Proceedings of the Workshop "Reasoning About Other Minds: Logical and Cognitive Perspectives"

Groningen, Monday July 11th, 2011

edited by Jan van Eijck and Rineke Verbrugge

The workshop theme

In recent years, the human ability to *reasoning about mental states of others* in order to explain and predict their behavior has come to be a highly active area of research. Researchers from a wide range of fields – from biology and psychology through linguistics to game theory and logic– contribute new ideas and results.

This interdisciplinary workshop, collocated with the Thirteenth International Conference on Theoretical Aspects of Rationality and Knowledge (TARK XIII), aims to shed light on models of social reasoning that take into account realistic resource bounds. People reason about other people's mental states in order to understand and predict the others' behavior. This capability to reason about others' knowledge, beliefs and intentions is often referred to as *theory of mind*.

Idealized rational agents are capable of recursion in their social reasoning, and can reason about phenomena like common knowledge. Such idealized social reasoning has been modeled by modal logics such as epistemic logic and BDI (belief, desire, intention) logics and by epistemic game theory. However, in real-world situations, many people seem to lose track of such recursive social reasoning after only a few levels. The workshop provides a forum for researchers that attempt to analyze, understand and model how resource-bounded agents reason about other minds.

Topics of interest of the workshop include but are not limited to the following:

- Logics modeling human social cognition;
- Computational cognitive models of theory of mind;
- Behavioral game theory;
- Bounded rationality in epistemic game theory;
- Relations between language and social cognition;
- Models of the evolution of theory of mind;
- Models of the development of theory of mind in children;
- Bounded rationality in multi-agent systems;
- Formal models of team reasoning;
- Theory of mind in specific groups, e.g., autism spectrum disorder;
- Complexity measures for reasoning about other minds.

Invited talks at the workshop

In addition to the contributed lectures and poster presentations, of which the articles are gathered in this volume, the workshop also presents three invited speakers:

- Petra Hendriks (University of Groningen): Bounded reasoning about others in language: Evidence from language acquisition;
- Barbara Dunin-Kęplicz (Warsaw University and Polish Academy of Sciences): Calibrating the expressiveness of collective notions;
- Chris Baker (Massachusetts Institute of Technology): Modeling human reasoning about beliefs, desires, goals and social relations.

Acknowledgments

We would like to thank all the people who helped to bring about the workshop *Reasoning About Other Minds: Logical and Cognitive Perspectives*. First of all, we thank all invited speakers, contributed speakers and poster presenters for ensuring a diverse and interesting workshop.

Special thanks are due to the members of the program committee for their professionalism and their dedication to select papers of quality and to provide authors with useful, constructive feedback during the in-depth reviewing process:

Program committee

- Johan van Benthem (University of Amsterdam and Stanford University)
- Robin Clark (University of Pennsylvania)
- Hans van Ditmarsch (University of Sevilla)
- Peter Gärdenfors (Lund University)
- Sujata Ghosh (University of Groningen)
- Noah Goodman (Stanford University)
- Bart Hollebrandse (University of Groningen)
- Eric Pacuit (Tilburg University and University of Maryland)
- Rohit Parikh (City University of New York)
- Jun Zhang (University of Michigan)

In addition, a number of experts outside the program committee also reviewed submissions to the workshop, for which we are very grateful. Thank you Krzysztof Apt, Boicho Kokinov, Barteld Kooi, Katja Mehlhorn, Ben Meijering, Olivier Roy, Sunil Simon, and Jakub Szymanik!

We would like to thank our colleagues in the TARK Organizing Committee for handling many organizational tasks: Virginie Fiutek, Sujata Ghosh, Barteld Kooi, Ben Meijering, Bryan Renne, Ben Rodenhäuser, Olivier Roy, Sonja Smets, Allard Tamminga, and Bart Verheij. The workshop website would not exist without the great help of Ben Meijering and Sujata Ghosh in Groningen.

Finally, we would like to express our gratitude to NWO for largely financing this workshop through Vici grant NWO 227-80-00, *Cognitive Systems in Interaction: Logical and Computational Models of Higher-Order Social Cognition.* We also acknowledge the other TARK sponsors for their financial support: the University of Groningen, CWI, KNAW, NWO, Springer, Gemeente Groningen, Provincie Groningen, and the Dutch Association for Logic (VvL).

