' to generate alternate explanations with; however, if one expands 'embodiment' to mean 'incorporation of other physiological processes beyond explicit deliberative reasoning' (similar to [18]), then GrooFiWorld counts as an example: In GrooFiWorld, agents groom when they are anxious, and they become anxious when participating in fights, or when they have not been groomed for a long time; in contrast, grooming itself reduces anxiety. This is based on observations of real primates, and represents a type of 'non-deliberative' embodied physiological process that is crucial to producing the suite of primate-like behavioral patterns exhibited by GrooFiWorld. For in-

stance, this mechanism explains why individuals appear to 'reconcile'; they are tense after fighting, and tend to be closer to their former opponents than to any other individuals.

In cognitive models, both embeddedness and embodiment are often less obvious. Early cognitively models consisted almost exclusively of internal representations, and even now, in studies of human cognition, the 'environment' is usually reduced to a computer screen [61], and 'the body' to eyesight and fingers to type with. However, even so, because the computer screen itself is also simulated, and a cognitive model must explicitly shift its 'attention' to the visual stimuli presented there, a cognitive model is automatically more embedded and embodied than many verbal theories are. This is illustrated by the fact that cognitive models have been used to study how memory can be offloaded into the environment [4], and by variants which are adapted to run on robots and interact with reality [61]. In the 'corvid model', both embeddedness and embodiment also play a role; cache distributions and stress are important aspects of its 'killjoy' explanation. However, the 'corvid model' is perhaps best understood as an example of the power of feedback, as will be explained in the following section.

2.2 Computational models exhibit feedbacks

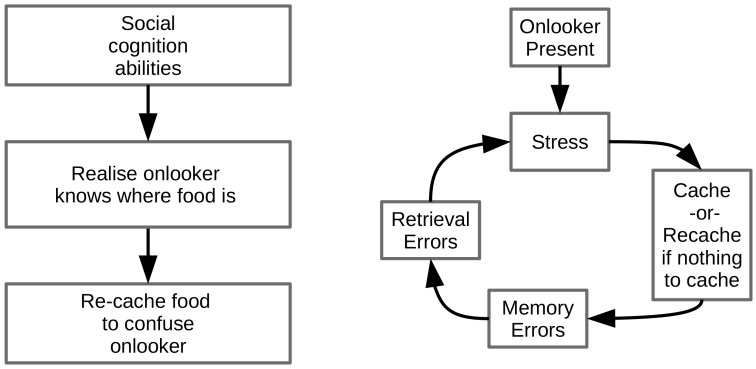


Figure 2. Illustration of the 'theory of mind' and 'stress' hypotheses for scrub re-caching. The 'theory of mind' hypothesis (left) is the one tested in the experiment by [26], while the 'stress hypothesis' (right) is the one featured in [63]. Notice how the 'theory of mind' hypothesis is essentially 'linear', directly linking intentions to actions, while the 'stress' hypothesis includes unintentional feedbacks - feedbacks only revealed through the use of a computational model.

The 'corvid model' generates the seemingly 'theory-of-mind-like' behavior of scrub jays through a series of simple behavioral rules and feedbacks. It

assumes that virtual jays cache their food, and that they remember where they have cached it. However, there is some error on this memory. In addition, it is assumed that virtual jays want to cache more when they are stressed, and that both the presence of an onlooker and not finding food where it was expected cause stress. If such a virtual jay is put through the experimental protocol used with scrub jays, it generates patterns of caching and re-caching that are remarkably similar to those of the real jays [63]. However, the virtual jay is not attempting to deceive the onlooker, nor is it aware of the onlooker's knowledge or mental state.

What happens instead is that, while a conspecific is watching, the virtual jay is stressed because it is being watched, and re-caches because it has no other food left to cache other than what it has previously buried. However, this digging up and moving of items then confuses its memory, so that later, once it is alone, it again becomes stressed because some of its items seem to be missing. This additional stress then again causes it to want to cache more, which it can only do by re-burying the items it has just recovered. Thus, one could claim that in the model the observed behavior has very little to do with sociality at all, given that any form of stress should generate the result (in reality, this does not seem to be the case [59], as will be discussed later).

This explanation for the scrub jays' behavior contrasts with the 'theory of mind hypothesis' that is typically given for their re-caching acts: That the cachers realise that their onlookers can see them, and that, therefore, they should move their items, so that other birds will no longer know where they are. The two explanations differ not only in the cognitive sophistication that they require of scrub jays, but in their basic structure: In the 'theory of mind hypothesis', what we see the birds do (move their items) is directly, or 'linearly', explained by what they are trying to do (make sure their conspecifics don't know where their items are). In contrast, in the 'stress hypothesis', the observable behavior is the result of an unintentional feedback loop between stress and memory errors (Figure 2).

Of course, there is no reason that hypotheses conceived *without* the help of computational models cannot contain feedbacks - but it is very difficult to think through such feedbacks without the help of a running computer programme. In contrast, in a cognitive or agent-based model, such feedbacks arise almost automatically. This is, in part, due to the fact that computational models encourage us to include more (and different) 'variables' than are typically considered in the original 'mental' or 'intuitive' hypotheses.

3 So what does this mean for 'scenario testing'?

What GrooFiWorld, the 'diet traditions' model, and the 'corvid model' all suggest is that complex cognition is unnecessary to generate specific sets of smart-looking behavior. However, both primates and corvids have unusually large brains [24], with high energy requirements; this investment must come with some enhancement in information processing (i.e. cognition). If computational models keep showing that this 'enhanced information processing' does

not seem to be necessary in the social realm, then we are left with two main implications: (1) Either the large brains of primates and corvids primarily serve some other, non-social purpose, or (2) computational models are oversimplifying things. Below we consider these and other implications of 'killjoy' results in more detail:

3.1 Big brains are not social

If we assume that the computational models are right, then perhaps we should conclude that the 'complex cognition' of humans and other big-brained creatures evolved primarily to solve some non-social problem, like tracking the availability of food sources [44,40], or mastering the technical tricks necessary to obtain them [8,7]. If this is true, then studying the evolution of social cognition in isolation is not likely to mirror reality in any meaningful way. We return to this possibility, and what it implies for future modeling efforts, in the following section. Of course, the computational models might not be right – a new study testing the predictions of the 'corvid model' [59], for instance, fails to confirm them – showing that a model's ability to replicate a given set of patterns does not mean that the underlying explanation itself is correct.

3.2 False-positives in comparative analyses

A more nuanced view is that computational models point to the possibility that some (but not necessarily all) of the cases where we have identified complex social cognition in animals may be false-positives. If so, we may need to reassess the overall phylogenetic patterns of complex social cognition. In turn, this will affect our assessment of which scenarios to focus on with respect to promising candidate scenarios in which social cognition could evolve.

For instance, if GrooFiWorld's explanations for primate social patterns are the whole story – if macaques can reconcile, trade favors, and form coalitions without any complex social cognition at all – then hypothetical evolutionary 'scenarios' based on the challenges of primate social life are unlikely to lead to any great insights. (Of course, it is also possible that GrooFiWorld's explanations are not the whole story, as even GrooFiWorld's creators explicitly acknowledge – but we will return to this possibility later).

3.3 False-positives in theoretical analyses

'Killjoy' models highlight that embodiment and embeddedness are important. They are what allow for the generation of feedbacks and self-reinforcing effects; without them, behavior can often seem to require more complex cognition than it might otherwise need. In our view, this means that simulating the evolution of any kind of cognition in 'a vacuum' is prone to result in false-positives. Yet, when simulating very complex behavior – such as mixed-motive negotiation – this is often what has to happen: The only aspect of the agent that is simulated is that which is directly necessary to accomplish the 'challenge' it is set, and the environment itself is reduced to that 'challenge' alone – see, for instance, de Weerd, Verbrugge and Verheij's studies on the usefulness of 'higher order theory of mind' [66,65] or Arbilly et al.'s [1] simulations of

the 'cognitive arms race' that may arise in species of cachers and pilferers, like corvids. When both the agents and the world are 'streamlined' to this degree, a given evolutionary 'scenario' may seem to favor the evolution of complex social cognition, when in fact embodiment and embeddedness could generate similar outcomes, if given a chance.

3.4 False-negatives due to over-simplification

On the other hand, if computational models are oversimplifying matters – for instance, if complex social cognition does play a role in primate social life, despite GrooFiWorld's ability to replicate many of the patterns typically taken as evidence that such complexity exists – then the question is: Why does oversimplification enable the models to generate patterns that should only arise with complex cognition?

Of course, simplifications are unavoidable and desirable when formulating models. However, when it comes to simulating the cognitive abilities of animals, simplifications can have important ramifications, in particular with respect to the 'frame problem'. The frame problem is essentially an artificial intelligence concept, which defines a problem in terms of its frame of reference, or the number of variables (and the combinatorial impact this has) that needs to be dealt with in order to solve it. The essential weakness of both computational models and experimental designs is that they alter the frame of reference (e.g. the salience of key aspects of a task), and therewith the complexity of the problem to be solved. Probably, given sufficient simplification, basic associative learning can solve any problem, given that it is fed the problem piece-meal, or the problem is reduced to some minimal core cause-effect component.

Take, for example, Harrison et al.'s [33] cognitive model of gaze following in chimpanzees. This simulation is based on an experiment [31] attempting to assess what chimpanzees understand about seeing; it pits a subordinate against a dominant chimpanzee in a competition for two pieces of food, one of which is visible to the subordinate only, and one of which both chimpanzees can see. It turns out that across a variety of setups [31,5,32] the subordinate chooses to first go to the food only it can see. This has been interpreted as evidence that the subordinate has some understanding of the dominant's visual perspective. But in Harrison et al.'s [33] cognitive model, the subordinate simply performs a visual search between the food and the dominant, and prefers to go to food where it finds an obstruction along the way. Thus, it is considering the situation only from its own perspective, not that of its competitor.

The 'virtual chimpanzee' acquires this strategy through reinforcement learning. This reinforcement learning takes place during ten full replicates of the experiment itself, where this built-in strategy becomes more and more preferred over a simpler, 'grab-and-go' alternative. This is supposed to mimic the experience that real chimpanzees get during their daily lives [33]. Although this is a convenient simplification, it effectively 'solves' the frame problem for the 'virtual chimpanzee'. A real chimpanzee would have to learn this strategy in a much less structured environment, while simultaneously filtering out all the

other things that were occurring at the same time and irrelevant to its food finding success. Thus, what looks like a simple learning problem in the 'competitive model' might in fact be a complex learning problem in real life, and therefore underestimate the added value of more complex social cognition.

The oversimplification of scenarios can therefore lead to an underestimation of the cognition that is required in different situations. This generates a problem of false-negatives with respect to the cognition employed by animals in cognitive tests, as well as with respect to assessing how cognition can be adaptive in various settings. The conundrum we are left with is therefore: *Computational models point to the problem of false-positives, but potentially suffer from the problem of false-negatives in terms of the cognitive sophistication required to navigate life's challenges*.

4 Future directions

So, can computational models help us understand the evolution of 'complex social cognition', as Verbrugge [64] proposes? Our answer is a resounding 'yes'! However, while computational models are clearly powerful tools for analysing complex systems, we should be aware of their limitations. Below we recap why computational models are useful, and propose two main ways with which to resolve the conundrum of false-positives and false-negatives.

4.1 The power of computational models

The main insight coming from computational models is that interactions between a multitude of variables and entities can generate unexpected patterns and dynamics via complex feedbacks. These dynamics can easily generate novel causal explanations for patterns that were previously interpreted 'linearly'. This is a general feature of computational models, above and beyond the specific explanations generated in response to specific research questions. Thus, we posit that, for any observed behavior, there is probably a computational model that can offer a different causal explanation than the linear one that seems self-evident. Clearly, science progresses by comparing alternative causal explanations, and without computational models, many possible causal explanations would be missed; for this reason alone, computational models are indispensable tools in studies of social cognition.

4.2 Allowing for side-effects to avoid false-positives

However, the unexpected patterns and dynamics that arise in computational models also suggest the danger of 'false-positives' in evolutionary 'scenario testing'. For example, if big brains did not evolve for social cognition, i.e. if we take the 'killjoy' lessons to heart, then social cognition could be part of more general purpose cognition. For example, if such cognition evolved for reasons other than sociality, then social cognition could be an evolutionary side-effect. In that case there may not be anything special about social learning, and it is just domain-general mechanisms functioning in social situations.

Allowing for 'side-effects' in evolutionary models is actually surprisingly dif-

ficult. First of all it requires multiple traits that are inter-related, so that if there is selection on one trait, as a side-effect this will cause another trait to change as well. Secondly, for side-effects to arise, they need to emerge via self-organization. The traits therefore cannot be directly implemented, but must be the product of a set of lower-level features that are built into the agents or the environment. Thirdly, one would probably wish to compare at least two kinds of selection pressure: one that leads to direct selection of social cognition, and one that allows social cognition to evolve as a side-effect. To achieve this requires sufficient environmental detail – true embeddedness, in other words.

The inter-relation of multiple traits, their emergence from lower-level features, and the existence of environmental detail will cause the model to be much more complex than models that focus on a particular predefined 'single-challenge' scenario. As a result of this increased complexity, models which allow for evolutionary side-effects tend to be agent-based models, where the agents themselves are relatively unsophisticated. What the models show, however, is that different kinds of behavior (foraging, grouping, movement) are intricately related, and their relationships are dependent on the cognition of individuals [51]. Cognition is therefore involved in most patterns of behavior (i.e. multiple higher-level traits). Changes in cognition are therefore expected to alter multiple patterns, and these may all affect whether the changes would be selected. These results suggest that social cognition may well be the result of selection pressures for general purpose cognition, and we should build models that allow for such a possibility.

4.3 Taking the frame problem seriously and addressing false-negatives

In order to deal with false-negatives due to oversimplification, we propose to build computational models that explicitly implement perception, motor skills, learning and attention-selection procedures. This is currently the domain of cognitive models. The challenge for the future is to embed such cognitive models more fully in more detailed simulated environments, where feedbacks and self-organisation can more easily emerge - essentially, to 'bridge the gap' between cognitive and agent-based models. In addition, as Verbrugge [64] suggests, it would be useful to generate theory about how to explicitly define the complexity of problems in computational models and experiments (e.g. via formal logics), and what performance on these problems therefore implies about an agent's cognitive abilities. Our prediction is that if oversimplification has lead to false-negatives, then explicit incorporation of perception, motor skills, learning and attention-selection procedures should also lead to more sophisticated cognition being necessary for computational agents to solve a given task.

5 Conclusion

We conclude that Verbrugge's proposed plan for tackling the evolution of higher-order social cognition has born fruit, but of a surprising, yet thought provoking flavor. It seems that one of the most useful properties of compu-

tational models lies in their ability to be 'killjoy', i.e. to explain cognitively-complex looking behavior as a side-effect of embodiment, embeddedness, and feedbacks; however, this may only be possible because the computational models simplify the world, and the cognition of agents in that world, to an unacceptable degree. The challenge for the future is to find out whether and when we are oversimplifying. This is true for models of cognitive experiments, models of social processes, as well as evolutionary models and is more generally relevant. Our proposal is to focus on evolutionary models with co-evolving traits and the potential for side-effects, as well as a more detailed and explicit implementation of the perception, motor skills, learning and attention-selection procedures that are necessary to solve problems in the real world.

References

[1] Arbilly, M., D. B. Weissman, M. W. Feldman and U. Grodzinski, *An arms race between producers and scroungers can drive the evolution of social cognition*, Behav. Ecol. **25** (2014), pp. 487–495.

[2] Arslan, B., N. A. Taatgen and R. Verbrugge, *Modeling developmental transitions in reasoning about false beliefs of others*, in: R. West and T. Stewart, editors, *Proceedings of the 12th International Conference on Cognitive Modelling* (2013), pp. 77–82.

[3] Beshers, S. N. and J. H. Fewell, *Models of division of labor in social insects*, Annu. Rev. Entomol. **46** (2001), pp. 413–440.

[4] Borst, J. P., T. A. Buwalda, H. van Rijn and N. A. Taatgen, *Avoiding the problem state bottleneck by strategic use of the environment*, Acta Psychol. (Amst) **144** (2013), pp. 373–379.

[5] Bräuer, J., J. Call and M. Tomasello, *Chimpanzees really know what others can see in a competitive situation*, Anim Cogn **10** (2007), pp. 439–448.

[6] Brooks, R., "Cambrian Intelligence: The Early History of the New AI," The MIT Press, Cambridge, MA, 1999.

[7] Byrne, R., *Cognition in great ape ecology: Skill-learning ability opens up foraging opportunities*, in: H. O. Box and K. R. Gibson, editors, *Mammalian Social Learning: Comparative and Ecological Perspectives* (1999), pp. 333–350.

[8] Byrne, R. W., *The technical intelligence hypothesis: An additional evolutionary stimulus to intelligence*, in: R. W. Byrne and A. Whiten, editors, *Machiavellian Intelligence II: Evaluations and Extensions* (1998), pp. 289–311.

[9] Byrne, R. W. and A. Whiten, "Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans," Oxford Univeristy Press, Oxford, 1988.

[10] Camazine, S., J. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz and E. Bonabeau, "Self-organization in Biological Systems," Princeton University Press, Princeton, 2001.

[11] Chapman, C. A. and L. M. Fedigan, *Dietary differences between neighboring* Cebus capucinus *groups: Local traditions, food availability or responses to food profitability?*, Folia Primatol **54** (1990), pp. 177–186.

[12] Cheney, D. L. and R. M. Seyfarth, "Baboon Metaphysics: The Evolution of a Social Mind" University of Chicago Press, Chicago, 2007.

[13] Clark, A., "Being There: Putting Brain, Body, and World Together Again," MIT Press, Cambridge, Mass., 1997.

[14] Clayton, N. S., J. M. Dally and N. J. Emery, *Social cognition by food-caching corvids. The western scrub-jay as a natural psychologist*, Philosophical Transactions of the Royal Society B: Biological Sciences **362** (2007), pp. 505–522.

[15] Dally, J. M., N. J. Emery and N. S. Clayton, *Cache protection strategies by western scrub-jays (Aphelocoma californica): Hiding food in the shade*, Proc. Biol. Sci. **271 Suppl 6** (2004), pp. S387–S390.