Amsterdam — Jan van Eijck
Groningen — Rineke Verbrugge

July 2011

Table of contents

Common Knowledge in Email Exchanges

Floor Sietsma[1] and Krzysztof R. Apt[1 2]

[1] Centre Mathematics and Computer Science (CWI), Amsterdam
[2] University of Amsterdam

Abstract. We consider a framework in which a group of agents communicates by means of emails, with the possibility of replies, forwards and blind carbon copies (BCC). We study the epistemic consequences of such email exchanges by introducing an appropriate epistemic language and semantics. Then we clarify when a group of agents acquires common knowledge of the formula expressing that an email was sent.

1 Introduction

1.1 Motivation

Email is by now a prevalent form of communication. Its advantages speak for themselves. However, we rarely pause to reflect on its undesired consequences. Just to mention a few.

One occasionally reads about scandals caused by email leaks, see, e.g., [2]. On a smaller scale, users of the *blind carbon copy* feature (BCC) are sometimes confronted with an undesired situation in which a BCC recipient of an email reveals his status to others by using the *reply-all* feature.

Recently, a main Dutch daily, NRC Handelsblad, reported, see [7], that Wouter Bos, the Deputy Prime minister in the previous Dutch government, revealed the extensive network of his contacts by sending out his new email address to hundreds of influential recipients whose email addresses were erronously put in the CC list instead of the BCC list. The list was leaked to the newspaper.

So when studying email exchanges a natural question arises: what are their knowledge-theoretic consequences? To put it more informally: after an email exchange took place, who knows what? To answer this question we study email exchanges by focusing on relevant features that we encounter in most email systems.

More specifically, we study the following form of email communication:

- each email has a sender, a non-empty set of regular recipients and a (possibly empty) set of blind carbon copy (BCC) recipients. Each of the recipients receives a copy of the message and is only aware of the regular recipients and not of the BCC recipients,
- in the case of a reply to or a forward of a message, the *unaltered* original message is included,
- in a reply or a forward, one can append new information to the original message one replies to or forwards.

As a result, the email exchanges, as studied here, are essentially different from other forms of communication, in particular from multicasting, i.e., sending a message to a group of recipients. Also, the resulting model of email communication differs from the ones that were studied in other papers in which only limited aspects of emails have been considered. These papers are discussed below.

1.2 Related work

The study of the epistemic effects of communication in distributed systems originated in the eighties and led to the seminal book [5]. The relevant literature, including [4], deals only with the customary form of communication, notably asynchronous send.

The epistemic effects of other forms of communication were studied in numerous papers. In particular, in [9] the communicative acts are assumed to consist of an agent j 'reading' an arbitrary propositional formula from another agent i. The idea of an epistemic contents of an email is implicitly present in [10], where a formal model is proposed that formalizes how communication changes the knowledge of a recipient of the message. In [3] a dynamic epistemic logic modelling effects of communication and change is introduced and extensively studied. [8] surveys these and related approaches and discusses the used epistemic, dynamic epistemic and doxastic logics. Further, in [12] an epistemic logic was proposed to reason about information flow w.r.t. underlying communication channels.

Most related to the work here reported are the following two references. [1] studied knowledge and common knowledge in a set up in which the agents send and forward propositional formulas in a social network. However, the forward did not include the original message which limited the scope of the resulting analysis. More recently, in [11] explicit messages are introduced in a dynamic epistemic logic to analyze a very similar setting, though it is assumed that the number of messages is finite and BCC is simulated as discussed in Section 6.

Finally, the concept of a causal relation between messages in distributed systems is due to [6].

1.3 Contributions

To study the relevant features of email communication we introduce in the next section a carefully chosen set of emails. We make a distinction between a *message*, which is sent to a public recipient list, and an *email*, which consists of a message and a set of BCC recipients. This distinction is relevant because a *forward* email contains only a message, without the list of BCC recipients. We also introduce the notion of a legal state that captures the fact that there is a *causal ordering* on the emails. For example, an email needs to precede any forward of it.

To reason about the knowledge of the agents after an email exchange has taken place we introduce in Section 3 an appropriate epistemic language. Its semantics takes into account the uncertainty of the recipients of an email about its set of BCC recipients and the ignorance about the existence of emails that one neither sent nor received. This semantics allows us to evaluate epistemic

formulas in legal states, in particular the formulas that characterize the full knowledge-theoretic effect of an email.

In Section 4 we present the main result of the paper, that clarifies when a group of agents can acquire common knowledge of the formula expressing the fact that an email has been sent. This characterization in particular sheds light on the epistemic consequences of BCC. The proof this result is given in Section 5. Then in Section 6 we show how BCC can be simulated using only messages without BCC recipients.

2 Preliminaries

2.1 Messages

In this section we define the notion of a message. We assume non-empty and finite sets of ***agents*** $Ag = \{1, \ldots, n\}$ and of ***notes*** P. Each agent has a set of notes he knows initially.

We make a number of assumptions. Firstly, we assume that the agents do not know which notes belong to the other agents. Furthermore, we assume that the agents only exchange emails about the notes. In particular, they cannot communicate epistemic formulas. We also assume that an agent can send a message to other agents containing a note only if he knows it initially or has learned it through an email he received earlier.

We inductively define ***messages*** as follows, where in each case we assume that $G \neq \emptyset$:

- $m := s(i, l, G)$; the message containing note l, sent by agent i to the group G,
- $m := f(i, l.m', G)$; the forwarding by agent i of the message m' with added note l, sent to the group G.

So the agents can send a message with a note or forward a message with a note appended. In the examples we assume that there exists a note **true** that is known by all agents and we identify **true**.m with m.

If m is a message, then we denote by $S(m)$ and $R(m)$, respectively, the singleton set consisting of the agent sending m and the group of agents receiving m. So for the above messages m we have $S(m) = \{i\}$ and $R(m) = G$. We do allow that $S(m) \subseteq R(m)$, i.e., that one sends a message to oneself.

Special forms of the forward messages can be used to model reply messages. Given $f(i, l.m, G)$, using $G = S(m)$ we obtain the customary *reply* message and using $G = S(m) \cup R(m)$ we obtain the customary *reply-all* message. (In the customary email systems there is syntactic difference between a forward and a reply to these two groups of agents, but the effect of both messages is exactly the same, so we ignore this difference.) In the examples we write $s(i, l, j)$ instead of $s(i, l, \{j\})$, etc.

2.2 Emails

An interesting feature of most email systems is that of the blind carbon copy (BCC). We would like to study the epistemic effects of sending an email with BCC recipients and will now include this feature in our presentation.

In the previous subsection we defined messages that have a sender and a group of recipients. Now we define the notion of an email which allows the additional possibility of sending a BCC of a message. The BCC recipients are not listed in the list of recipients, therefore we have not included them in the definition of a message. Formally, by an ***email*** we mean a construct of the form m_B, where m is a message and $B \subseteq Ag$ is a possibly empty set of BCC recipients. Given a message m we call each email m_B a ***full version*** of m.

Since the set of BCC recipients is 'secret', it does not appear in a forward. That is, the forward of an email m_B with added note l is the message $f(i, l.m, G)$ or an email $f(i, l.m, G)_C$, in which B is not mentioned. However, this forward may be sent not only by a sender or a regular recipient of m_B, but also by a BCC recipient. Clearly, the fact that an agent was a BCC recipient of an email is revealed at the moment he forwards its message.

A natural question arises: what if someone is both a regular recipient and a BCC recipient of an email? In this case, no one (not even this BCC recipient himself) would ever notice that this recipient was also a BCC recipient since everyone can explain his knowledge of the message by the fact that he was a regular recipient. Only the sender of the message would know that this agent was also a BCC recipient. This fact does not change anything and hence we assume that for any email m_B we have $(S(m) \cup R(m)) \cap B = \emptyset$.

2.3 Legal States

Our goal is to analyze knowledge of agents after some email exchange took place. To this end we need to define a possible collection of sent emails.

First of all, we shall assume that every message is used only once. In other words, for each message m there is at most one full version of m, i.e., an email of the form m_B. The rationale behind this decision is that a sender of m_B and $m_{B'}$ might equally well send a single email $m_{B \cup B'}$. This assumption can be summarized as a statement that the agents do not have 'second thoughts' about the recipients of their emails. It also simplifies subsequent considerations.