[16] Dally, J. M., N. J. Emery and N. S. Clayton, *Cache protection strategies by western scrub-jays, Aphelocoma californica: Implications for social cognition*, Animal Behaviour **70** (2005), pp. 1251–1263.
[17] Dally, J. M., N. J. Emery and N. S. Clayton, *Food-caching western scrub-jays keep track of who was watching when*, Science **312** (2006), pp. 1662–1665.
[18] Damasio, A. R., "Descartes' Error: Emotion, Reason and the Human Brain," Grosset/Putnam, 1994.
[19] Dean, L. G., G. L. Vale, K. N. Laland, E. Flynn and R. L. Kendal, *Human cumulative culture: A comparative perspective*, Biol Rev Camb Philos Soc (2013).
[20] DeAngelis, D. and V. Grimm, *Individual-based models in ecology after four decades*, F1000 Prime Reports, **6** (2014).
[21] Dennett, D. C., *Intentional systems in cognitive ethology: The "Panglossian paradigm" defended*, Behavioral and Brain Sciences **6** (1983), pp. 343–390.
[22] Dunbar, R. I. M., *The social brain hypothesis*, Evolutionary anthropology **6** (1998), pp. 178–190.
[23] Emery, N. J. and N. S. Clayton, *Effects of experience and social context on prospective caching strategies by scrub jays*, Nature **414** (2001), pp. 443–446.
[24] Emery, N. J. and N. S. Clayton, *The mentality of crows: Convergent evolution of intelligence in corvids and apes*, Science **306** (2004), pp. 1903–1907.
[25] Emery, N. J. and N. S. Clayton, *Comparative social cognition*, Annu Rev Psychol **60** (2009), pp. 87–113.
[26] Emery, N. J., J. M. Dally and N. S. Clayton, *Western scrub-jays (Aphelocoma californica) use cognitive strategies to protect their caches from thieving conspecifics*, Animal Cognition **7** (2004), pp. 37–43.
[27] Evers, E., H. de Vries, B. Spruijt and E. Sterck, *Better safe than sorry: Socio-spatial group structure emerges from individual variation in fleeing, avoidance or velocity in an agent-based model*, PloS One **6** (2011), p. e26189.
[28] Frank, R. E. and J. B. Silk, *Impatient traders or contingent reciprocators?*, Behaviour **146** (2009), pp. 1123–1135.
[29] Gomes, C. M., R. Mundry and C. Boesch, *Long-term reciprocation of grooming in wild West African chimpanzees*, Proc. R. Soc. Lond. B **276** (2009), pp. 699–706.
[30] Harcourt, A. H. and F. B. M. De Waal, "Coalitions and Alliances in Humans and Other Animals," Oxford University Press, New York, 1992.
[31] Hare, Call, Agnetta and Tomasello, *Chimpanzees know what conspecifics do and do not see*, Anim Behav **59** (2000), pp. 771–785.
[32] Hare, B., J. Call and M. Tomasello, *Do chimpanzees know what conspecifics know?*, Anim Behav **61** (2001), pp. 139–151.
[33] Harrison, A. and J. G. Trafton, *Gaze-following and awareness of visual perspective in chimpanzees*, in: A. Howes, D. Peebles and R. Cooper, editors, *Proceedings of the Ninth International Conference on Cognitive Modeling*, Manchester, UK, 2009, pp. 292–297.
[34] Hemelrijk, C. K., *An individual-orientated model of the emergence of despotic and egalitarian societies*, Proceedings of the Royal Society of London B **266** (1999), pp. 361–369.
[35] Hemelrijk, C. K., *Understanding social behaviour with the help of complexity science (Invited article)*, Ethology **108** (2002), pp. 655–671, doi: 10.1046/ j.1439-0310.2002.00812.x.
[36] Hemelrijk, C. K. and H. Hildenbrandt, *Schools of fish and flocks of birds: their shape and internal structure by self-organization*, Interface Focus **2** (2012), pp. 726–737.
[37] Hemelrijk, C. K. and I. Puga-Gonzalez, *An individual-oriented model on the emergence of support in fights, its reciprocation and exchange*, PLoS ONE **7** (2012), p. e37271.
[38] Hemelrijk, C. K., J. Wantia and K. Isler, *Female dominance over males in primates: Self-organisation and sexual dimorphism*, PLoS ONE **3** (2008), p. e2678.
[39] Heyes, C. M., *Social learning in animals: Categories and mechanisms*, Biol. Rev. **69** (1994), pp. 207–231.
[40] Janmaat, K. R. L., L. Polansky, S. D. Ban and C. Boesch, *Wild chimpanzees plan their breakfast time, type, and location*, Proc. Natl. Acad. Sci. U.S.A. (2014).
[41] van Maanen, L. and R. Verbrugge, *A computational model of second-order social reasoning*, in: D. Salvucci and G. Gunzelmann, editors, *Proceedings of the 10th International Conference on Cognitive Modelling* (2010), pp. 259–264.

[42] Meijering, B., N. A. Taatgen, H. van Rijn and R. Verbrugge, *Reasoning about mental states in sequential games: As simple as possible, as complex as necessary*, in: R. West and T. Stewart, editors, *Proceedings of the 12th International Conference on Cognitive Modelling,* (2013), pp. 173–178.
[43] Miller, S. A., *Children's understanding of second-order mental states.*, Psychol Bull **135** (2009), pp. 749–773.
[44] Milton, K., "Foraging behaviour and the evolution of primate intelligence. In Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans, Byrne, R. W., Whiten, A. (Eds)," Oxford Univeristy Press, Oxford, 1988 pp. 285–305.
[45] Moll, H. and M. Tomasello, *Cooperation and human cognition: The Vygotskian intelligence hypothesis*, Philos. Trans. R. Soc. Lond., B, Biol. Sci. **362** (2007), pp. 639–648.
[46] van Overwalle, F., *Social cognition and the brain: A meta-analysis*, Human Brain Mapping **30** (2009), pp. 829–858.
[47] Pfeifer, R. and C. Scheier, "Understanding Intelligence." MIT Press, 1999, I–XIX, 1–697 pp.
[48] van der Post, D. J. and P. Hogeweg, *Resource distributions and diet development by trial-and-error learning*, Behav Ecol Sociobiol **61** (2006), pp. 65–80.
[49] van der Post, D. J. and P. Hogeweg, *Diet traditions and cumulative cultural processes as side-effects of grouping*, Anim Behav **75** (2008), pp. 133–144.
[50] van der Post, D. J. and D. Semmann, *Patch depletion, niche structuring and the evolution of cooperative foraging*, BMC Evolutionary Biology **11** (2011), p. 335.
[51] van der Post, D. J., R. Verbrugge and C. K. Hemelrijk, *The evolution of different forms of sociality: Behavioral mechanisms and eco-evolutionary feedback*, PloS One (under review).
[52] Premack, D. and G. Woodruff, *Does the chimpanzee have a theory of mind?*, Behavioral and Brain Sciences **1** (1978), pp. 515–526.
[53] Puga-Gonzalez, I., H. Hildenbrandt and C. K. Hemelrijk, *Emergent patterns of social affiliation in primates, a model*, PLoS Computational Biology **5** (2009), p. e1000630.
[54] Richerson, P. J. and R. Boyd, "Not by Genes Alone. How Culture Transformed Human Evolution," The University of Chicago Press, Chicago, 2005.
[55] Schino, G. and F. Aureli, *Reciprocal altruism: Partner choice, cognition and emotions.*, Advances in the Study of Behavior **39** (2009), pp. 45–69.
[56] Shettleworth, S. J., *Clever animals and killjoy explanations in comparative psychology*, Trends in Cognitive Sciences **14** (2010), pp. 477–481.
[57] Sun, R., "The Cambridge Handbook of Computational Psychology," Cambridge University Press, New York, NY, USA, 2008.
[58] Tennie, C., J. Call and M. Tomasello, *Ratcheting up the ratchet: On the evolution of cumulative culture*, Philosophical Transactions: Biological Sciences **364** (2009), pp. 2405–2415.
[59] Thom, J. M. and N. S. Clayton, *Re-caching by western scrub-jays (Aphelocoma californica) cannot be attributed to stress*, PLoS ONE **8** (2013), p. e52936.
[60] Tomasello, M., M. Carptenter, J. Call, T. Behne and H. Moll, *Understanding and sharing intentions: The origins of cultural cognition*, Behavioral and Brain Sciences **28** (2005), pp. 675–735.
[61] Trafton, J. G., L. Hiatt, A. Harrison, F. Tamborello, S. Khemlani and A. Schultz, *ACT-R/E: An embodied cognitive architecture for human-robot interaction*, Journal of Human-Robot Interaction **2** (2013), pp. 3–55.
[62] van der Vaart, E., R. Verbrugge and C. K. Hemelrijk, *Corvid caching: Insights from a cognitive model*, J Exp Psychol Anim Behav Process **37** (2011), pp. 330–340.
[63] van der Vaart, E., R. Verbrugge and C. K. Hemelrijk, *Corvid re-caching without 'theory of mind': A model*, PLoS ONE **7** (2012), p. e32904.
[64] Verbrugge, R., *Logical and social cognition: The facts matter, and so do computational models*, Journal of Philosophical Logic **38** (2009), pp. 649–680.
[65] de Weerd, H., R. Verbrugge and B. Verheij, *Higher-order theory of mind in negotiations under incomplete information*, in: G. Boella, E. Elkind, B. T. R. Savarimuthu, F. Dignum and M. K. Purvis, editors, *PRIMA*, Lecture Notes in Computer Science **8291** (2013), pp. 101–116.
[66] de Weerd, H., R. Verbrugge and B. Verheij, *How much does it help to know what she knows you know? An agent-based simulation study*, Artif. Intell. **199** (2013), pp. 67–92.

The Importance of Accounting for Heterogeneity of Strategy Use

Maartje Raijmakers

University of Amsterdam

1 Introduction

Research in cognitive development traditionally has a strong focus on higher-order reasoning, compared to adult research. One of the important questions in developmental research is the level of reasoning that children of different ages have. That is, an important subject of study is whether children apply increasingly complex strategies in solving problems. I will argue that the analysis of heterogeneity of behavioral data in terms of categorically different strategies is important for adult cognitive science as well.

Jean Piaget, as a founder of cognitive developmental research, introduced the importance of specific, diagnostic tests to assess children's level of reasoning [38]. At the time he was registering IQ-tests in the lab of Binetin Paris (1919-1921) he noticed that the errors children made in IQ-tests were very specific, different from adults, and not just a random choice other than the correct answer. This observation led to the hypothesis that children, when they do not have a correct strategy to solve problems, use an alternative, sub-optimal strategy. That is, children have their own logic in reasoning.

The idea that error-patterns in responses are diagnostic of a reasoning strategy was a very fruitful idea. Robert Siegler [36] was one of the first researchers who has been taken this idea further by developing a methodology to assess children's reasoning strategies based on error patterns for a carefully designed series of items, the rule-assessment methodology (RAM). To illustrate the importance of accounting for heterogeneity in strategy use, I will first explain RAM and mention a few cognitive domains for which strategies were detected. Subsequently, I will discuss pitfalls of RAM and the alternative of latent variable techniques.

2 Strategies in cognition

Children importantly improve their reasoning abilities from toddlerhood to adolescence. For some reasoning domains this improvement appears to consist of the application of increasingly complex strategies. Henceforth I will use the general definition of strategy by Rickard (2004, p. 65):

> "the term strategy is used merely to denote a unique series of mental steps toward a solution and does not necessarily have direct implications regarding

intention or awareness".

A benchmark domain for cognitive strategies in development, is reasoning about torque (or moment of force) as measured with the balance scale task [17]. The balance scale task tests for the understanding of the way two physical variables, weight of the blocks placed at both sides of a balance and their distance to the fulcrum, relate to the balance of the balance scale. Children typically develop a more advanced understanding of a balance scale during childhood, before they are explicitly taught the rule of torque. At first, children only consider weights at both sides of the balance to be important. At a later stage, children do take distance of blocks into account but only when weights are equal at both sides. At a next stage children combine the variables, weight and distance in more complex ways when reasoning about torque [39]. These reasoning strategies are defined as a series of mental steps and are categorically different. That is, children do rarely show behavior in between these strategies.

The balance scale is a benchmark task, but strategy detection was done for many more cognitive developmental domains. Below, I will give a list (very much inexhaustive because it is mostly but not exclusively related to our lab) of cognitive domains for which age-related increasingly complex strategies were detected: proportional reasoning tasks, such as the balance scale task [39,18], the shadow-size task [33], and the buoyancy task [35,10], mental models of physical and biological phenomena [41,11,8], causal reasoning [32], recursive reasoning [29], free classification [26], card sorting [47,2], analogical reasoning [14], working memory [19], transitive reasoning [3], decision making [16], and category learning [34,43].

In addition to cross-sectional studies, the literature shows that heterogeneity in strategy use is also present for many domains within young adulthood. The cognitive domains for which strategy-use was found include (again a list which is very much inexhaustive and mostly but not exclusively related to our lab): working memory [31], category learning [22,40,29], feedback learning [1,24], phoneme perception [46], and artificial grammar learning [44].

So, in a wide range of cognitive domains heterogeneity of performance is best described as age-related strategy use. This finding is not limited to wide age ranges, or to strategies that can be expressed verbally [44,46].

3 Strategies and brain activity

A popular way of studying human cognition is searching for the brain areas that show task-specific activity. In these studies, cognitive behavior is related to specific brain areas by functional MRI (magnetic resonance imaging) studies. The most common way to look at the brain-behavior relation is to contrast brain activity during different events. For example, in a study with a simple feedback-learning task, brain activity after positive feedback is subtracted from brain activity after negative feedback to detect areas that are specifically active after negative feedback [6]. Individual differences in the brain-behavior

relation are mostly studied by correlating performance (e.g., the number of successful learning events) with specific activity (e.g., negative feedback related activity) for specific brain areas or the whole brain [6]. A different, less common approach of studying individual differences is by analyzing single-trial brain activity over time (e.g., [7]).

In developmental research one typically studies whether task-related activity in specific brain areas is related to age. The suggestion is that maturation of these areas explains improvement of performance with development. For example, Crone et al. [6] conclude:

> "These findings demonstrate that changes in separable neural systems underlie developmental differences in flexible performance adjustment."

A different way of analyzing these brain behavior data is by applying strategy analysis to the behavioral data first before relating behavior to imaging data. For example, we performed a strategy analysis of the trial-by-trial responses of a simple feedback-learning task in participants between 8 and 25 years old [24]. It appeared that we could distinguish four different strategies, with different levels of efficiency in feedback learning. Strikingly, strategies were distributed over age groups, such that adults were part of the three most advanced strategy groups and children were part of the three less advanced strategy groups. Although performance was strongly related to age, there was thus a great overlap between children, adolescents and adults in strategy use. The next step in the analysis was relating strategy use to task-related brain activity in specific areas. A mediation analysis showed that for specific brain areas, for example the dorsal lateral prefrontal cortex, the strong correlations between age and activity were largely mediated by strategy use.

This type of studies highlights that if one is interested in the direct relation between performance and neural activity studying individual differences in terms of strategy use is important (see also [5]).

4 How to detect latent strategies?

The data from which strategies need to be detected are typically responses to a series of multiple-choice items. To this end, Siegler [36] defined the Rule Assessment Methodology (RAM). RAM starts with defining the strategies that participants are expected to use (Step 1). Next, a set of items is constructed that optimally discriminates between those expected strategies (Step 2). Finally, the data are matched with the expected response patterns (Step 3). For each observed response pattern the strategy with the best match is selected provided that observed and expected response patterns match sufficiently. The criterion for a sufficient match is laid down arbitrarily, independent of the data. For example, the general idea that Siegler and Chen [37] used to attribute a rule to a child is that at least 78% of the observed responses (i.e., for 7 out of 9 items), is consistent with a rule. Unfortunately, there are several drawbacks of this method that are nicely illustrated in [21]. First, RAM does not provide statistical grounds to decide whether the data should be regarded as a continuous

variation in responses or a limited number of categorically different response patterns. This appeared to be an issue when analyzing responses of a computational model of learning the balance scale task [27,25]. Applying RAM, it was wrongly concluded that the behavior of the computational model agrees with applying increasingly complex strategies. Second, unexpected strategies cannot be detected with RAM, because strategies need to be defined before they can be detected. For the balance scale task, for example, more advanced statistical techniques made it possible to detect the so-called sum strategy or buggy rule, that is, comparing the sum of weight and distance at both sides of the balance scale, instead of the product [18]. Unexpected strategies were also found for the shadow size task [33]. Third, the criterion for a match between an observed response pattern and a strategy is chosen arbitrarily in RAM and independent of the data. If strategies are applied with very little error (by adults, for example), the criterion should be very strict for an optimal assignment of strategies. However, young children might apply strategies with many mistakes. In this case a strict criterion would result in a suboptimal assignment of strategies and thus unnecessary many misclassifications. With simulation studies Van der Maas and Straatemeier [21] illustrate this issue.

A first indication of the presence of strategies is given by the frequency distribution of the data, such as the sum scores of a homogeneous subset of items. If the distribution shows multiple modes instead of one single mode, the presence of categorically different response patterns is very likely. Latent variable techniques provide a statistically grounded alternative to pattern matching to detect strategies [18]. Model selection techniques aid to decide on the number of present strategies and whether strategies are present at all. Based on a latent variable model that fits the data, observed response data are assigned to categorically different strategies.

Different types of latent variable models are available for different types of data. In the case the data consist of responses to a series of multiple-choice items, latent class analysis [23] is the technique to apply. But additional techniques are available for more complex data. For example, if the data consist of repeated measurements with the same set of items, latent Markov analysis could also estimate the transitions between strategies over time. In this way one could detect strategy changes during learning [29]. If the application of multiple strategies is mixed within individuals during the course of a task the detecting of strategies is even more challenging. Van Maanen, De Jong, and Van Rijn [20] provide an interesting approach for this situation by modeling reaction time distributions. In case the items have an internal structure that is supposed to be important, for example items of a task for floating and sinking systematically differ in weight and volume [10,44], latent regression analysis [15] could be used to detect strategies. Finally, there are also techniques available to analyze continuous instead of nominal responses. The package of depmixS4 [45] is available in the R-computing environment to use these latent variable techniques, but there are other packages available as well (e.g., [42]).

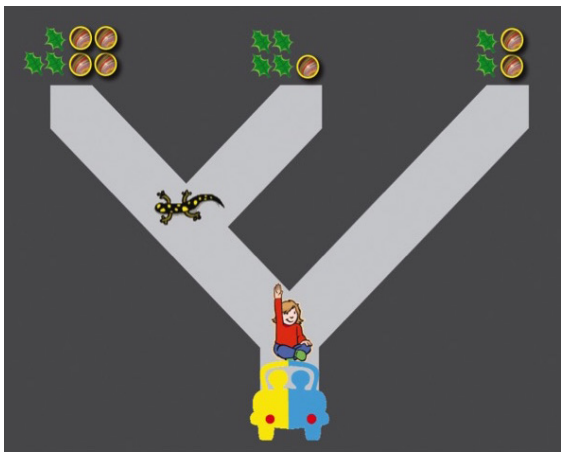


Figure 3. A first order-reasoning item from the traveling game, which is based on Flobbe et al. [9]. While seeing items like this, the child is told that it travels with a lizard in a car to collect marbles. The goal is to collect as many marbles as possible. However, at each cross, either the child or the lizard (as indicated) could choose which way to go. The lizard is very smart and plays as good as possible accounting for smart choices made by the child. (from [28]).

5 An example: Strategies in zero, first and second-order reasoning [1]

In a developmental study we were interested in how children advanced in playing strategic games. The strategic game we adapted for this reason was first used by Hedden and Zhang [13] and was later studied in a developmental context by Flobbe et al. [9]. Here, by way of illustration, I will only present the methodology for detecting strategies for solving first-order reasoning items. Figure 1 explains our implementation of the game. The methodology for detecting latent strategies consists of 3 steps: 1) Proposing plausible strategies (possibly we find additional strategies); 2) designing diagnostic items; 3) analyzing children's response patterns.

5.1 Step 1: Defining possible strategies

S1: The optimal strategy for playing first-order items like displayed in Figure 1 is to account for the optimal choices of the lizard, which results in a choice for going right at the first junction.

When considering relevant weaknesses of children's cognitive abilities or wrong interpretations of the task one could expect alternative, suboptimal strategies.

[1] Adapted from from [28].

Strategy	Priors	Conditional Probabilities		
		I1	I2	I3
S1: Optimal	.38	.94	.94	.94
S2: 0-order	.19	.96	.04	.96
S4: Go right	.04	.03	.97	.03
Guess	.39	.79	.47	.62

Table 5. Latent class model of responses to first-order items. The table shows the estimated parameters of the latent class model of responses to first-order items from 129 children (5—12 years of age). Priors indicate the class size; conditional probabilities indicate the accuracy of an item type given that one belongs to a specific class. The left-most column shows a possible interpretation of the strategy.

S2: A zero-order strategy. In Figure 1 this would results in choice for going left at the first junction, because this road leads to the largest number of marbles if not accounting for the lizard's choice.

S3: As Flobbe et al. [9] found, one could also make a choice to get the largest relative gain (number of marbles divided by the sum of marbles and leaves) instead of the absolute gain (number of marbles). In the case of Figure 1 this would result in a choice for going right.

S4: One could also avoid uncertainties and thus always choose the site that directly leads to the marbles (avoiding a choice by the lizard). That is, going right in the case of Figure 1.

Actually, more strategies were predicted but these unnecessarily complicate the example (see [28], for more details).

5.2 Step 2: Designing diagnostic items

The item displayed in Figure 1 is diagnostic for S2, but S1, S3, and S4 predict the same response, that is, "go right". Hence, items with different characteristics need to be designed in order to distinguish all strategies. To this end, we designed three item types:

I1: Items that are solvable by zero-order strategies and the correct response is going left.

I2: items with the correct choice to the right, while the largest number of marbles is at the left side, as in Figure 1.

I3: Items that are correctly answered by choosing the absolute gain and going left, but that would be incorrectly answered if one would choose the relative gain instead of the absolute gain.

5.3 Step 3: Latent Class Analysis

Latent class analysis was applied to the sum scores of item types I1, I2 and I3. The model displayed in Table 1 is the resulting best fitting, most parsimonious model. Four different classes were found, from which three are consistent with

one of the expected strategies, S1, S2, and S4. The consistency of applying one of these strategies was very high (at least .94). S3, optimizing the relative gain was not found in our version of the strategic game. In addition to the expected strategies, we found one response pattern that was not easy to interpret but could be guessing. In conclusion, this example shows how one could construct test items that are diagnostic to the expected strategies and with this test one could even detect a subgroup that was not anticipated.

6 Conclusion

The conclusion from this line of research on strategies use is that humans apply categorically different strategies to perform on the same cognitive task. This is the case for higher order tasks, such as playing strategic games, proportional reasoning, feedback learning, rule-application, etc. For these tasks, participants are mostly able to express their strategy in words. Remarkably, categorically different strategies were also detected for tasks that are believed to involve implicit memory, such as implicit learning tasks, and phoneme perception.

Most research involving strategy detection concerns developmental studies, and hence age-related application of strategies. Hence, age-performance relations are largely explained by strategy use. Nevertheless, individual differences within age groups are large. The most advanced strategies that are applied by a sub group of six-years old children also exist among twelve-years olds. The same counts for twelve-years old children and young adults. Moreover, studies with adults only show that also among adults individual differences in behavior are sometimes best described as strategies.

What is the consequence of this observation for research on cognition? Most importantly, the average behavior might give a wrong representation of what individual participants actually do. A typical developmental question is whether children of a certain age are able to solve problems of a certain difficulty. For example, are toddlers able to do causal reasoning [12] or can children do second-order recursive reasoning [9]? Group averages of scores above chance are taken as evidence that children are able to perform the reasoning. However, there are at least two reasons to be cautious with this conclusion. First, if children within the group apply different strategies, maybe a small subgroup of children is able to perform the reasoning and the rest are applying different strategies. This was found for the case of causal reasoning in two years old children [32]. Second, (a subgroup of) children may apply a strategy that results in above chance performance, which is nevertheless qualitatively different from the expected reasoning strategy. For example, latent class analysis of balance scale data showed that many children use a sum strategy instead of applying the torque principle in solving balance scale problems [18]. Hence, if one is interested in the performance on cognitive tasks a careful strategy analysis of behavior is important.

Two specific types of studies once more illustrate the importance of strategy analysis. First, the performance on a behavioral task is an important issue in relating behavior to brain activity. After all, the idea of drawing this relation is

that the brain activity reflects the mental steps that are executed by the individuals. Therefore, in comparing brain activity and behavior one would like to compare homogeneous strategy groups [4]. Second, also for comparing simulation models with human behavior, the strategy analysis of human behavior might be important [30]. The idea of defining a cognitive model is that performance of individuals is modeled, and not only that average behavior of the model agrees with average human behavior. Hence, the level of latent strategies is also an important level of analysis in comparing models and human behavior.

Therefore, I would conclude that the analysis of heterogeneity of behavioral data in terms of categorically different strategies is important for cognitive science in general.

References

[1] Andersen, L. M., I. Visser, E. A. Crone, P. C. Koolschijn and M. E. Raijmakers, *Cognitive strategy use as an index of developmental differences in neural responses to feedback*, Developmental Psychology (2014).

[2] van Bers, B. M., I. Visser, T. J. van Schijndel, D. J. Mandell and M. E. Raijmakers, *The dynamics of development on the Dimensional Change Card Sorting task*, Developmental Science **14** (2011), pp. 960–971.

[3] Bouwmeester, S., J. K. Vermunt and K. Sijtsma, *Development and individual differences in transitive reasoning: A fuzzy trace theory approach*, Developmental Review **27** (2007), pp. 41–74.

[4] Brown, T. T., H. M. Lugar, R. S. Coalson, F. M. Miezin, S. E. Petersen and B. L. Schlaggar, *Developmental changes in human cerebral functional organization for word generation*, Cerebral Cortex **15** (2005), pp. 275–290.

[5] Church, J. A., S. E. Petersen and B. L. Schlaggar, *The "Task B problem" and other considerations in developmental functional neuroimaging*, Human brain mapping **31** (2010), pp. 852–862.

[6] Crone, E. A., K. Zanolie, L. Van Leijenhorst, P. M. Westenberg and S. A. Rombouts, *Neural mechanisms supporting flexible performance adjustment during development*, Cognitive, Affective, & Behavioral Neuroscience **8** (2008), pp. 165–177.

[7] Duann, J.-R., T.-P. Jung, W.-J. Kuo, T.-C. Yeh, S. Makeig, J.-C. Hsieh and T. J. Sejnowski, *Single-trial variability in event-related BOLD signals*, Neuroimage **15** (2002), pp. 823–835.

[8] van Es, S. E., T. J. van Schijndel, R. Franse and M. E. Raijmakers, *Children's thoughts on unborn babies: Representational redescription in preconceptions of children on fetal development*, in: *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009, pp. 112–117.

[9] Flobbe, L., R. Verbrugge, P. Hendriks and I. Krämer, *Children's application of theory of mind in reasoning and language*, Journal of Logic, Language and Information **17** (2008), pp. 417–442.

[10] Franse, R., T. van Schijndel and M. Raijmakers, *Children's thinking about sinking: How predictions tell a different story than justifications* (2014), presentation at the 44th Annual Conference of the Jean Piaget Society, San Francisco.

[11] Frappart, S., M. Raijmakers and V. Frède, *What do children know and understand about universal gravitation? structural and developmental aspects*, Journal of Experimental Child Psychology **120** (2014), pp. 17–38.

[12] Gopnik, A., C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir and D. Danks, *A theory of causal learning in children: Causal maps and bayes nets*, Psychological Review **111** (2004), pp. 3–32.

[13] Hedden, T. and J. Zhang, *What do you think i think you think?: Strategic reasoning in matrix games*, Cognition **85** (2002), pp. 1–36.

[14] Hosenfeld, B., H. L. van der Maas and D. C. van den Boom, *Detecting bimodality in the analogical reasoning performance of elementary schoolchildren*, International Journal of Behavioral Development **20** (1997), pp. 529–547.

[15] Huang, G.-H. and K. Bandeen-Roche, *Building an identifiable latent class model with covariate effects on underlying and measured variables*, Psychometrika **69** (2004), pp. 5–32.

[16] Huizenga, H. M., E. A. Crone and B. J. Jansen, *Decision-making in healthy children, adolescents and adults explained by the use of increasingly complex proportional reasoning rules*, Developmental Science **10** (2007), pp. 814–825.

[17] Inhelder, B. and J. Piaget, "The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures," Basic Books, New York, 1958.

[18] Jansen, B. R. and H. L. van der Maas, *Statistical test of the rule assessment methodology by latent class analysis*, Developmental Review **17** (1997), pp. 321–357.

[19] Koppenol-Gonzalez, G. V., S. Bouwmeester and J. K. Vermunt, *The development of verbal and visual working memory processes: A latent variable approach*, Journal of Experimental Child Psychology **111** (2012), pp. 439–454.

[20] van Maanen, L., R. de Jong and H. van Rijn, *How to assess the existence of competing strategies in cognitive tasks: A primer on the fixed-point property*, PloS one **9** (2014), p. e106113.

[21] van der Maas, H. L. and M. Straatemeier, *How to detect cognitive strategies: Commentary on 'Differentiation and integration: Guiding principles for analyzing cognitive change'*, Developmental Science **11** (2008), pp. 449–453.

[22] Maddox, W. T., J. Pacheco, M. Reeves, B. Zhu and D. M. Schnyer, *Rule-based and information-integration category learning in normal aging*, Neuropsychologia **48** (2010), pp. 2998–3008.

[23] McCutcheon, A. L., "Latent Class Analysis," Sage, Beverly Hills, 1987.

[24] Peters, S., P. C. M. Koolschijn, E. A. Crone, A. C. Van Duijvenvoorde and M. E. Raijmakers, *Strategies influence neural activity for feedback learning across child and adolescent development*, Neuropsychologia **62** (2014), pp. 365–374.

[25] Quinlan, P. T., H. L. van der Maas, B. R. Jansen, O. Booij and M. Rendell, *Re-thinking stages of cognitive development: An appraisal of connectionist models of the balance scale task*, Cognition **103** (2007), pp. 413–459.

[26] Raijmakers, M. E., B. R. Jansen and H. L. van der Maas, *Rules and development in triad classification task performance*, Developmental Review **24** (2004), pp. 289–321.

[27] Raijmakers, M. E., S. van Koten and P. Molenaar, *On the validity of simulating stagewise development by means of PDP networks: Application of catastrophe analysis and an experimental test of rule-like network performance*, Cognitive Science **20** (1996), pp. 101–136.

[28] Raijmakers, M. E., D. J. Mandell, S. E. van Es and M. Counihan, *Children's strategy use when playing strategic games*, Synthese **191** (2014), pp. 355–370.

[29] Raijmakers, M. E., V. D. Schmittmann and I. Visser, *Costs and benefits of automatization in category learning of ill-defined rules*, Cognitive psychology **69** (2014), pp. 1–24.

[30] van Rijn, H., M. van Someren and H. van der Maas, *Modeling developmental transitions on the balance scale task*, Cognitive Science **27** (2003), pp. 227–257.

[31] Sanfratello, L., A. Caprihan, J. M. Stephen, J. E. Knoefel, J. C. Adair, C. Qualls, S. L. Lundy and C. J. Aine, *Same task, different strategies: How brain networks can be influenced by memory strategy*, Human Brain Mapping **35** (2014), pp. 5127–5140.

[32] van Schijndel, T. J. P. and M. E. J. Raijmakers, *Causal reasoning in 2 to 5-year-olds: Individual differences in strategies for responding to causal reasoning problems*, Biennial Meeting of the Society for Research in Child Development, Denver 2009 (2009).

[33] van Schijndel, T. J., I. Visser, B. M. van Bers and M. E. Raijmakers, *Preschoolers perform more informative experiments after observing theory-violating evidence*, Journal of Experimental Child Psychology **131** (2015), pp. 104–119.

[34] Schmittmann, V. D., H. L. van der Maas and M. E. Raijmakers, *Distinct discrimination learning strategies and their relation with spatial memory and attentional control in 4-to 14-year-olds*, Journal of Experimental Child Psychology **111** (2012), pp. 644–662.

[35] Schneider, M. and I. Hardy, *Profiles of inconsistent knowledge in children's pathways of conceptual change.*, Developmental Psychology **49** (2013), pp. 1639–1649.

[36] Siegler, R. S., *Cognition, instruction, development, and individual differences*, in: A. M. Lesgold, J. W. Pellegrino, S. Fokkema and R. Glaser, editors, *Cognitive Psychology and Instruction*, Springer, 1978 pp. 389–403.
[37] Siegler, R. S. and Z. Chen, *Differentiation and integration: Guiding principles for analyzing cognitive change*, Developmental Science **11** (2008), pp. 433–448.
[38] Siegler, R. S. and S. Ellis, *Piaget on childhood*, Psychological Science (1996), pp. 211–215.
[39] Siegler, R. S., S. Strauss and I. Levin, *Developmental sequences within and between concepts*, Monographs of the society for research in child development (1981), pp. 1–84.
[40] Speekenbrink, M., D. A. Lagnado, L. Wilkinson, M. Jahanshahi and D. R. Shanks, *Models of probabilistic category learning in Parkinson's disease: Strategy use and the effects of L-dopa*, Journal of Mathematical Psychology **54** (2010), pp. 123–136.
[41] Straatemeier, M., H. L. van der Maas and B. R. Jansen, *Children's knowledge of the earth: A new methodological and statistical approach*, Journal of Experimental Child Psychology **100** (2008), pp. 276–296.
[42] Vermunt, J. K. and J. Magidson, *Latent GOLD 4.0 user's guide* (2005), statistical Innovations, Belmont, MA. Retrieved February 23, 2010 from `http://www.statisticalinnovations.com/index.html`.
[43] Visser, I. and M. E. Raijmakers, *Developing representations of compound stimuli*, Frontiers in Psychology **3** (2012).
[44] Visser, I., M. E. Raijmakers and E. M. Pothos, *Individual strategies in artificial grammar learning*, The American Journal of Psychology (2009), pp. 293–307.
[45] Visser, I. and M. Speekenbrink, *depmixS4: An R-package for hidden Markov models*, Journal of Statistical Software **36** (2010), pp. 1–21.
[46] Wanrooij, K., P. Escudero and M. E. Raijmakers, *What do listeners learn from exposure to a vowel distribution? an analysis of listening strategies in distributional learning*, Journal of Phonetics **41** (2013), pp. 307–319.
[47] Zelazo, P. D., *The dimensional change card sort (DCCS): A method of assessing executive function in children*, Nature Protocols **1** (2006), pp. 297–301.

Infinitary Hybrid Logic and the Lindelöf Property

Gerard R. Renardel de Lavalette [1]

Johann Bernoulli Institute for Mathematics and Computer Science
University of Groningen

Abstract

This short note is about hybrid logic with infinitary rules, interpreted in named models. It is shown that there is a rule κ such that the property 'entails κ' is not Lindelöf: if some collection of rules $\mathcal{R}$ entails κ, then there is not always a countable subset of $\mathcal{R}$ that entails κ.

Keywords: Hybrid logic, infinitary rules, named models, Lindelöf.

1 Introduction

This paper is a modest contribution to the theory of hybrid logic. It grew from the collaboration of Rineke Verbrugge, Barteld Kooi and the present author investigating strong completeness (the property $\Gamma \models \varphi \;\Rightarrow\; \Gamma \vdash \varphi$) for non-compact logics. Recall that, in compact logics where $\Gamma \models \varphi$ implies that $\Gamma_0 \models \varphi$ for some finite $\Gamma_0 \subseteq \Gamma$, strong completeness directly follows from ordinary completeness, where Γ is empty. This collaboration led to the publications [5] and [6] on infinitary propositional dynamic logic, and [2], [3] on Khyb_ω, infinitary hybrid logic. In [3] we demonstrate that any extension of Khyb_ω with $\mathcal{R}$, a countable collection of infinitary rules, is strongly complete. Our methods and results are inspired by those in [7], [12]. With this general result, we obtained strong completeness for several non-compact hybrid logics: ancestral logic, reachability logic, cycle logic and BCC (bounded chain condition) logic.

Then we tried to extend our result to hybrid provability logic, which is characterized by conversely wellfounded frames (a definition is given in Section 4). These frames can be characterized by an uncountable collection of rules. We found no way to extend the strong completeness result to uncountable sets of rules, so we wondered whether there might be a countable set of rules that characterizes conversely wellfounded frames. We have not found such a set, but we discovered the following property:

$$\text{there are a rule } \kappa \text{ and a collection of rules } \mathcal{R} \text{ such that } \mathcal{R} \models^{\mathsf{n}} \kappa, \quad \text{but there is no countable subset } \mathcal{R}_0 \subseteq \mathcal{R} \text{ with } \mathcal{R}_0 \models^{\mathsf{n}} \kappa. \tag{11}$$

[1] Email: g.r.renardel.de.lavalette@rug.nl

Here $\models^{\mathsf{n}}$ denotes entailment in named models. We shall prove (11) in Section 5.

Observe that (11) looks like a compactness property where 'finite' is replaced by 'countable'. In topology, this weakening of the compactness property (every open cover has a finite subcover) is called the Lindelöf property, named after the Finnish mathematician Ernst Leonard Lindelöf (1870 - 1946). A well-known example of a Lindelöf space is $\mathbb{R}$, the real line. We generalize this as follows. A property P of sets X is called *Lindelöf* whenever $P(X)$ implies $P(X_0)$ for some countable subset $X_0 \subseteq X$. Now we may paraphrase (11) by: in Khyb_ω, the property $\mathcal{R} \models^{\mathsf{n}} \kappa$ of rule sets $\mathcal{R}$ is not Lindelöf.

A short overview of hybrid logic is given in Section 2. In Section 3 we define the language, semantics and proof system of Khyb_ω, hybrid logic with infinitary rules. Section 4 contains several results about the characterization of conversely wellfounded models and frames, which lead to the main result in Section 5. We end with a concluding remark in Section 6.

2 What is hybrid logic?

Hybrid logic is an extension of modal logic with propositional variables called *nominals* i that refer to possible worlds. As an atomic formula, i is true in exactly one possible world. Moreover, a nominal i and a formula φ can be combined with the satisfaction operator to form the formula $\mathsf{i} : \varphi$ with the intended meaning "φ holds at the world named i". Nowadays, this is written as $@_{\mathsf{i}}\varphi$ using the *at*-operator $@$. This extension was elaborated for the first time in the context of tense logic by Arthur Prior in the late 1960s: see [11]. Almost 20 years later, hybrid logic was reinvented by Solomon Passy and Tinko Tinchev in their paper [10] on Propositional Dynamic Logic. Another ten years later, Valentin Goranko introduced in [8] the nominal binder $\downarrow\mathsf{i}$ that binds i to the actual world: $\downarrow\mathsf{i}\varphi$ means "φ holds when i is interpreted as the actual world".

Hybrid logic is an active area of research. See [1] for an overview, and [4] for more information about its proof theory.

3 Khyb_ω, hybrid logic with infinitary rules

Given a countable set of propositional variables $p \in P$, and a countably infinite set of nominals $\mathsf{i} \in \mathsf{I}$, we define the language of hybrid logic by

$$\varphi ::= \bot \mid p \mid \mathsf{i} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \Box\varphi \mid @_{\mathsf{i}}\varphi \mid \downarrow\mathsf{i}\varphi$$

In $\downarrow\mathsf{i}\varphi$, $\downarrow\mathsf{i}$ binds all free occurrences of nominal i in φ. $\mathsf{fnom}(\varphi)$ denotes the collection of free nominals in φ.

We extend the language with rules. A *rule* is an expression of the form Γ/φ, where Γ is a (possibly infinite) collection of formulas and φ is a formula. We use $\mathcal{R}$ to range over sets of rules.

3.1 Semantics

A model M for hybrid logic is a model for modal logic extended with a valuation for the nominals. So $M = (W, R, V, A)$, with possible worlds in $W \neq \emptyset$,

accessibility relation $R \subseteq W \times W$, propositional valuation $V : P \to \wp(W)$ and nominal valuation $A : \mathsf{I} \to W$. $\langle W, R\rangle$ is called the *frame* of M. A *named model* is a model where every world has a name, i.e. the nominal valuation A is surjective. We write MOD for the class of models, and NMOD for the class of named models.

The interpretation of formulae is defined as follows:

$$
\begin{array}{ll}
(M,w) \not\models \bot & \\
(M,w) \models p & \text{iff } w \in V(p) \\
(M,w) \models \mathsf{i} & \text{iff } w = A(\mathsf{i}) \\
(M,w) \models \neg\varphi & \text{iff } (M,w) \not\models \varphi \\
(M,w) \models \varphi \wedge \psi & \text{iff } (M,w) \models \varphi \text{ and } (M,w) \models \psi \\
(M,w) \models \Box\varphi & \text{iff } (M,v) \models \varphi \text{ for all } v \text{ with } (w,v) \in R \\
(M,w) \models @_{\mathsf{i}}\varphi & \text{iff } (M, A(\mathsf{i})) \models \varphi \\
(M,w) \models \downarrow\mathsf{i}\varphi & \text{iff } (M[\mathsf{i} := w], w) \models \varphi
\end{array}
$$

Here $M[\mathsf{i} := w]$ is M with nominal valuation A replaced by $A[\mathsf{i} := w]$ which sends i to w and other nominals $\mathsf{j} \neq \mathsf{i}$ to $A(\mathsf{j})$.

We extend the definition of the entailment relation $\models$ as follows:

$$
\begin{array}{ll}
(M,w) \models \Gamma & \equiv (M,w) \models \varphi \text{ for all } \varphi \in \Gamma \\
M \models \Gamma/\varphi & \equiv (M,w) \models \Gamma \;\Rightarrow\; (M,w) \models \varphi \text{ for all } w \in W_M \\
\langle W, R\rangle \models \Gamma/\varphi & \equiv M \models \Gamma/\varphi \text{ for all models } M = \langle W, R, V, A\rangle \\
\Gamma \models \varphi & \equiv M \models \Gamma/\varphi \text{ for all } M \in \mathsf{MOD} \\
M \models \mathcal{R} & \equiv M \models \Gamma/\varphi \text{ for all } \Gamma/\varphi \in \mathcal{R} \\
\mathcal{R} \models \Gamma/\varphi & \equiv M \models \mathcal{R} \;\Rightarrow\; M \models \Gamma/\varphi \text{ for all } M \in \mathsf{MOD} \\
\mathcal{R} \models^{\mathsf{n}} \Gamma/\varphi & \equiv M \models \mathcal{R} \;\Rightarrow\; M \models \Gamma/\varphi \text{ for all } M \in \mathsf{NMOD}
\end{array}
$$

3.2 Proof system

The proof system Khyb_ω is based on sequents of the form $\Gamma \vdash \varphi$. The axioms and rules are

Taut	$\vdash \varphi$ if φ is an instance of a propositional tautology	
MP	$\varphi, \varphi \to \psi \vdash \psi$	(modus ponens)
$\mathbf{K}_\Box$	$\vdash \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$	(distribution)
$\mathbf{K}_@$	$\vdash @_{\mathsf{i}}(\varphi \to \psi) \to (@_{\mathsf{i}}\varphi \to @_{\mathsf{i}}\psi)$	(distribution)
$\mathbf{SD}_@$	$\vdash @_{\mathsf{i}}\varphi \to \neg@_{\mathsf{i}}\neg\varphi$	(self-dual)
Intr	$\vdash \mathsf{i} \wedge \varphi \to @_{\mathsf{i}}\varphi$	(introduction)
$\mathbf{T}_@$	$\vdash @_{\mathsf{i}}\mathsf{i}$	(reflexivity)
Agree	$\vdash @_{\mathsf{i}}@_{\mathsf{j}}\varphi \leftrightarrow @_{\mathsf{j}}\varphi$	(agree)
Back	$\vdash \Diamond@_{\mathsf{i}}\varphi \to @_{\mathsf{i}}\varphi$	(back)
DA	$\mathsf{i} \vdash \downarrow\mathsf{j}\varphi \leftrightarrow \varphi[\mathsf{j} := \mathsf{i}]$	(downarrow)

Name	$\vdash \downarrow i @_i \varphi \to \varphi$ provided $i \notin \mathsf{fnom}(\varphi)$	(name)
BG	$i \vdash \Box \downarrow j @_i \Diamond j$ provided $i \neq j$	(bounded generalization)
$\mathbf{SNec}_{\Box}$	if $\Gamma \vdash \varphi$, then $\Box\Gamma \vdash \Box\varphi$	(strong necessitation)
$\mathbf{SNec}_{@}$	if $\Gamma \vdash \varphi$ then $@_i\Gamma \vdash @_i\varphi$	(strong necessitation)
$\mathbf{SNec}_{\downarrow}$	if $\Gamma \vdash \varphi$ then $\downarrow i\Gamma \vdash \downarrow i\varphi$	(strong necessitation)
InfCut	if $\Gamma \vdash \Delta$ and $\Gamma', \Delta \vdash \varphi$ then $\Gamma, \Gamma' \vdash \varphi$	(infinitary cut)
Ded	if $\Gamma, \varphi \vdash \psi$ then $\Gamma \vdash \varphi \to \psi$	(deduction)

For $\mathcal{R}$ a rule set, $\mathsf{Khyb}_\omega + \mathcal{R}$ is defined straightforwardly as the extension of the proof system given above with sequents $\Delta \vdash \psi$ for all $\Delta/\psi \in \mathcal{R}$. We say that $\Gamma \vdash_{\mathcal{R}} \varphi$ holds if $\Gamma \vdash_{\mathcal{R}} \varphi$ is derivable in this extended proof system. In [3], we showed that Khyb_ω is sound, i.e. that $\Gamma \vdash \varphi$ implies $\Gamma \models \varphi$. Soundness easily transfers to all extensions of Khyb_ω with rule sets. We also proved that, for countable rule sets $\mathcal{R}$, we have strong completeness:

every $(\mathsf{Khyb}_\omega + \mathcal{R})$-consistent set of formulas is satisfiable in a named model in which $\mathcal{R}$ is valid.

Combining this with soundness, we have for all countable rule sets $\mathcal{R}$

$$\Gamma \vdash_{\mathcal{R}} \varphi \;\Leftrightarrow\; \Gamma \models_{\mathcal{R}} \varphi \;\Leftrightarrow\; \Gamma \models^{\mathsf{n}}_{\mathcal{R}} \varphi$$

4 Conversely wellfounded relations

Recall that a relation R on a set X is *conversely wellfounded* iff it has no infinite ascending paths:

$$\forall f \in X^{\mathbb{N}} \exists m\ (f(m), f(m+1)) \notin R$$

The next lemma shows that, for infinite X, $X^{\mathbb{N}}$ in this definition cannot be replaced by a countable collection of functions.

Lemma 4.1 *Let $\{f_n \mid n \in \mathbb{N}\} \subseteq \mathbb{N}^{\mathbb{N}}$. Then there are a relation $R \subseteq \mathbb{N}^2$ and a function $g \in \mathbb{N}^{\mathbb{N}}$ such that $\forall n \exists m\ (f_n(m), f_n(m+1)) \notin R$, but $\forall m\ (g(m), g(m+1)) \in R$. So R is not conversely well-founded.*

Proof. First we define

R admits f iff $\forall n\ (f(n), f(n+1)) \in R$
R blocks f iff f does not admit R, ie. $\exists n\ (f(n), f(n+1)) \notin R$

We will construct an injection $g \in \mathbb{N}^{\mathbb{N}}$ and $R = R_g = \{(g(m), g(m+1)) \mid m \in \mathbb{N}\}$ such that R admits g and blocks all f_n, which proves the lemma.
It is evident that R_g admits g by definition. We observe that the injectivity of g entails that

$$R_g \text{ admits } f \text{ iff } \exists k \forall m\ f(m) = g(m+k)$$

So to realize that R_g blocks all f_n, it suffices to have $\forall n \forall k \exists m f_n(m) \neq g(m+k)$ or equivalently

$$\forall n \forall k \exists p \geq k \; f_n(p-k) \neq g(p) \tag{12}$$

Let $\alpha : \mathbb{N}^2 \to \mathbb{N}$ be defined by $\alpha(n,k) = ((n+k)^2 + 3n + k)/2$. It is easily verified that α is a bijection, so there are inverses $\underline{,}\gamma$ with $\forall nk(\alpha(\underline{n}), \gamma(n)) = n \wedge (\alpha(n,k)) = n \wedge \gamma(\alpha(n,k)) = k)$. Moreover, we have $\forall nk \; \alpha(n,k) \geq k$, which implies $\forall n \; \gamma(n) \leq n$. Define g by

$$g(p) = \min(\mathbb{N} - \{g(k) \mid k < p\} - \{f_{\beta(p)}(p - \gamma(p))\})$$

Then g is injective and $\forall p \; f_{\beta(p)}(p - \gamma(p)) \neq g(p)$, so $\forall n \forall k \; f_n(\alpha(n,k) - k) \neq g(\alpha(n,k))$. Together with $\forall nk \; \alpha(n,k) \geq k$ this implies (12). This ends the proof. □

We define rules $\rho(f)$, for every $f \in \mathsf{I}^{\mathbb{N}}$:

$$\rho(f) = \{@_{f(n)} \Diamond f(n+1) \mid n \in \mathbb{N}\}/\bot$$

and the rule $\kappa(p)$:

$$\kappa(p) = \{\Box^n(\Box p \to p) \mid n \in \mathbb{N}\}/p$$

The next lemma indicates to what extent conversely wellfounded models and frames can be characterized by these rules. We shall use the Axiom of Dependent Choice:

$$X \neq \emptyset \;\&\; \forall x \in X \exists y \in X A(x,y) \to \exists f \in X^{\mathbb{N}} \forall n \; A(f(n), f(n+1)) \tag{13}$$

This axiom follows from the Axiom of Choice, and it implies the Axiom of Countable Choice.

Lemma 4.2 (i) $\{\rho(f) \mid f \in \mathsf{I}^{\mathbb{N}}\}$ *characterizes conversely wellfounded named models, i.e. for named models* M

$$M \models \{\rho(f) \mid f \in \mathsf{I}^{\mathbb{N}}\} \;\Leftrightarrow\; R_M \textit{ is conversely wellfounded}$$

(ii) *No countable subset of* $\{\rho(f) \mid f \in \mathsf{I}^{\mathbb{N}}\}$ *characterizes conversely wellfounded named models.*

(iii) $\kappa(p)$ *characterizes conversely wellfounded frames, i.e.*

$$\langle W, R\rangle \models \kappa(p) \;\Leftrightarrow\; R \textit{ is conversely wellfounded}$$

(iv) $\kappa(p)$ *does not characterize conversely wellfounded named models.*

Proof.

(i) Let $M = \langle W, R, V, A\rangle$ be a named model, so A is surjective. Then $\{A \circ f \mid f \in \mathsf{I}^{\mathbb{N}}\} = W^{\mathbb{N}}$, and also $M \models \rho(f) \;\Leftrightarrow\; \exists m \; (A(f(m)), A(f(m+1))) \notin R$. So $M \models \{\rho(f) \mid f \in \mathsf{I}^{\mathbb{N}}\}$ iff $\forall g \in W^{\mathbb{N}} \exists m (g(m), g(m+1)) \notin R$, i.e. iff R conversely wellfounded.

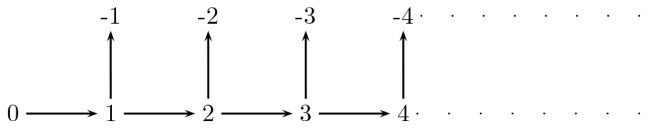


Figure 4. A not conversely wellfounded frame where $\kappa(\bot)$ holds.

(ii) Let $F \subseteq \mathsf{I}^{\mathbb{N}}$ be countable, and let $A : \mathsf{I} \to \mathbb{N}$ be bijective. Then there is a countable $G \subseteq \mathbb{N}^{\mathbb{N}}$ with $F = \{A^{-1} \circ g \mid g \in G\}$. By Lemma 4.1, there is an $R \subseteq \mathbb{N}^2$ that is not conversely wellfounded and blocks all $g \in G$. Then $M = \langle \mathbb{N}, R, V, A\rangle$ with arbitrary $V : P \to \wp(\mathbb{N})$ is a named model that is not conversely wellfounded such that $M \models \rho(f)$ for all $f \in F$.

(iii) We prove both directions of the equivalence via contraposition.

$\Rightarrow$: assume that R is not conversely wellfounded, so there is an $f \in W^{\mathbb{N}}$ with $\forall n\ (f(n), f(n+1)) \in R$. Let M be a model $\langle W, R, V, A\rangle$ with $V(p) = W - \mathsf{rg}(f)$. Now $(M, f(0)) \models \Box^n(\Box p \to p)$ evaluates to $\forall v \in \mathsf{rg}(f)((f(0), v) \in R^n \to \exists u \in \mathsf{rg}(f)\ (v, u) \in R)$, and this is true for all n. However, $(M, f(0)) \not\models p$. So there is a model M with frame $\langle W, R\rangle$ where $\kappa(p)$ does not hold, hence $\langle W, R\rangle \not\models \kappa(p)$.

$\Leftarrow$: assume there is a model $M = \langle W, R, V, A\rangle$ with $M \not\models \kappa(p)$. Then there is a $w \in W$ with $(M, w) \models \Box^n(\Box p \to p)$ for all n, and $(M, w) \not\models p$. This evaluates to $\forall v \in X \exists u \in X\ (v, u) \in R$ with $X = \{v \mid (w, v) \in R^* \ \&\ v \notin V(p)\}$, and $w \notin V(p)$, which implies $X \neq \emptyset$. Here R^* is, as usual, the reflexive transitive closure of R, i.e. the least relation that contains R and that is reflexive and transitive. Applying the Axiom of Dependent Choice (13) now yields an $f \in X^{\mathbb{N}}$ with $\forall n\ (f(n), f(n+1)) \in R$, so R is not conversely wellfounded.

(iv) Consider the named model $M = \langle \mathbb{Z}, R, V, A\rangle$ with $R = \{(k, k+1) \mid k \geq 0\} \cup \{(k, -k) \mid k > 0\}$, $V(p) = \emptyset$ and $A : \mathsf{I} \to W$ surjective. See Figure 4. It is clear that R is not conversely wellfounded, for there is an infinite chain $0, 1, 2, \ldots$ of R-steps. We shall show that $M \models \kappa(\bot)$, using that $\kappa(\bot)$ is equivalent to the rule $\{\Box^n \Diamond \top \mid n \in \mathbb{N}\}/\bot$. One easily verifies

$$\begin{array}{l} \text{if } w < 0 \text{ then } w \not\models \Diamond\top (= \Box^0\Diamond\top) \\ \text{if } w = 0 \text{ then } \forall n\ w \not\models \Box^{n+2}\Diamond\top \\ \text{if } w > 0 \text{ then } \forall n\ w \not\models \Box^{n+1}\Diamond\top \end{array}$$

So for all $w \in \mathbb{Z}$ there is a n with $(M, w) \not\models \Box^n\Diamond\top$. This implies $M \models \kappa(\bot)$, hence $M \models \kappa(p)$ (for $V(p) = \emptyset$).

□

5 The main result

Now the announced result is within reach.

Theorem 5.1 *The following hold:*

$$\{\rho(f) \mid f \in \mathsf{I}^{\mathbb{N}}\} \models^n \kappa(p) \tag{14}$$

$$\{\rho(f) \mid f \in F\} \not\models^n \kappa(p) \quad \textit{if } F \subseteq \mathsf{I}^{\mathbb{N}} \textit{ countable} \tag{15}$$

So the property $\mathcal{R} \models^n \kappa(p)$ of rule sets $\mathcal{R}$ is not Lindelöf.

Proof. By Lemma 4.2.(i) we have that $M \models \{\rho(f) \mid f \in \mathsf{I}^{\mathbb{N}}\}$ iff M is conversely wellfounded. By Lemma 4.2.(iii) we have that all conversely wellfounded models entail $\kappa(p)$. This proves (14).
By Lemma 4.2.(ii) there is a named model $M = \langle W, R, V, A\rangle$ with $M \models \{\rho(f) \mid f \in F\}$ and R not conversely wellfounded. By Lemma 4.2.(iii) there is a model $M' = \langle W, R, V', A'\rangle$ with frame $\langle W, R\rangle$ such that $M' \not\models \kappa(p)$. Now define $M'' = \langle W, R, V', A\rangle$. This is a named model, for A is surjective. We observe $M'' \models \{\rho(f) \mid f \in F\}$, for the rules $\rho(f)$ contain no propositional variables, and M, M'' only differ in the propositional valuation. Moreover, we have $M'' \not\models \kappa(p)$, for the rule $\kappa(p)$ contains no nominals and M', M'' only differ in the nominal valuation. This proves (15). □

6 Concluding remarks

We established a mildly surprising property: when going from finite hybrid logic to the infinitary system Khyb_ω, we do not only lose the compactness property but also the weaker Lindelöf property. Thus fails our second attempt to prove strong completeness for infinitary hybrid provability logic, i.e. $\mathsf{Khyb}_\omega+$ the following uncountable collection of rules:

$$\{\{@_{\mathsf{i}_n} \Diamond \mathsf{i}_{n+1} \mid n \in \mathbb{N}\}/\bot \mid \mathsf{i}_0, \mathsf{i}_1, \ldots \text{ a sequence of nominals}\} \tag{16}$$

As was mentioned in [3], the first naive attempt to prove this collapsed on the observation that there is no countable set of rules containing finitely many nominals that characterizes converse wellfoundedness. This observation follows from the embedding of Khyb_ω in $L_{\omega_1\omega}$, predicate logic with countably infinite conjunctions and disjunctions, and from the undefinability of wellordering in $L_{\omega_1\omega}$ (a consequence of Lopez-Escobar's undefinability result in [9]).

So we still do not know whether $\mathsf{Khyb}_\omega + (16)$ is strongly complete with respect to named conversely wellfounded models. A third attempt to prove this would be: extend the main result of [3] to arbitrary sets of rules, not only countable sets. In other words: prove that any extension of Khyb_ω with a (possibly uncountable) collection $\mathbb{R}$ of rules is strongly complete. We conjecture that this is true, but up to now we have not been able to find a proof.

References

[1] Areces, C. and B. ten Cate, *Hybrid logics*, in: P. Blackburn, J. van Benthem and F. Wolter, editors, *Handbook of Modal Logic*, Elsevier, 2007 pp. 821–868.

[2] Barteld Kooi, Gerard R. Renardel de Lavalette and Rineke Verbrugge, *Strong completeness for non-compact hybrid logics*, in: R. Schmidt, I. Pratt-Hartmann and M. Reynolds, editors, *AiML2004 — Advances in Modal Logic* (2004), pp. 212 – 223.

[3] Barteld Kooi, Gerard Renardel de Lavalette and Rineke Verbrugge, *Hybrid logics with infinitary proof systems*, Journal of Logic and Computation **16** (2006), pp. 161 – 175.

[4] Braüner, T., "Hybrid Logic and its Proof-Theory," Applied Logic Series **37**, Springer, 2011.

[5] Gerard R. Renardel de Lavalette, Barteld Kooi and Rineke Verbrugge, *A strongly complete proof system for propositional dynamic logic*, in: P. Balbiani, N.-Y. Suzuki and F. Wolter, editors, *AiML2002 — Advances in Modal Logic* (2002), pp. 377–393.

[6] Gerard Renardel de Lavalette, Barteld Kooi and Rineke Verbrugge, *Strong completeness and limited canonicity for PDL*, Journal of Logic, Language and Information **17** (2008), pp. 69–87, see also the Erratum in vol. 18 (2009) pp. 291–292.

[7] Goldblatt, R., "Mathematics of Modality," CSLI Lecture Notes **43**, CSLI Publications, Stanford, California, 1993.

[8] Goranko, V., *Temporal logic with reference pointers*, in: *in Proceedings of the 1st International Conference on Temporal Logic (Lecture Notes in Artificial Intelligence: Volume 827)* (1994), pp. 133–148.

[9] Lopez-Escobar, E. G. K., *On defining well-orderings*, Fundamenta Mathematicae **59** (1966), pp. 13–21, 299–300.

[10] Passy, S. and T. Tinchev, *Quantifiers in combinatory PDL: Completeness, definability, incompleteness*, in: *Fundamentals of Computation Theory FCT 85 (Lecture Notes in Computer Science, Volume 199)* (1985), pp. 512–519.

[11] Prior, A. N., "Past, Present and Future," Oxford University Press, 1967.

[12] Segerberg, K., *A model existence theorem in infinitary propositional modal logic*, Journal of Philosophical Logic **23** (1994), pp. 337–367.

Understanding Irony in Autism: The Role of Context and Prosody

Iris Scholten, Eerin Engelen & Petra Hendriks

University of Groningen

1 Introduction

Irony is a figure of speech that can be used to express the opposite of what is literally said. For example, the sentence "That was fun!" implies that the speaker had a great time, but when intended ironically (such as after a boring party) the same sentence expresses the exact opposite. The point of irony is to indicate that a proposition that a speaker may normally endorse is in fact not endorsed by the speaker, for example because it is false or might be unlikely given the situation [25]. To make sure the hearer will understand the ironic intention of the speaker, speakers can use cues to get their intention across. For example, they can use a specific facial expression or body language. In this study, we will focus on two linguistic cues: context and prosody.

The context in which something is said is considered to be one of the most important cues for the recognition of irony [15,7,10]. Context can set up a particular expectation, which is in conflict with the content of the ironic statement. This conflict can help the hearer to recognize the ironic intention of the statement. As an example, consider the following situation: John is on holiday. He discovers that his bags have gotten lost at the airport and the hotel he booked is full. When he says to Mary: "This must be my lucky day!," this utterance is so obviously in conflict with the context that Mary should normally be able to recognize the irony.

A second important cue is the prosody used in the ironically intended expression. According to [6], there is a typical prosody that implies the intention of irony and therefore could invite the recognition of irony by hearers. This typical prosody in ironic expressions involves two high peaks: one peak around the second word (usually the verb, in languages such as English and Dutch) and one peak at the end of the expression. There may be other prosodic features that are associated with an ironic intention, such as an exaggerated monotone intonation or overly enthusiastic exclamations, but these features might be harder to recognize [27].

An important factor in the use of irony is the ambiguity of the ironically intended expression. Even though context and prosody can provide very clear cues for the ironic interpretation, it is still possible that the hearer does not pick up on the ironic intention and instead interprets the expression literally. The clearer the discrepancy between the speaker's description of the situation and

the actual state of affairs, the easier it is to recognize the irony [25]. The same may be true for prosody: when the intonational pattern described above is used, it might become easier to recognize the ironic intention. If prosody is indeed a clear indicator for irony, context might not even be needed to recognize irony and prosody may be enough to understand the speaker's intention.

Understanding irony is a skill that seems particularly difficult for individuals with Autism Spectrum Disorder (ASD). ASD is a congenital neural developmental disorder that is characterized by qualitative deficits in social interaction and communication and by limited, repetitive or stereotypical behaviors, interests or activity patterns [2], ranging from mild to severe [3]. It is claimed that individuals with ASD do not have difficulties with language per se, but rather with the pragmatic functions of language [4,14,20,18]. As a consequence, they may have problems in taking advantage of the contextual cues that indicate indirect, figurative or ironic language use. Furthermore, they may have difficulties grasping the suprasegmental aspects of language, such as prosody, rhythm and accents [22]. These aspects of language are very important for the understanding of irony, since contextual and intonational cues contribute to the recognition of irony. A deficit in understanding these cues may therefore lead to problems in recognizing ironic intentions.

Various studies have shown difficulties in the understanding of irony in children and adults with ASD in comparison to typically developing peers [16,1,17,18,24]. According to [16], children with ASD are unable to recognize the ironic intention of the speaker when being asked about why someone says something. [17] found that children with ASD, when being asked for someone's true reason for saying something like "Great job!", often give a literal meaning or merely rephrase the expression produced.

Several explanations have been proposed for the difficulties children and adults with ASD have in recognizing irony. For example, these difficulties may be due to a deficit in Theory of Mind [5]. Theory of Mind (ToM) is the ability to understand and predict behavior based on one's own beliefs and the beliefs of others [26,21]. To understand irony, it is important that the hearer is able to apply higher-order ToM reasoning. First-order ToM reasoning (ToM-1) is the ability to attribute beliefs, thoughts and desires to someone else and to understand that these beliefs, thoughts and desires influence this person's behavior. For example, John is able to apply first-order ToM reasoning if he understands that Mary utters the sentence "That was fun!" when talking about a party because Mary believes the party was great. Higher-order reasoning (ToM-2 and further) involves the beliefs someone else has about another person and their predictions about this other person. For example, Mary is able to apply higher-order ToM reasoning if she understands that John believes that she believes the party was great. Thus, ToM-2 is needed to be aware of the fact that someone else has beliefs about you [23]. Crucially, the beliefs of this other person might be different from your own beliefs. That is, while John may believe that Mary believes the party was great, Mary might in fact have found the party quite boring. Understanding irony requires the ability to apply higher-order ToM reason-

ing because it requires understanding a thought about an attributed thought [11]. When Mary ironically says "That was fun!", she mentions a thought and at the same time expresses her attitude towards this thought. Understanding the irony in this utterance requires that the listener not only understands the thought, but also understands the speaker's attitude towards this thought.

Studies that tested children on a ToM-2 task and a separate irony task confirm the suggestion that the understanding of an ironic intention and ToM-2 reasoning are closely related [28,11,9]: children who fail on a ToM-2 task are also less capable of understanding ironic expressions. It is well-established that children with ASD have more difficulty with ToM reasoning than typically developing children (e.g., [5,11–13,4,9,19]). Therefore, their poor performance in understanding irony could be caused by their difficulties with ToM. Alternatively, their difficulties in understanding irony may also be due to their problems in understanding the cues for irony.

In contrast to children's understanding of irony, not much is known about adolescents' understanding of irony. Adolescents with ASD are expected to be linguistically more advanced than children with ASD. Does this mean that they are fully capable of using linguistic cues such as context and prosody to recognize the ironic intention of the speaker?

The present study aims to investigate whether adolescents with ASD are able to recognize and understand irony in the same way as their typically developing peers. In particular, we wish to find out whether they use the linguistic cues of context and prosody in the same way. To this end, we carried out an irony recognition task with a group of Dutch-speaking adolescents with ASD and a control group of typically developing adolescents, in which we manipulated context and prosody.

2 Methods

2.1 Participants

Thirteen adolescents with Autism Spectrum Disorders (mean age 15.5, age range 14–20, 10 male) were recruited from Scholengemeenschap De Ambelt in Zwolle (a school for secondary special education, cluster 4). The inclusion criteria for the sample were based on parental information about the clinical diagnosis, that was confirmed by the participants. One additional adolescent was tested but later excluded from the analysis because of lack of confirmation of the clinical diagnosis. There were no participants with a double diagnosis, such as the combination of ASD and ADHD. The control group consisted of fourteen typically developing adolescents (mean age 14.4, age range 11–20, 5 male); twelve of these adolescents were recruited through Scouting Group Don Bosco in Geldrop and two others were recruited through the researchers' personal network. All parents and/or caretakers of the participants gave written informed consent for their participation in this study.

2.2 Design and materials

The experiment manipulated two factors: context (inviting an ironic versus a non-ironic interpretation) and prosody (inviting an ironic versus a non-ironic interpretation), yielding a design with four conditions (see Table 1). There were eight items per condition, resulting in 32 items in total. From these items, four counterbalanced randomized lists were constructed. Each participant heard 16 items in total and 4 items per condition.

The 32 items consisted of short stories followed by three test questions. Each story started with an introductory sentence, followed by a concrete event that further specified the situation, and concluded with an evaluative statement. The experiment tested these items in four conditions: the neutral condition, the prosody condition, the context condition, and the combination condition. In the neutral condition, neither the story context nor the prosody of the evaluative statement invited an ironic interpretation of the evaluative statement. In the prosody condition, the evaluative statement had a prosody that invited an ironic interpretation, while the story context was compatible with a non-ironic interpretation of the evaluative statement. In the context condition, the story context invited an ironic interpretation of the evaluative statement, while the prosody of the evaluative statement was neutral. In the combination condition, finally, both the story context and the prosody of the evaluative statement invited an ironic interpretation of the evaluative statement.

Prosody was manipulated by distinguishing between two patterns of pronunciation for the evaluative statement at the end of the stories. In one pattern, the sentence was uttered with a typical ironic intonation in which there

	− Context	+ Context
− Prosody	*John's long-time wish is to get a scooter. Today is his birthday. When he enters the garage, he sees a brand new, shiny scooter. He says to his parents:* ***"What a great gift."***	*Sara has a job interview at the local grocery store today. She feels relaxed and responds to the questions very well. She gets hired. When she comes home, she says to her father:* ***"It went very badly."***
+ Prosody	*Tim spends the entire summer working in a clothing store. The customers constantly muddle the clothes and there is never a moment of relaxation. When his girlfriend stops by he says to her:* ***"This is the worst job ever."***	*Peter promises his wife to clean the house. When he tries to dust the mantelpiece he accidentally knocks over the favorite vase of his wife. It shatters into a thousand pieces as it hits the floor. His wife hears the noise and says:* ***"Great, well done."***

Table 6. Design of the experiment, with a sample item for each condition (translated from Dutch)

were two high peaks: one around the second word and one at the end of the sentence (see [6]). In the other pattern, the sentence was uttered with neutral intonation. The intonation patterns were verified using Praat, a computer software package to analyse speech, and were pre-tested with seven typically developing adults, who listened to the statements out of context and rated these statements on a five-point-scale for level of irony. Statements that were rated as not clearly ironic or not clearly non-ironic (with average ratings between 1 and 4) were recorded again with a more distinct pronunciation.

Context was manipulated by distinguishing between stories that are consistent with the subsequent evaluative statement (e.g., the story at the top left in Table 6) and stories that are inconsistent with the evaluative statement (e.g., the story at the top right in Table 6, in which the positive expectation of the story is inconsistent with the negative value of the evaluative statement). The second type of story invites an ironic interpretation of the evaluative statement, whereas the first type of story does not. We pre-tested these stories with six typically developing adult participants, who listened to the stories without evaluative statements and were asked to indicate which emoticon matched best with the main character of the story. Stories for which not at least five out of the six participants chose the target emoticon were adapted. We thus made sure that the stories in our experiment were all unambiguously interpreted as either positive or negative.

All stories were followed by three questions: a question about the emotion of the main character in the story, a first-order ToM question (ToM-1) about the emotion of the main character and a second-order ToM question (ToM-2) about the belief of the secondary character about the emotion of the main character. For example, after the story at the bottom left in Table 6 about Tim, the following three questions were asked:

(i) Which emoticon do you think matches best with Tim?

(ii) Do you think Tim thinks this is the worst job ever?

(iii) Does Tim's girlfriend think that Tim thinks this is the worst job ever?

Participants were instructed to answer the first question by pointing to one of four emoticons, which were presented on a piece of paper. These emoticons were selected on the basis of a pre-test: an online questionnaire. In this pre-test, the respondents ($n = 93$, all different from the participants in the present study) were presented with three emoticons and one emotion and were instructed to select the emoticon that they thought represented the emotion most accurately. A sample question (translated from Dutch) was: Which emoticon expresses the emotion ANGRY best, according to you? The participant had to choose between three emoticons from the Emoji of smartphones that can be used to indicate the emotion mentioned, in this case angry. For the present study, the four emoticons were chosen that – according to the results of the pre-test – best represented the four emotions used in the test: *happy*, *angry*,

scared and *sad* (see Figure 5 for the black and white versions of the colored emoticons).

Figure 5. Emoticons that were used in the test, from left to right: happy, angry, scared and sad.

The answer to the two ToM questions, illustrated by (ii) and (iii) above, could be *yes* or *no*. Participants' responses to each of the three question types (i.e., the choice of emoticon and the yes/no answers) were scored as either *ironic* or *not ironic*. For example, if the participant chose the sad, angry or scared emoticon in response to the statement "This is the worst job ever" in the context in Table 6, that would be scored as not ironic. On the other hand, if the participant chose the happy emoticon, that would be scored as ironic. Likewise, if the participant answered *yes* on the ToM-1 question in (ii) or the ToM-2 question in (iii), that would be scored as not ironic, and if they answered *no*, that would be scored as an ironic response. The responses per condition and per question type were analysed separately. This resulted in a mixed design with 12 variables. Performance on these variables was based on 4 items each and was converted into percentages of ironic responses.

In typically developing adolescents, we expect the combination condition to lead to more ironic responses than the other three conditions, because two cues that are important for recognizing and interpreting irony are present in this condition. If adolescents with ASD ignore contextual cues, prosodic cues, or both, when listening to utterances that are intended ironically, as is suggested by the literature, we expect them to give fewer ironic responses in the condition employing these cues than their typically developing peers. For both groups, least ironic responses are expected on the neutral condition, because this condition provides no cues for an ironic interpretation. As context is considered a stronger cue than prosody, both groups are also expected to give more ironic responses in the context condition than in the prosody condition. If context is a prerequisite for an ironic interpretation, the prosody condition may in fact not invite any ironic responses at all.

2.3 Procedure

All stories were recorded using Adobe Audition and played during test sessions using iTunes on a laptop with speakers. The participants listened to 2 practice stories and 16 experimental stories in a quiet room; the students of De Ambelt were tested in a room at school and the scouting youth was tested in a room in the scouting building. Two researchers were present during the test sessions.

	Emotion		ToM-1		ToM-2	
	Mean	SD	Mean	SD	Mean	SD
Neutral	0	0	3.9	9.4	3.9	9.4
Prosody	17.3	23.7	34.6	26.1	26.9	23.9
Context	75.0	17.7	76.9	16.0	61.5	24.2
Combination	57.7	15.8	65.4	28.0	65.4	28.0

Table 7. Mean percentages of ironic responses and standard deviations for the ASD group ($n = 13$) per condition (Neutral, Prosody, Context, Combination) and question type (Emotion, ToM-1, ToM-2).

	Emotion		ToM-1		ToM-2	
	Mean	SD	Mean	SD	Mean	SD
Neutral	0	0	3.6	9.1	5.4	10.6
Prosody	17.9	22.8	26.8	26.8	25.0	25.9
Context	80.4	17.5	82.1	15.3	55.4	24.4
Combination	80.4	14.4	87.5	19.0	75.0	25.9

Table 8. Mean percentages of ironic responses and standard deviations for the control group ($n = 14$) per condition (Neutral, Prosody, Context, Combination) and question type (Emotion, ToM-1, ToM-2).

One of the researchers made notes on the scoring forms and operated the laptop and the voice recorder. All sessions were recorded with a voice recorder. We started with a pre-test assessing whether the participants were familiar with the emoticons used in the test. Next, participants listened to the pre-recorded stories while looking at pictures. The pictures did not display any of the characters in the stories, that could be associated with emotions, but merely showed emotion-neutral objects mentioned in the story to help the participants focus on the task. Test sessions took approximately 12 minutes.

3 Results

Table 2 and Table 3 list the means and standard deviations on the four conditions and three question types for adolescents with ASD and the control group of adolescents without ASD. The results are also shown graphically in Figure 2 and Figure 3 below.

A mixed ANOVA was performed with *Group* (ASD, control) as the between-subjects factor, and *Condition* (neutral, prosody, context, combination) and *Question Type* (emotion, ToM-1, ToM-2) as within-subjects factors. There were significant main effects for *Condition* ($F(3, 75) = 186, p < .001, \eta^2 = .881$) and *Question Type* ($F(2, 50) = 5.81, p = .005, \eta^2 = .188$) on the mean percentages of ironic responses. There were significant interactions between *Condition* and *Question Type* ($F(3.8, 94.9) = 6.15, p < .001, \eta^2 = .197$) (Greenhouse-Geisser corrected) and between *Condition* and *Group* ($F(3, 75) = 3.45, p = .021, \eta^2 =$

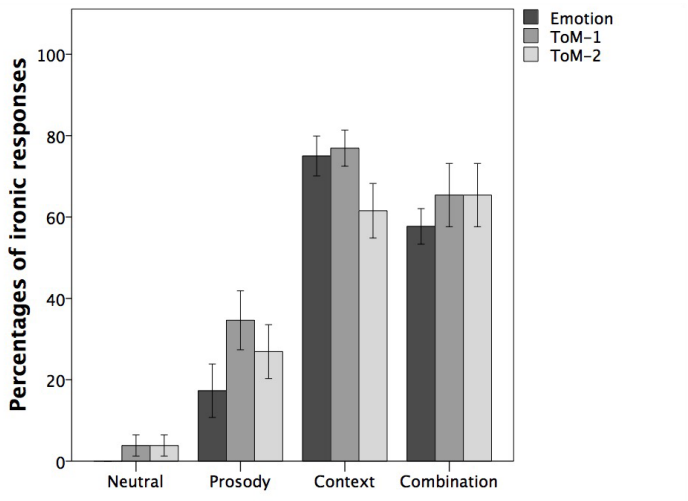


Figure 6. Mean percentages of ironic responses and standard deviations for the ASD group ($n = 13$) per condition (Neutral, Prosody, Context, Combination) and question type (Emotion, ToM-1, ToM-2).

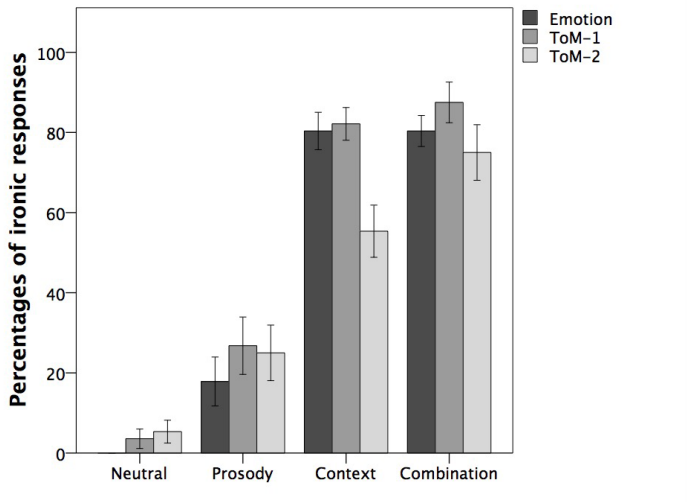


Figure 7. Mean percentages of ironic responses and standard deviations for the control group ($n = 14$) per condition (Neutral, Prosody, Context, Combination) and question type (Emotion, ToM-1, ToM-2).

.121). Both groups interpreted utterances that were accompanied by contextual cues or a combination of contextual and prosodic cues as more ironic than utterances that were not accompanied by any of these cues or were accompanied by prosodic cues only. Furthermore, for all four conditions, both groups had more ironic responses on the ToM-1 questions than on the ToM-2 questions.

To further inspect the interaction effect of *Group* with *Condition*, four one-way MANOVAs [8] were run separately for each condition with *Group* as the fixed factor and the three question types as dependent variables. There was a significant difference in ironic responses based on the participant's diagnosis for the combination condition ($F(3, 23) = 4.69, p = .011$; Wilk's $\Lambda = .621, \eta^2 = .379$). The group with ASD gave significantly less ironic responses on the combination condition than the control group. Follow-up tests on the *Combination Condition* furthermore revealed that *Group* had a significant effect on the emotion question ($F(1, 25) = 15.18, p = .001, \eta^2 = .378$) (Bonferroni corrected). Post hoc tests revealed that the group with ASD gave significantly less ironic responses on the emotion question in the combination condition ($M = 57.7, SD = 15.8$) than the control group ($M = 80.4, SD = 14.5$). From this we can conclude that for adolescent with ASD the presence of both contextual and prosodic cues leads to significantly less ironic interpretations than for adolescent without ASD, especially on the emotion question.

4 Discussion

In this study, we investigated whether adolescents with ASD have difficulty understanding irony. Hypothesizing that individuals with ASD are less capable of recognizing and interpreting irony than their typically developing peers, we furthermore wanted to find out in what way their recognition of irony depends on linguistic factors. To investigate this, we compared adolescents with ASD and typically developing adolescents on their interpretation of short stories in which prosodic and contextual cues for irony were manipulated.

If young individuals with ASD have difficulty understanding irony, we expect them to recognize the ironic intention in our stories less well than their typically developing peers. We found that, overall, the adolescents with ASD did not recognize the ironic intention less often than their typically developing peers. However, they did so when the ironic intention was indicated by both prosody and context. In that case, they gave fewer ironic responses on the emotion question than typically developing adolescents. Thus, adolescents with ASD have more difficulty than their typically developing peers to recognize the ironic intention of a statement that has an ironic prosody and at the same time is preceded by a context that is inconsistent with the positive or negative value of the statement.

Could the observed lower performance by the adolescents with ASD be attributed to their suboptimal use of prosodic or contextual cues? Both the adolescents with ASD and their typically developing peers interpreted stories in which the only cue to the speaker's ironic intention was the prosodic structure of the sentence differently than they did stories without any cues. The

presence of prosodic cues led to more ironic interpretations than the absence of any linguistic cues. Although we did not expect the prosody of a sentence alone to lead to an ironic interpretation, we found that prosody can invite an ironic interpretation. Furthermore, we found that adolescents with ASD, like their typically developing peers, use such prosodic cues in their interpretation of irony.

Also the presence of contextual cues was found to lead to more ironic interpretations compared to when there were no cues, both for adolescents with ASD and for typically developing adolescents. In both groups, the percentage of ironic interpretations was much larger when irony was signalled by context than when it was signalled by prosody. In fact, for both groups the combination of prosodic and contextual cues did not lead to more ironic interpretations than the presence of only contextual cues. This suggests that for adolescents with ASD as well as for adolescents without ASD the most important cue for recognizing irony is context. Thus, the difference between adolescents with ASD and their typically developing peers in the recognition of an ironic intention of a statement does not seem to be due to their insensitivity to prosody or context, which are the two most important linguistic cues for irony. Despite their sensitivity to prosodic and contextual cues for irony, it is possible that adolescents with ASD are less efficient in using these cues or perhaps have difficulty integrating two different cues.

Even in the conditions with the highest percentages of ironic interpretations, the adolescents' ironic interpretations generally did not rise above 80%. An exception are the responses by the typically developing adolescents on the ToM-1 questions in the context condition and the combination condition. Our study did not include an adult group, so we cannot be certain whether adolescents in general are not adult-like yet in their recognition of irony and their use of linguistic cues, or whether their performance with irony is adult-like. However, it is quite likely that adults are not perfect in their recognition of the ironic intentions of a speaker either.

When both prosodic and contextual cues were present, adolescents with ASD gave fewer ironic responses than their typically developing peers on emotion questions, but not on ToM-1 or ToM-2 questions. This does not mean that adolescents with ASD have no difficulty with ToM reasoning. In our study, the responses to the two ToM questions are dependent on the response on the emotion question. Therefore, our study did not test participants' ToM reasoning independently of their recognition of irony. To further investigate the relation between the recognition of irony and ToM reasoning, participants should be tested on an irony task as well as a separate ToM task. We leave this for further research.

Acknowledgement

We thank Scholengemeenschap De Ambelt in Zwolle and Scouting Group Don Bosco in Geldrop for their hospitality. In particular, we greatly appreciate the help of René Reith, Rianne Harms and the other teachers of De Ambelt and

Kirsten van der Geer of the scouting group. Furthermore, we are grateful to all participants and parents for their cooperation in this study. Laurie Stowe and Wander Lowie are thanked for their advice on statistical issues, and two anonymous reviewers for their useful comments.

References

[1] Adachi, T., T. Koeda, S. Hirabayashi, Y. Maeoka, M. Shiota, E. Charles Wright and A. Wada, *The metaphor and sarcasm scenario test: A new instrument to help differentiate high functioning pervasive developmental disorder from attention deficit/hyperactivity disorder*, Brain and Development **26** (2004), pp. 301–306.

[2] American Psychiatric Association, "Diagnostic and Statistical Manual of Mental Disorder: DSM-IV-TR," American Psychiatric Association, 2000.

[3] American Psychiatric Association, "Diagnostic and Statistical Manual of Mental Disorder: DSM 5," American Psychiatric Association, 2013.

[4] Baron-Cohen, S., *Hey! it was just a joke! understanding propositions and propositional attitudes by normally developing children and children with autism*, Israel Journal of Psychiatry and Related Sciences **34** (1997), pp. 174–178.

[5] Baron-Cohen, S., A. M. Leslie and U. Frith, *Does the autistic child have a "theory of mind"?*, Cognition **21** (1985), pp. 37–46.

[6] Bryant, G. A., *Prosodic contrasts in ironic speech*, Discourse Processes **47** (2010), pp. 545–566.

[7] Capelli, C. A., N. Nakagawa and C. M. Madden, *How children understand sarcasm: The role of context and intonation*, Child Development **61** (1990), pp. 1824–1841.

[8] Field, A., "Discovering Statistics using SPSS," Sage Publications, London, 2009.

[9] Filippova, E. and J. W. Astington, *Further development in social reasoning revealed in discourse irony understanding*, Child Development **79** (2008), pp. 126–138.

[10] Giora, R., *Understanding figurative and literal language: The graded salience hypothesis*, Cognitive Linguistics **8** (1997), pp. 183–206.

[11] Happé, F. G., *Communicative competence and theory of mind in autism: A test of relevance theory*, Cognition **48** (1993), pp. 101–119.

[12] Happé, F. G., *An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults*, Journal of Autism and Developmental Disorders **24** (1994), pp. 129–154.

[13] Happé, F. G., *The role of age and verbal ability in the theory of mind task performance of subjects with autism*, Child Development **66** (1995), pp. 843–855.

[14] Happé, F. G., *Central coherence and theory of mind in autism: Reading homographs in context*, British Journal of Developmental Psychology **15** (1997), pp. 1–12.

[15] Jorgensen, J., G. A. Miller and D. Sperber, *Test of the mention theory of irony.*, Journal of Experimental Psychology: General **113** (1984), pp. 112–120.

[16] Kaland, N., A. Møller-Nielsen, K. Callesen, E. L. Mortensen, D. Gottlieb and L. Smith, *A new advanced test of theory of mind: Evidence from children and adolescents with Asperger syndrome*, Journal of Child Psychology and Psychiatry **43** (2002), pp. 517–528.

[17] MacKay, G. and A. Shaw, *A comparative study of figurative language in children with autistic spectrum disorders*, Child Language Teaching and Therapy **20** (2004), pp. 13–32.

[18] Martin, I. and S. McDonald, *An exploration of causes of non-literal language problems in individuals with asperger syndrome*, Journal of Autism and Developmental Disorders **34** (2004), pp. 311–328.

[19] Massaro, D., A. Valle and A. Marchetti, *Irony and second-order false belief in children: What changes when mothers rather than siblings speak?*, European Journal of Developmental Psychology **10** (2013), pp. 301–317.

[20] Norbury, C. and D. Bishop, *Inferential processing and story recall in children with communication problems: a comparison of specific language impairment, pragmatic language impairment and high-functioning autism*, International Journal of Language & Communication Disorders **37** (2002), pp. 227–251.

[21] Perner, J. and H. Wimmer, *"John thinks that Mary think that..." attribution of second-order beliefs by 5-to 10-year-old children*, Journal of Experimental Child Psychology **39** (1985), pp. 437–471.
[22] Rutter, M., L. Mawhood and P. Howlin, *Language delay and social development*, in: P. Fletcher and D. Hall, editors, *Specific Speech and Language Disorders in Children: Correlates, Characteristics and Outcomes*, Whurr, London, 1992 pp. 63–78.
[23] Verbrugge, R. and L. Mol, *Learning to apply theory of mind*, Journal of Logic, Language and Information **17** (2008), pp. 489–511.
[24] Wang, A. T., S. S. Lee, M. Sigman and M. Dapretto, *Neural basis of irony comprehension in children with autism: the role of prosody and context*, Brain **129** (2006), pp. 932–943.
[25] Wilson, D. and D. Sperber, *Explaining irony*, in: *Meaning and Relevance*, Cambridge University Press, 2012 pp. 123–145.
[26] Wimmer, H. and J. Perner, *Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception*, Cognition **13** (1983), pp. 103–128.
[27] Winner, E., "The Point of Words: Children's Understanding of Metaphor and Irony," Harvard University Press, Cambridge, MA, 1988.
[28] Winner, E. and S. Leekam, *Distinguishing irony from deception: Understanding the speaker's second-order intention*, British Journal of Developmental Psychology **9** (1991), pp. 257–270.

Oracle bites Theory

Albert Visser [1]

Philosophy, Faculty of Humanities
Utrecht University

Abstract

In the context of Elementary Arithmetic (EA) we know that already an extremely weak arithmetical theory like R proves every true Σ_1-sentence. Thus, it would seem that adding the true Σ_1-sentences to the axiom set of a given theory adds nothing. However, Elementary Arithmetic cannot prove this 'obvious fact'. We show that under the assumption of the negation of Σ_1-collection, the weak theory PA^- plus the true Σ_1-sentences is inconsistent.
It follows that, in EA plus the negation of Σ_1-collection, any consistent extension U of PA^- is not closed under finite conjunctions: there is a conjunction of theorems of U such that U plus that conjunction is inconsistent.
A corollary of our main insight is that Σ_1-collection is, over Elementary Arithmetic, equivalent to the restricted consistency of PA^- plus the true Σ_1-sentences.
In Appendix C, we prove slightly modified results for the weaker theory R. A consequence of these results is that, over EA, Σ_1-collection is equivalent to the consistency of the theory axiomatized by the theorems of R.
In Appendix A, we provide a lay person's summary of the results of the paper.

Keywords: collection principles, reflection principles, completeness principles, elementary arithmetic

1 Introduction

It is told of a German monk who came back to tell his friend about heaven that he only spoke the words *totaliter aliter*. The worlds of metamathematics in which we start from the negation of a cherished assumption are also fairly *aliter*. Fortunately, in their case, we are often able to say much more than just affirming their alterity. The study of such a world is beneficial since it gives us a better feeling of where the cherished assumption is needed. In this paper, we study one such world: a world in which we start from the negation of Σ_1-collection.

In the paper we will consider the connection between Σ_1-collection and provability with an oracle for Σ_1-truth in the context of EA or Elementary

[1] I am grateful to Zofia Adamowicz, Lev Beklemishev, Emil Jeřábek, Joost Joosten and Leszek Kołodziejczyk for helpful comments, corrections and additional insights. I thank the two anonymous referees for their insightful comments.

Arithmetic. It is well known that, in the context of EA, all theories extending the very weak arithmetic R prove all true Σ_1-sentences. So, it would seem that adding these sentences to the axiom set of a given theory is a harmless addition that could only result in speeding up some proofs. However, the verification of this 'obvious fact' essentially depends on the presence of Σ_1-collection. We will show that the 'obvious fact' fails as badly as possible if we add the negation of Σ_1-collection to EA. Under these assumptions, the weak theory PA^- plus all true Σ_1-sentences becomes inconsistent.

We note that it follows that, in EA plus the negation of Σ_1-collection, consistent extensions U of PA^- are not closed under finite conjunctions of theorems: there is a finite conjunction of theorems such that adding it to the theory makes it inconsistent. In other words: the theory axiomatized by the set of theorems of U is inconsistent.

The methods of the paper allow us to see that, over EA, Σ_1-collection is equivalent with the restricted consistency of PA^- plus all true Σ_1-sentences.

Remark 1.1 There is a great and beautiful open problem in the global area of this paper: is the theory $\mathrm{I}\Delta_0 + \neg\mathsf{Exp} + \neg\,\Sigma_1\text{-}\mathsf{coll}$ consistent? See [14] and [2]. Regrettably, I do not see any relevance of the results of the present paper for this problem.

Remark 1.2 The present paper is, in a sense, a sequel of Section 5 of my paper [10].

In Appendix A we give a description for the lay person of some salient results of the paper.

2 Basics

In Appendix B we will give a bit more detail on the various theories discussed in the paper. The reader is referred to [4] for more details and more discussion.

2.1 Theories

The theories we study in this paper are extensions of the weak arithmetic R of [9] in (definitional extensions of) the arithmetical language. See Appendix B for the axioms of R. See also [12] for discussion and further references on R.

Remark 2.1 Since we will work in Elementary Arithmetic as our ambient meta-theory, we have the luxury of the totality of exponentiation available. This means that we do not have to worry about the big disjunctions in the axiom set that Tarski, Mostowski and Robinson call Ω_4 (R4 of Appendix B). If we would work in the context of S^1_2 these disjunctions would be too big. There it would be better to replace Ω_4 by the axioms:
$x \leq \underline{0} \leftrightarrow x = \underline{0}$ and $x \leq \underline{n+1} \leftrightarrow (x \leq \underline{n} \vee x = \underline{n+1})$.

Our theories are given by arithmetical predicates that define the axiom set. We allow axiom sets defined by more complex formulas than e.g. Δ^{b}_1. Suppose e.g. the theory U is axiomatized by $\alpha(x)$ and a set X of sentences is given by $\beta(x)$. Then, $U + X$ is the theory given by $\alpha(x) \vee \beta(x)$. We implicitly assume that the formulas defining the axioms sets of familiar theories are chosen in some obvious way.

A salient theory in the paper is PA^-, the theory of discretely ordered commutative semirings with a least element. This theory is mutually interpretable with Robinson's Arithmetic Q. However, PA^- has the additional good property that it is sequential as was shown in [5]. In Appendix B, we give the axioms of PA^-. We refer the reader to [6] and [5] for information on PA^-.

A second salient theory is S^1_2, the theory of p-time computability introduced in [3]. In Appendix B, we give a bit more detail concerning S^1_2. See also [4].

A third salient theory is Elementary Arithmetic EA, i.e. $\mathrm{I}\Delta_0 + \mathsf{Exp}$. Harvey Friedman calls this theory EFA, for: Elementary Function Arithmetic. We describe this theory in Appendix B.

Since we are going to enter a world in which Σ_1-collection fails, we will also be interested in two theories that are, in the real world, extensionally the same as a given theory U but which might be extensionally different in that other world. The first theory is the theory $U + \mathfrak{S}$, where

$$\mathfrak{S} := \{S \in \Sigma_1\text{-}\mathsf{sent} \mid \mathsf{true}(S)\}$$

and true is the arithmetized Σ_1-truth predicate. It is well known that the elementary properties of this truth predicate are verifiable in EA. We note that the axiom set of $U + \mathfrak{S}$ is given by a Σ_1-formula. The second theory is the theory $\mathsf{thm}(U)$, which is the theory axiomatized by the theorems of U. We note that the axiom set of $\mathsf{thm}(U)$ is also given by a Σ_1-formula. We will show that, according to EA, the theories $U + \mathfrak{S}$ and $\mathsf{thm}(U)$ prove the same theorems. It follows, in EA, that thm is a closure operation (modulo sameness of the theorem set).

We will use modal notation $\Box$ for (formalized) provability. We employ the dot notation for variables occurring freely inside boxes. For example, $\Box_U P(\dot{x})$ means: the number resulting by substituting the Gödel number of the numeral of x for $\underline{\ulcorner v \urcorner}$ in $\underline{\ulcorner P(v) \urcorner}$ has the property prov_U. More formally, this could be written as:

$$\Box_U P(\dot{x}) :\leftrightarrow \mathsf{prov}_U(\mathsf{sub}(\mathsf{num}(x), \underline{\ulcorner v \urcorner}, \underline{\ulcorner P(v) \urcorner})).$$

2.2 Formula Classes

We define the following formula classes in the arithmetical language.

- Δ_0 is the class of formulas in which all quantifiers are bounded.
- $\Sigma_1 := \Sigma_{1,0}$ is the class of formulas of the form $\exists \boldsymbol{x}\, S_0(\boldsymbol{x}, \boldsymbol{y})$, where S_0 is in Δ_0.
- $\Sigma_{1,n+1}$ is the class of formulas of the form $\exists \boldsymbol{x}\, \forall \boldsymbol{y} < \boldsymbol{t}\, S_0(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, where S_0 is $\Sigma_{1,n}$. Here the bounded quantification is subject to the usual restriction that the quantified variable y does not occur in the bounding term.
- $\Sigma_{1,\infty}$ is the union of the $\Sigma_{1,n}$.

We define $\Pi_{1,n}$ analogously. Our main interesest in the present paper will be in these classes for $n \leq 2$.

2.3 Collection

The Σ_1-Collection Principle, Σ_1-coll, is:

- $\vdash \forall x < a\, \exists y\, Sxy\boldsymbol{z} \to \exists b\, \forall x < a\, \exists y < b\, Sxy\boldsymbol{z}$, where $Sxy\boldsymbol{z}$ is Σ_1.

The principle Σ_1-coll follows from the special case that S is in Δ_0. Suppose $Sxy\boldsymbol{z}$ is of the form is $\exists \boldsymbol{w}\, S_0\boldsymbol{w}xy\boldsymbol{z}$ with $S_0 \in \Delta_0$. Now consider the formula $S_0^+xy\boldsymbol{z} :\leftrightarrow \exists \boldsymbol{w}, u < y\, S_0\boldsymbol{w}xu\boldsymbol{z}$. It is easy to see that collection for S follows from collection for S_0^+ in PA^-.

We show that, in EA, Σ_1-Collection can be compressed to a single sentence. In EA, we have a Σ_1-satisfaction predicate $\mathsf{sat}(\sigma, s)$. Here σ stands for a sequence of numbers and s represents a Σ_1-formula. This satisfaction predicate has the form $\exists w\, \mathsf{sat}_0(w, \sigma, s)$, where sat_0 is in Δ_0. The usual construction yields that the witness for s (given σ) is below w, since w contains this witness as a component. We consider the following principle:

$$(\dagger)\ \ \forall x < a\, \exists w\, \mathsf{sat}_0(w, \langle x\rangle * \sigma, s) \to \exists b\, \forall x < a\, \exists w < b\, \mathsf{sat}_0(w, \langle x\rangle * \sigma, s).$$

Consider any Σ_1-formula $Sxy\boldsymbol{z}$. Let $S'x\boldsymbol{z} := \exists y\, Sxy\boldsymbol{z}$. Assuming (†), we find:

$$\begin{aligned}\mathsf{EA} \vdash \forall x < a\, \exists y\, Sxy\boldsymbol{z} &\to \forall x < a\, \exists w\, \mathsf{sat}_0(w, \langle x\rangle * \langle \boldsymbol{z}\,\rangle, \underline{\ulcorner S'\urcorner})\\ &\to \exists b\, \forall x < a\, \exists w < b\, \mathsf{sat}_0(w, \langle x\rangle * \langle \boldsymbol{z}\,\rangle, \underline{\ulcorner S'\urcorner})\\ &\to \exists b\, \forall x < a\, \exists y < b\, Sxy\boldsymbol{z}\end{aligned}$$

The last step uses that y is bounded by w. We note that our reduction uses the presence of parameters. It was shown in [7] that, over EA, parameter-free Σ_1-collection cannot be finitely axiomatized. Hence the use of parameters is essential here.

In [1], it is shown that the Σ_1-collection principle is $\Pi_{1,1}$. We give a proof here. Let $S_0 \in \Sigma_0$. We consider the following properties. (We suppress the parameters and the dependence of the P on S_0.)

$P_0(a)$ $\ \forall x \le a\, \exists y\, S_0xy \to \exists b\, \forall x\ \le a\, \exists y \le b\, S_0xy$.

$P_1(a)$ $\ \exists u \le a\, \forall v\, (S_0uv \to \forall x\ \le a\, \exists y \le v\, S_0xy)$.

Lemma 2.2 *We have:*

a. $\mathsf{PA}^- \vdash P_1(a) \to P_0(a)$.

b. $\mathrm{I}\Delta_0 \vdash P_0(a) \to P_1(a)$.

Proof. We prove (a). Reason in PA^-. Suppose $P_1(a)$ and $\forall x \le a\, \exists y\, S_0xy$. Pick u as promised by $P_1(a)$. By our second assumption, there is a v such that S_0uv. Thus, by $P_1(a)$, we get the conclusion of $P_0(a)$ with $b := v$.

We prove (b). Reason in $\mathrm{I}\Delta_0$. Assume $P_0(a)$. In case $\exists x \le a\, \forall y\, \neg\, S_0xy$, we immediately have $P_1(a)$. So suppose $\forall x \le a\, \exists y\, S_0xy$. By $P_0(a)$, we have, for some b, that $\forall x \le a\, \exists y \le b\, S_0xy$. By the Δ_0-minimum principle, there is a minimum such b, say $b^\star$. If we had $\forall x \le a\, \exists y\ < b^\star\, S_0xy$, then we would have $\forall x \le a\, \exists y \le b^\star - 1\, S_0xy$, contradicting the minimality of $b^\star$. So, we get $\exists x \le a\, \forall y < b^\star\, \neg\, S_0xy$. Let $u^\star \le a$ be such that $\forall y < b^\star\, \neg\, S_0u^\star y$. Consider any

v and suppose $S_0u^\star v$. It follows that $v \geq b^\star$. Since, for any $x \leq a$ there is a $y \leq b^\star$ with S_0xy, there also is a $y \leq v$ such that S_0xy. So, we find $P_1(a)$. $\Box$

Since P_1 is $\Pi_{1,1}$ it follows that Σ_1-collection is $\Pi_{1,1}$ in $\mathrm{I}\Delta_0$. The following lemma will be used in the proof of our main result Theorem 4.1. We remind the reader that a virtual class is *progressive* if it contains 0 and is closed under successor.

Lemma 2.3 *Both P_0 and P_1 are progressive in* PA^-.

Proof. We leave the proof for P_0 to the reader. We prove the claim for P_1. We remind the reader of the definition of P_1:

$$P_1(a) :\leftrightarrow \exists u \leq a\, \forall v\, (S_0uv \to \forall x\ \leq a\, \exists y \leq v\, S_0xy).$$

We reason in PA^-. Clearly we have $P_1(0)$. Suppose $P_1(a)$. Let b be the promised witness and suppose S_0bv. In case we have (i) $\exists y \leq v\, S_0(a+1)y$, we easily find that $b \leq (a+1)$ and $\forall x\ \leq (a+1)\, \exists y \leq v\, S_0xy$. So b is also a witness for $P_1(a+1)$. Suppose (ii) $\forall y \leq v \neg S_0(a+1)y$. We claim that, in this case, $a+1$ is a witness for $P_1(a+1)$. Clearly $a+1 \leq a+1$. Suppose $S_0(a+1)w$. Then, by (ii) $w > v$. So $\forall x\ \leq a\, \exists y \leq w\, S_0xy$. Moreover, $\exists y \leq w\, S_0(a+1)y$, since this is witnessed by w itself. So, in both cases, we have a witness for $P_1(a+1)$. $\Box$

For more information about the $\Sigma_{1,n}$ classes and Σ_1-collection, see e.g. [11].

3 Completeness

A central component of the proof of our main result is the $\Sigma_{1,1}$-completeness of extensions of R plus the true Σ_1-sentences. In this section we will treat some basic facts concerning both $\Sigma_{1,0}$-completeness and $\Sigma_{1,1}$-completeness.

We can distinguish three versions of Σ_1-completeness ($\Sigma_{1,0}$-completeness) for a theory U.

- *Local or sentential Σ_1-completeness:* for all Σ_1-sentences S, we have: $\vdash S \to \Box_U S$.
- *Uniform Σ_1-completeness:* for all Σ_1-formulas $S\boldsymbol{x}$, we have: $\vdash \forall \boldsymbol{x}\, (S\boldsymbol{x} \to \Box_U S\dot{\boldsymbol{x}})$.
- *Global Σ_1-completeness:* $\vdash \forall S \in \mathsf{sent}(\Sigma_1)\, (\mathsf{true}(\dot{S}) \to \Box_U S)$. Note that for global Σ_1-completeness, we need an ambient theory like EA in order to provide the basic facts concerning the relevant Σ_1-truth predicate.

As is well known, all three versions hold over EA when U is EA-verifiably an extension of R. In this paper we zoom in on uniform Σ_1-completeness and uniform $\Sigma_{1,1}$-completeness. Of course, uniform $\Sigma_{1,1}$-completeness is defined just like uniform $\Sigma_{1,0}$-completeness. Undoubtedly more could be said about global $\Sigma_{1,1}$-completeness too, but developing this would distract us too much from the main line of the paper. We will use simply 'Σ_1-completeness' for *uniform Σ_1-completeness*.

Our first order of business is to provide a counter example for EA-verifiable $\Sigma_{1,1}$-completeness. Our counter example is also a counter example to sentential $\Sigma_{1,1}$-completeness. The method used is due to Paris and Kirby. See [8] or [6].

Theorem 3.1 *For some sentence S in $\Sigma_{1,1}$, we have:* $\mathsf{EA} \nvdash S \to \Box_{\mathsf{EA}} S$.

Proof. We remind the reader that we can transform any Σ_1-formula $S(x)$ into a Σ_1-formula $S^\circ(x)$ that has the following properties:

i. $\{x \mid S^\circ(x)\} \subseteq \{x \mid S(x)\}$,

ii. if $\{x \mid S(x)\}$ is non-empty, then so is $\{x \mid S^\circ(x)\}$,

iii. $\{x \mid S^\circ(x)\}$ has at most one element.

Suppose $S(x)$ is $\exists \boldsymbol{y}\, S_0(\boldsymbol{y}, x)$, where S_0 is Δ_0. We take:

- $S_1(z) :\leftrightarrow \exists \boldsymbol{y} \leq z\, (z = \langle x, \boldsymbol{y} \rangle \wedge S_0(\boldsymbol{y}, x))$.
- $S^\circ(x) :\leftrightarrow \exists z\, (S_1(z) \wedge \forall w < z \neg S_1(w) \wedge (z)_0 = x)$.

Thus, we can work with Σ_1-definitions of elements which have the normal form S° without worrying about the uniqueness clause. Using the Σ_1-truth predicate true, we can construct a Σ_1-predicate def, where $\mathsf{def}(s, x)$ codes: s is a Σ_1-definition of x. Let:

$$S^\star := \exists p\, (\, \mathsf{proof}_{\mathsf{PA}}(p, \bot) \wedge \forall q < p \neg \mathsf{proof}_{\mathsf{PA}}(q, \bot) \wedge \forall x \leq p\, \exists s < p\ \mathsf{def}(s, x)).$$

We note that $S^\star$ is $\Sigma_{1,1}$.

Consider any model $\mathcal{N}$ of $\mathsf{PA} + \mathsf{incon}(\mathsf{PA})$. Let $\mathcal{M}$ be the submodel given by the Σ_1-definable elements of $\mathcal{N}$. By the results of [8] (see also [6]), we have: $\mathcal{M} \models \mathsf{EA}$. Moreover, clearly, $\mathcal{M} \models S^\star$.

To prove our theorem it suffices to show that $\mathcal{M} \not\models \Box_{\mathsf{EA}} S^\star$. We note that if $\mathcal{M} \models \Box_{\mathsf{EA}} S^\star$, then $\mathcal{N} \models \Box_{\mathsf{EA}} S^\star$. Since $\mathcal{N}$ is a model of PA and since PA proves reflection for EA, we have $\mathcal{N} \models S^\star$. Quod non, since PA proves the relevant version of the Pigeon Hole Principle. □

Remark 3.2 The above proof is a variation of a proof in Section 5 of [10].

We show that EA does prove uniform $\Sigma_{1,1}$-completeness for $\Box_{\mathsf{R}+\mathfrak{S}}$ and, *ipso facto* for all EA-verifiable extensions of $\mathsf{R} + \mathfrak{S}$.

We can describe an $\mathsf{R} + \mathfrak{S}$-proof p as follows. It is a *preproof*, i.e. a proof from R plus assumptions that are Σ_1-sentences such that its assumptions, say the elements of $\mathsf{ass}(p)$, are all Σ_1-true.

Theorem 3.3 *We have, for all $\Sigma_{1,1}$-formulas $S\boldsymbol{x}$,* $\mathsf{EA} \vdash \forall \boldsymbol{x}\, (S\boldsymbol{x} \to \Box_{\mathsf{R}+\mathfrak{S}} S\dot{\boldsymbol{x}})$.

Proof. We give our proof for the case that the initial blocks of quantifiers consist of just one quantifier. The more general case is similar. Let $S\boldsymbol{x}$ be a formula of the form $\exists y \forall z < t(y, \boldsymbol{x})\, S_0(z, y, \boldsymbol{x})$, where S_0 is $\Sigma_{1,0}$.

We reason in EA. Suppose $S\boldsymbol{a}$. Find b such that we have:

$$(\ddagger)\ \ \forall z < t(b, \boldsymbol{a})\, S_0(z, b, \boldsymbol{a}).$$

We can easily construct an R-preproof p of $\forall z < t(\dot{b}, \dot{\boldsymbol{a}})\, S_0(z, \dot{b}, \dot{\boldsymbol{a}})$ from assumptions $S_0(0, \dot{b}, \dot{\boldsymbol{a}})$, $\ldots$, $S_0(t(\dot{b}, \dot{\boldsymbol{a}})\dot{-}1, \dot{b}, \dot{\boldsymbol{a}})$. By (‡) all these assumptions are Σ_1-true. $\Box$

We can refine Theorem 3.3 a bit. Let ρ be the complexity measure *depth of quantifier alternations*. Inspecting the proof of Theorem 3.3, we see that, for a fixed $S\boldsymbol{x}$, the ρ-complexity of the formulas in the $\mathsf{R}+\mathfrak{S}$-proof of $S\boldsymbol{x}$ is bounded by a fixed standard number depending only on the ρ-complexity of $S\boldsymbol{x}$. The same holds when we exchange $\mathsf{R}+\mathfrak{S}$ for e.g. $\mathsf{PA}^- + \mathfrak{S}$, since in this last theory the verifications of the axioms of R are of a fixed restricted complexity.

We write $\Box_{U,n}$ for provability in U from axioms bounded by n with a proof only involving formulas of ρ-complexity below n. By the above considerations, we find:

Theorem 3.4 *Consider a $\Sigma_{1,1}$-formula $S\boldsymbol{x}$. We can find a standard n such that $\mathsf{EA} \vdash \forall \boldsymbol{x}\, (S\boldsymbol{x} \to \Box_{\mathsf{R}+\mathfrak{S},n} S\dot{\boldsymbol{x}})$. Similarly, for $U + \mathfrak{S}$, for a theory U in which the verifications of the axioms of R are EA-verifiably of standardly bounded complexity, like all theories containing PA^-.*

Remark 3.5 Note that in the witnesses of $\Box_{U+\mathfrak{S},n} A$ only Σ_1-sentences with complexity less than n will be used. So we could as well write something like $\Box_{U+\mathfrak{S}_n,n} A$.

Instead of restricted provability we could as well have used cut-free provability, tableaux provability or Herbrand provability.

4 Oracle Provability and Collection

In this section we study provability with a Σ_1-oracle in the context of EA.

Theorem 4.1 $\mathsf{EA} \vdash \neg\, \Sigma_1\text{-}\mathsf{coll} \to \Box_{\mathsf{PA}^- + \mathfrak{S}} \bot$.

Proof. We reason in EA. Suppose $\neg\, \Sigma_1$-coll. Thus, for some a and s, we have $\neg\, P_0(a, s)$. Here s appears in the role of an appropriate parameter. It follows that $\neg\, P_1(a, s)$ (Lemma 2.2). Since $\neg\, P_1(a, s)$ is $\Sigma_{1,1}$, it follows that $\Box_{\mathsf{PA}^- + \mathfrak{S}} \neg\, P_1(\dot{a}, \dot{s})$ (Theorem 3.3). Since $\{x \mid P_1(x, s)\}$ is progressive in PA^- (Lemma 2.3), it follows by induction that $\forall x\, \Box_{\mathsf{PA}^-} P_1(\dot{x}, \dot{s})$. To make the induction work in EA we show that the witnessing proofs can be given an multi-exponential bound. Thus, we find: $\Box_{\mathsf{PA}^-} P_1(\dot{a}, \dot{s})$. *A fortiori* we have $\Box_{\mathsf{PA}^- + \mathfrak{S}} P_1(\dot{a}, \dot{s})$. Hence, $\Box_{\mathsf{PA}^- + \mathfrak{S}} \bot$. $\Box$

Remark 4.2 With a bit more care we can prove the analogue of Theorem 4.1 also for $\mathsf{Q} + \mathfrak{S}$.

Theorem 4.1 allows us to characterize provability with a Σ_1-oracle for extensions of PA^-.

Theorem 4.3 *Suppose U is, EA-verifiably, an extension of PA^-. We have:*
$\mathsf{EA} \vdash \forall A\, (\Box_{U+\mathfrak{S}} A \;\leftrightarrow\; (\neg\, \Sigma_1\text{-}\mathsf{coll} \vee \Box_U A))$.

Proof. We reason in EA.

Left to Right. Suppose $\Box_{U+\mathfrak{S}} A$. If we have $\neg\, \Sigma_1$-coll, we are done. So, suppose Σ_1-coll. In this case we can transform any $U + \mathfrak{S}$-proof p into an U-proof q, by inserting proofs of the Σ_1-sentences occurring as axioms in p. By Σ_1-coll, the witnesses of these sentences are all bounded by some number

r. Moreover, a proof of a Σ_1-sentence witnessed by a number below r, is bounded by 2^{2^r}. Thus, the transformation from p to q exists given the presence of exponentiation.

Right to Left. If $\neg\,\Sigma_1$-coll, then $\Box_{\mathsf{PA}^- + \mathfrak{S}}\bot$, and, hence, $\Box_{U+\mathfrak{S}}\bot$ and, *a fortiori*, $\Box_{U+\mathfrak{S}}A$. Moreover, if $\Box_U A$, then, trivially, $\Box_{U+\mathfrak{S}}A$. □

Problem 4.4 *What is the combined provability logic of $\Box_{\mathsf{EA}}$ and $\Box_{\mathsf{EA}+\mathfrak{S}}$? And a slight variation: what is the combined provability logic of $\Box_{\mathsf{EA}}$ and $\Box_{\mathsf{EA}+\mathfrak{S}}$ with a constant for Σ_1-coll?*

We formulate an immediate corollary of Theorem 4.1. Clearly, in EA, we have, for all A, if $\Box_{\mathsf{PA}^- + \mathfrak{S}}A$, then $\Box_{\mathsf{thm}(\mathsf{PA}^-)}A$. It follows that:

Corollary 4.5 $\mathsf{EA} \vdash \neg\,\Sigma_1\text{-coll} \to \Box_{\mathsf{thm}(\mathsf{PA}^-)}\bot$.

From this, we easily derive:

Theorem 4.6 *Suppose U is, EA-verifiably, an extension of PA^-. We have:*
$\mathsf{EA} \vdash \forall A\,(\Box_{\mathsf{thm}(U)}A \;\leftrightarrow\; (\neg\,\Sigma_1\text{-coll} \vee \Box_U A))$.

It follows that $\mathsf{EA} \vdash \forall A\,(\Box_{\mathsf{thm}(U)}A \;\leftrightarrow \Box_{U+\mathfrak{S}}A)$.

Inspecting the proofs of Theorems 4.1 and 4.3 (replacing the application of Theorem 3.3 by Theorem 3.4), we find:

Theorem 4.7 *Suppose U is, EA-verifiably, an extension of PA^-. We have, for a sufficiently large number n:*
$\mathsf{EA} \vdash \forall A \in \mathsf{sent}_n\,(\Box_{U+\mathfrak{S},n}A \;\leftrightarrow\; (\neg\,\Sigma_1\text{-coll} \vee \Box_{U,n}A))$.

We remind the reader that, for any n, $\mathsf{EA} \vdash \mathsf{con}_n(\mathsf{PA}^-)$. (This follows e.g. from the results of [13], noting the fact that restricted provability and tableaux provability are multi-exponentially connected and the mutual interpretability of Q and PA^-). If we take $U := \mathsf{PA}^-$ and $A := \bot$ in Theorem 4.7, we obtain:

$$\mathsf{EA} \vdash \Box_{\mathsf{PA}^- + \mathfrak{S},n}\bot \;\leftrightarrow\; (\neg\,\Sigma_1\text{-coll} \vee \Box_{\mathsf{PA}^-,n}\bot).$$

And, hence, $\mathsf{EA} \vdash \Box_{\mathsf{PA}^- + \mathfrak{S},n}\bot \leftrightarrow \neg\,\Sigma_1$-coll. Thus, we may conclude:

Theorem 4.8 *We have, for sufficiently large n,* $\mathsf{EA} \vdash \mathsf{con}_n(\mathsf{PA}^- + \mathfrak{S}) \leftrightarrow \Sigma_1\text{-coll}$.

It is not difficult to see that we may replace PA^- in the statement of the theorem by e.g. Q or S^1_2. Thus, we have found that Σ_1-coll is EA-provably equivalent to a reasonably non-contrived consistency statement. We note that in the statement of the theorem con_n can also be replaced by cut-free consistency, tableaux consistency or Herbrand consistency.

In the appendix we will sketch a proof that shows that with minor differences in formulation we can replace PA^- by R.

Remark 4.9 As referee I points out our results show the non-verifiability of Craig's trick in EA. (See e.g. [4],Theorem 2.29, Chapter III.) This is clear since Craig's trick allows the transformation of a recursively enumerable axiom set to a elementarily decidable one. One easily sees that Craig's trick is verifiable in $\mathsf{EA} + \Sigma_1$-coll.

Theorems 4.8 and C.4 constitute positive answers of sorts for $n = 1$ to Problem 5 of the section on Reflection Principles of Lev Beklemishev's list of questions on `http://www.mi.ras.ru/~bekl/problems.html`:

Are the collection principles $\mathrm{B}\Sigma_n$ equivalent to some form of reflection over EA?

It seems reasonably hopeful to extend the results to $n > 1$ using adaptations of the methods employed above. However, I did not try this.

As pointed out by referee I, the above answers to Beklemishev's question are not fully satisfactory, since $\mathsf{con}_n(\mathsf{PA}^- + \mathfrak{S})$ is the restricted consistency statement of a theory that is not axiomatized by an elementarily decidable set of axioms as is usually the case in Beklemishev's work.

Of course, we can recast $\mathsf{con}_n(\mathsf{PA}^- + \mathfrak{S})$ more in the style of a reflection principle by noting that:

$$\mathsf{EA} \vdash \mathsf{con}_n(\mathsf{PA}^- + \mathfrak{S}) \leftrightarrow \forall X \subseteq \Pi_1\text{-}\mathsf{sent}\,(\Box_{\mathsf{PA}^-} \bigvee X \to \exists P \in X\, \mathsf{true}^*(P)).$$

Here X ranges over finite sets and true^* is the Π_1-truth predicate. In the recast version, reflection takes more the form of a combination of reflection and a disjunction property, so again one could take exception to this form of the result as providing an answer to Beklemishev's question. Referee I suggests the following theorem as giving an unobjectionable answer. We publish it here with his/her kind permission.

Theorem 4.10 *i.* $\mathsf{EA} + \mathsf{con}(\mathsf{PA}^-) \vdash \mathsf{Rfn}_{\Pi_{1,1}}(\mathsf{PA}^-) \leftrightarrow \Sigma_1\text{-}\mathsf{coll}$.

ii. $\mathsf{EA} \vdash \mathsf{Rfn}_{n,\Pi_{1,1}\cap\Gamma_n}(\mathsf{PA}^-) \leftrightarrow \Sigma_1\text{-}\mathsf{coll}$, *for sufficiently large* n. *Here* Γ_n *is the class of formulas with depth of quantifier alternations less than or equal to* n.

Here:

- $\mathsf{Rfn}_{\Pi_{1,1}}(\mathsf{PA}^-)$ is the scheme $\forall \boldsymbol{x}\,(\Box_{\mathsf{PA}^-} P(\dot{\boldsymbol{x}}) \to P(\boldsymbol{x}))$, where $P \in \Pi_{1,1}$.
- $\mathsf{Rfn}_{n,\Pi_{1,1}}(\mathsf{PA}^-)$ is the scheme $\forall \boldsymbol{x}\,(\Box_{\mathsf{PA}^-,n} P(\dot{\boldsymbol{x}}) \to P(\boldsymbol{x}))$, where $P \in \Pi_{1,1} \cap \Gamma_n$. Here Γ_n is the set of formulas with depth of quantifier alternations less than or equal to n.

Proof. We prove (i). Reason in $\mathsf{EA} + \mathsf{con}(\mathsf{PA}^-)$. Suppose $\mathsf{RFN}_{\Pi_{1,1}}(\mathsf{PA}^-)$. Since, $\{y \mid P_1(y, \boldsymbol{x})\}$ is progressive in y in PA^- (Lemma 2.3), we have $\forall y, \boldsymbol{x}\, \Box_{\mathsf{PA}^-} P_1(\dot{y}, \dot{\boldsymbol{x}})$. Since Π_1 is $\Pi_{1,1}$, by reflection, we find $\forall y, \boldsymbol{x}\, P_1(y, \boldsymbol{x})$, and, hence, we have Σ_1-coll.

In the other direction, suppose Σ_1-coll and, for some $\boldsymbol{x}$, $\Box_{\mathsf{PA}^-} P(\dot{\boldsymbol{x}})$. Suppose $\neg P(\boldsymbol{x})$. We rewrite $\neg P(\boldsymbol{x})$, using only predicate logic, to a sentence of the form:

$$\exists \boldsymbol{y}\, \forall \boldsymbol{z} < \boldsymbol{t}\, \exists \boldsymbol{u}\, S^\circ(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{x}).$$

Here S° is Δ_0. By Σ_1-coll, we find:

$$\exists \boldsymbol{y}, w\, \forall \boldsymbol{z} < \boldsymbol{t}\, \exists \boldsymbol{u} < w\, S^\circ(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{x}).$$

It follows that:

$$\Box_{\mathsf{PA}^-} \exists \boldsymbol{y}, w\, \forall \boldsymbol{z} < \boldsymbol{t}\, \exists \boldsymbol{u} < w\, S^\circ(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \dot{\boldsymbol{x}}).$$

Hence, $\Box_{\mathsf{PA}^-}\bot$. Quod non. We may conclude that $P(\boldsymbol{x})$.

The proof of (ii) is similar. □

Problem 4.11 *Let PrL_Σ be predicate logic in signature Σ. Can we give a perspicuous axiomatization of $X := \{A \in \mathsf{sent}_\Sigma \mid \mathsf{EA} + \neg\,\Sigma_1\text{-}\mathsf{coll} \vdash \Box_{\mathsf{thm}(\mathsf{PrL}_\Sigma)}A\}$? We note, for example, that for any translation τ of the language of arithmetic in $\mathcal{L}_\Sigma$, we have that $\neg\,(\bigwedge \mathsf{PA}^-)^\tau$ is in X. So, X strictly extends PrL_Σ. On the other hand X should be contained in $\mathsf{PrL}^{\mathsf{fin}}_\Sigma$, the principles valid in all finite models of signature Σ. Consider any finite model $\mathcal{M}$ of signature Σ. We can think of the theory of $\mathcal{M}$ as axiomatized by a standardly finite model description A. The theory EA can verify that proofs from A can be replaced by a proofs with an elementary bound. Hence $\mathsf{thm}(A)$ will EA-verifiably prove the same theorems as A. It follows, externally, that X is contained in the set of sentences that are true in $\mathcal{M}$.*

References

[1] Adamowicz, Z. and C. Dimitracopoulos, *On a problem concerning parameter free induction*, Mathematical Logic Quarterly **37** (1991), pp. 363–366.

[2] Adamowicz, Z., L. Kołodziejczyk and J. Paris, *Truth definitions without exponentiation and the Σ_1 collection scheme*, The Journal of Symbolic Logic **77** (2012), pp. 649–655.

[3] Buss, S., "Bounded Arithmetic," Bibliopolis, Napoli, 1986.

[4] Hájek, P. and P. Pudlák, "Metamathematics of First-Order Arithmetic," Perspectives in Mathematical Logic, Springer, Berlin, 1993.

[5] Jeřábek, E., *Sequence encoding without induction*, Mathematical Logic Quarterly **58** (2012), pp. 244–248.

[6] Kaye, R., "Models of Peano Arithmetic," Oxford Logic Guides, Oxford University Press, 1991.

[7] Kaye, R., J. Paris and C. Dimitracopoulos, *On parameter-free induction schemas*, Journal of Symbolic Logic **53** (1988), pp. 1082–1097.

[8] Paris, J. and L. Kirby, *Σ_n-collection schemas in arithmetic*, in: A. Macintyre, L. Pacholski and J. Paris, editors, *Logic Colloquium '77* (1978), pp. 199–209.

[9] Tarski, A., A. Mostowski and R. Robinson, "Undecidable theories," North–Holland, Amsterdam, 1953.

[10] Visser, A., *The formalization of interpretability*, Studia Logica **51** (1991), pp. 81–105.

[11] Visser, A., *Peano Corto and Peano Basso: A study of local induction in the context of weak theories*, Mathematical Logic Quarterly **60** (2014), pp. 92–117.
URL `http://dx.doi.org/10.1002/malq.201200102`

[12] Visser, A., *Why the theory* R *is special*, in: N. Tennant, editor, *Foundational Adventures. Essays in honour of Harvey Friedman*, College Publications, UK, 2014 pp. 7–23, originally published online by Templeton Press in 2012. See `http://foundationaladventures.com/`.

[13] Wilkie, A. and J. Paris, *On the scheme of induction for bounded arithmetic formulas*, Annals of Pure and Applied Logic **35** (1987), pp. 261–302.

[14] Wilkie, A. and J. Paris, *On the existence of end extensions of models of bounded induction*, in: *Logic, Methodology and Philosophy of Science VIII* (*Moscow 1987*) (1989), pp. 143–161.

Appendix

A Lay Person's Summary

In this section, we briefly explain in somewhat simpler terms what this paper is about. The paper is a study of features of certain subsystems of the strong system Peano Arithmetic PA. [2] Given the fact that we can do nearly all numerical reasoning we can think of in Peano Arithmetic, why look at weaker systems at all? The reason is that a lot of the fine structure of reasoning cannot be explicitly studied when we just consider a strong system like PA. Take, for example, the version of the Pigeon Hole Principle that says that if we put k objects in n boxes and $k > m$, then some box will contain more than one object. This principle can be proven by induction. However, this reduction does not well reflect the fundamental character of the principle in finite combinatorics, where it appears as a basic principle of thought. It is therefore interesting to study the status of the Pigeon Hole Principle in the context of weaker theories where it can function as a basic axiom.

A second reason to study weaker systems is that we may be interested in extracting algorithmic information from proofs. A proof in a weaker system will usually yield a better algorithm.

In the present paper we study a version of the (finite) Collection Principle. Suppose we have a function f from a finite set X to the natural numbers. In that case the values of the function are bounded by some natural number n. We study a special case of this principle, to wit Σ_1-collection. It says that if we have a computable function g on the set of numbers $\leq k$, then the values of g have a bound n.

To get the study of the Σ_1-collection Principle off the ground, we have to work in an meta-theory that is weaker than the strong theory Peano Arithmetic PA. After all, PA proves Σ_1-collection, so we are barred from inspection what happens if it fails, but for the uninteresting observation that such failure leads to a contradiction over PA. For this reason we work over the weaker theory Elementary Arithmetic EA. [3] There are many reasons why EA is a good choice. One such reason is that we have a rather simple model construction due to Paris and Kirby that gives us models of EA plus the negation of Σ_1-collection. Another reason is that over EA the principle Σ_1-collection is finitely axiomatizable.

The Σ_1-collection Principle is almost everywhere present in our numerical reasoning. It is so obvious that most of the time we do not notice we are using it. One of the examples that the present paper zooms in on is as follows. Suppose we have an effectively axiomatized theory U and a finite set of X of theorems of U. Let $U + X$ be the theory axiomatized by the axioms of U plus the elements of X. Then, U and $U + X$ prove the same theorems. Clearly, this insight is the basis of our use of already proven theorems to prove new ones.

---

[2] See Appendix B for a description of PA.

[3] See Appendix B for a description of EA.

Let's call this insight *the Lemma Principle*.

In the paper we show that without Σ_1-collection we cannot derive the Lemma Principle over our chosen ambient theory EA. Here is some intuition: suppose we have an $(U + X)$-proof p^* of A. How do we convert this into a U-proof? Well, we take, for every B in X, a U-proof p_B of B and we add p_B above the assumption B in p^*. But what if the p_B could grow arbitrarily large so that no finite proof would result of this construction? To ensure that the p_B do not get out of hand, we need Σ_1-collection.

In fact, for a wide range of theories U, we can prove something strange: *if we assume the negation of Σ_1-collection*, then we can find a finite set of theorems X of U such that $U + X$ is inconsistent. Among the theories within the range of this result are such familiar, trusted theories as EA and PA. Hence, if Σ_1-collection fails, a theory and its finite extension with some theorems can be very different in their consequences.

An important program in current metamathematics if Reverse Mathematics: over a given basic theory we show that different salient principles are equivalent. In the present paper we prove a result of this form: under the assumption of the consistency of U, we can prove that the Lemma Principle for U is equivalent to Σ_1-collection. [4]

B Axioms of Important Systems

The system R was introduced in [9]. It is a very weak system that is essentially undecidable. A theory is essentially undecidable if every consistent extension of it is undecidable. The language of R is the language of arithmetic with 0, S (successor), $+$ and $\cdot$. We write

$$\underline{n} := \overbrace{\mathsf{S}\cdots\mathsf{S}}^{n\times}0.$$

The theory R is axiomatized as follows.

R1. $\vdash \underline{n} + \underline{m} = \underline{n + m}$

R2. $\vdash \underline{n} \cdot \underline{m} = \underline{n \cdot m}$

R3. $\vdash \underline{n} \neq \underline{m}$, for $n \neq m$

R4. $\vdash x \leq \underline{n} \to \bigvee_{m \leq n} x = \underline{m}$

R5. $\vdash x \neq \underline{n} \vee \underline{n} \leq x$

One easily shows that R is not finitely axiomatizable. For more information on R we refer the reader to [12].

The system Q was introduced in [9]. It is a weak finitely axiomatized theory in the language of arithmetic that extends R (and, hence, is essentially undecidable). The theory Q is axiomatized as follows.

[4] In Appendix C we provide examples of theories U such that this results holds for U and such that EA proves the consistency of U.

Q1. $\vdash \mathsf{S}x = \mathsf{S}y \to x = y$

Q2. $\vdash \mathsf{S}x \neq 0$

Q3. $\vdash x = 0 \vee \exists y\, x = \mathsf{S}y$

Q4. $\vdash x + 0 = x$

Q5. $\vdash x + \mathsf{S}y = \mathsf{S}(x + y)$

Q6. $\vdash x \cdot 0 = 0$

Q7. $\vdash x \cdot \mathsf{S}y = x \cdot y + x$

A slightly stronger theory than Q is PA^-. This theory has the advantage that it has better metamathematical properties than Q and that is has a more mathematical 'feel' to it. It is the preferred basic theory of researchers in the area of non-standard models of arithmetic. PA^- is the theory of theory of discretely ordered commutative semirings with a least element. It is given in the arithmetical language plus the relation symbol $\leq$. The theory PA^- is given by the following axioms.

PA^-1. $\vdash x + 0 = x$

PA^-2. $\vdash x + y = y + x$

PA^-3. $\vdash (x + y) + z = x + (y + z)$

PA^-4. $\vdash x \cdot 1 = x$

PA^-5. $\vdash x \cdot y = y \cdot x$

PA^-6. $\vdash (x \cdot y) \cdot z = x \cdot (y \cdot z)$

PA^-7. $\vdash x \cdot (y + z) = x \cdot y + x \cdot z$

PA^-8. $\vdash x \leq y \vee y \leq x$

PA^-9. $\vdash (x \leq y \wedge y \leq z) \to x \leq z$

PA^-10. $\vdash x + 1 \not\leq x$

PA^-11. $\vdash x \leq y \to (x = y \vee x + 1 \leq y)$

PA^-12. $\vdash x \leq y \to x + z \leq y + z$

PA^-13. $\vdash x \leq y \to x \cdot z \leq y \cdot z$

PA^-14. $\vdash x \leq y \to \exists z\, x + z = y$

Emil Jeřábek in his paper [5] employs a version without the subtraction axiom (PA^-14).

The theory S^1_2 was first given in [3]. It was tailored for the study of p-time computability. It is an ideal theory for the arithmetization of syntax. It would take us too far afield to give a full description of S^1_2. I just will give some salient features. We start with the arithmetical language including $\leq$ and add two new function symbols $|\cdot|$ and $\#$. Here $|x|$ is the length of the binary representation of x and $\#$, the smash function, is $x, y \mapsto 2^{|x|\cdot|y|}$. This function grows faster than multiplication but slower than exponentiation. We have a basic arithmetic

like PA^- plus suitable axioms for $|\cdot|$ and $\#$. Moreover, we have an induction axiom of the form:

$$(A0 \wedge \forall x\,(Ax \to (A(2x+1) \wedge A(2x+2))).$$

Here A is Σ_1^{b} which means that it is of the form $\exists y \leq t\, A_0 xy$, where A_0 is Δ_0^{b}. A formula is Δ_0^{b} if all its quantifiers are sharply bounded, i.e. they are of the form $\forall z \leq |u|$ or $\exists z \leq |u|$. An important feature of S_2^1 is the fact that the theory is finitely axiomazable.

The theory Elementary Arithmetic or EA or EFA or $\mathrm{I}\Delta_0 + \mathsf{exp}$ is given as follows. It is a theory in the arithmetical language with axioms (Q1,2,4,5,6,7) plus Δ_0-induction, where a formula is Δ_0 if all its quantifiers are bounded. We can show that the graph of exponentiation can be written as a Δ_0-formula. We have as final axiom exp which states that exponentiation is total. An important feature of EA is the fact that the theory is finitely axiomazable.

Finally, Peano Arithmetic or PA is the theory in the language of arithmetic axiomatized by (Q1,2,4,5,6,7) plus full induction.

C Downtuning our Results to R

We first prove the desired result for the theory NSN. This is *the theory of a non-standard number*. Here are the axioms of NSN.

NSN1. $\vdash \mathsf{S}x \neq 0$

NSN2. $\vdash \mathsf{S}x = \mathsf{S}y \to (x = y \vee \mathsf{S}x = c)$

NSN3. $\vdash \mathsf{S}c = c$

NSN4. $\vdash c \neq \underline{n}$

NSN5. $\vdash x + 0 = x$

NSN6. $\vdash x + \mathsf{S}y = \mathsf{S}(x + y)$

NSN7. $\vdash x \times 0 = 0$

NSN8. $\vdash x \times \mathsf{S}y = x \times y + x$

NSN9. $\vdash (A0 \wedge \forall x\,(Ax \to A\mathsf{S}x)) \to \forall x\, Ax$

The theory NSN is locally finite. This means that every finitely axiomatized subtheory of NSN has a finite model. We can easily verify that the usual ordering linearly orders the NSN-numbers with minimum 0 and maximum c. Addition and Multiplication are like ordinary addition and multiplication — only they are cut off at c.

Let $P_1(a, s)$ be the formula given above Lemma 2.2 for the instance of Σ_1-collection that implies all other instances. We have $\mathsf{NSN} \vdash \forall s\,\forall a\, P_1(a, s)$. We can see this in two ways. First, trivially, we have $\mathsf{NSN} \vdash \forall s\,\forall a\, P_0(a, s)$. It follows by the reasoning of the proof of Lemma 2.2 that $\mathsf{NSN} \vdash \forall s\,\forall a\, P_1(a, s)$. Alternatively, we verify as in Lemma 2.3 that P_1 is progressive in a and prove the desired result by induction.

If we work in EA plus the negation of Σ_1-collection, we find $\Box_{\mathsf{NSN}}\forall s\,\forall a\,P_1(a,s)$. Also, since $\neg\,P_1(a,s)$, for some a and s, and since NSN contains R, we find that $\Box_{\mathsf{NSN}+\mathfrak{S}}\neg\,P_1(\dot{a},\dot{s})$. It follows that $\Box_{\mathsf{NSN}+\mathfrak{S}}\bot$. Thus, we have:

Theorem C.1 *i.* $\mathsf{EA}\vdash\neg\,\Sigma_1\text{-}\mathsf{coll}\to\Box_{\mathsf{NSN}+\mathfrak{S}}\bot$.

ii. $\mathsf{EA}\vdash\Box_{\mathsf{NSN}+\mathfrak{S}}A\leftrightarrow(\neg\,\Sigma_1\text{-}\mathsf{coll}\vee\Box_{\mathsf{NSN}}A)$.

iii. $\mathsf{EA}\vdash\Box_{\mathsf{thm}(\mathsf{NSN})}A\leftrightarrow(\neg\,\Sigma_1\text{-}\mathsf{coll}\vee\Box_{\mathsf{NSN}}A)$.

We can use the fact that EA verifies that NSN is locally finite, to prove that $\mathsf{EA}\vdash\mathsf{con}(\mathsf{NSN})$. It follows that:

Theorem C.2 *i.* $\mathsf{EA}\vdash\mathsf{con}(\mathsf{NSN}+\mathfrak{S})\leftrightarrow\Sigma_1\text{-}\mathsf{coll}$.

ii. $\mathsf{EA}\vdash\mathsf{con}(\mathsf{thm}(\mathsf{NSN}))\leftrightarrow\Sigma_1\text{-}\mathsf{coll}$.

We turn to the treatment of R. We refer the reader to our paper [12] for background on R. By the main result of [12] it follows that R interprets NSN, say via an interpretation K. Clearly, NSN extends R. Since EA verifies that NSN is locally finite, we can verify the correctness of the construction of K in EA. This can be seen by inspecting the construction in [12]. A consequence is that we can find a multi-exponential bound on the transformation of an NSN-proof of A into an R-proof of A^K.

We work in EA. Consider a proof p of $\bot$ in $\mathsf{NSN}+\mathfrak{S}$. We can transform p into a proof $p^\star$ of $\bot$ in $\mathsf{R}+\mathfrak{S}^K$, where $\mathfrak{S}^K:=\{S^K\mid\mathsf{true}(S)\}$, using the fact that we have multi-exponential bounds for the R-proofs of A^K, where A is an axiom of NSN. Thus we find:

Theorem C.3 $\mathsf{EA}\vdash\neg\,\Sigma_1\text{-}\mathsf{coll}\to\Box_{\mathsf{R}+\mathfrak{S}^\kappa}\bot$.

It follows that:

Theorem C.4 $\mathsf{EA}\vdash\mathsf{con}(\mathsf{thm}(\mathsf{R}))\leftrightarrow\Sigma_1\text{-}\mathsf{coll}$.

We note the important difference with Theorem 4.5: we can work here with ordinary consistency instead of restricted consistency.