One could argue that there is a total ordering on the emails entailed by the time at which they were sent. However, the fact that an email was sent at a certain time does not imply that it was also read at that time. All what we can assert it that the email was read after it was sent. Further, the order in which an agent reads the emails he received is undetermined. This explains why we do not impose a linear ordering on the emails and we do not give the messages time stamps.

However, we have to impose some ordering on the sets of emails. For example, we need to make sure that the agents only send information they actually know.

Moreover, a forward can only be sent after the original email was sent. We will introduce the minimal partial ordering that takes care of such issues.

First, we define by structural induction the ***factual information*** $FI(m)$ contained in a message m as follows:

$$FI(s(i,l,G)) := \{l\},$$
$$FI(f(i,l.m,G)) := FI(m) \cup \{l\}.$$

We will use the concept of a ***state*** to model the effect of an email exchange. A state $s = (E, L)$ is a tuple consisting of a finite set E of emails that took place and a sequence $L = (L_1, \ldots, L_n)$ of sets of notes for all agents. The idea of these sets is that each agent i initially knows the notes in L_i. We use E_s and L_s to denote the corresponding elements of a state s, and $L_1, \ldots, L_n$ to denote the elements of L.

We say that a state $s = (E, L)$ is ***legal*** w.r.t. a strict partial ordering (in short, an spo) $\prec$ on E if it satisfies the following conditions:

L.1: for each email $f(i,l.m,G)_B \in E$ an email $m_C \in E$ exists such that $m_C \prec f(i,l.m,G)_B$ and $i \in S(m) \cup R(m) \cup C$,
L.2: for each email $s(i,l,G)_B \in E$, where $l \notin L_i$, an email $m_C \in E$ exists such that $m_C \prec s(i,l,G)_B$, $i \in R(m) \cup C$ and $l \in FI(m)$,
L.3: for each email $f(i,l.m',G)_B \in E$, where $l \notin L_i$, an email $m_C \in E$ exists such that $m_C \prec f(i,l.m',G)_B$, $i \in R(m') \cup C$ and $l \in FI(m')$.

Condition L.1 states that the agents can only forward messages they previously received. Conditions L.2 and L.3 state that if an agent sends, a note that he did not initially know, then he must have learned it by means of an earlier email.

We say that a state s is legal iff it is legal w.r.t. some spo. Given a legal state s, by its ***causality ordering*** we mean the smallest (so the least constraining) spo w.r.t. which s is legal.

So a state is legal if every forward was preceded by its original message, and for every note sent in an email there is an explanation how the sender of the email learned this note.

3 Epistemic language and its semantics

We want to reason about the knowledge of the agents after an email exchange has taken place. For this purpose we use a language $\mathfrak{L}$ of communication and knowledge defined as follows:

$$\varphi ::= m \mid i \blacktriangleleft m \mid \neg\varphi \mid \varphi \wedge \varphi \mid C_G\varphi$$

The formula m expresses the fact that m has been sent in the past, with some unknown group of BCC recipients. The formula $i \blacktriangleleft m$ expresses the fact that agent i was involved in a full version of the message m, i.e., he was either

the sender, a recipient or a BCC recipient. The formula $C_G\varphi$ denotes common knowledge of the formula φ in the group G.

We use the usual abbreviations $\vee$, $\rightarrow$ and $\leftrightarrow$ and we use $K_i\varphi$ as an abbreviation of $C_{\{i\}}\varphi$. The fact that an email with a certain set of BCC recipients was sent can be expressed in our language by the following abbreviation:

$$m_B ::= m \wedge \bigwedge_{i \in S(m) \cup R(m) \cup B} i \blacktriangleleft m \wedge \bigwedge_{i \notin S(m) \cup R(m) \cup B} \neg i \blacktriangleleft m$$

Note that this formula expresses the fact that the message m was sent with exactly the group B as BCC recipients, which captures precisely the intended meaning of m_B.

We now provide a semantics for this language interpreted on legal states, inspired by the epistemic logic and the history-based approaches of [9] and [10]. For every agent i we define an indistinguishability relation $\sim_i$, where we intend $s \sim_i s'$ to mean that agent i cannot distinguish between the states s and s'. We first define this relation on the level of emails as follows (recall that we assume that senders and regular recipients are not BCC recipients):

$$m_B \sim_i m'_{B'}$$

iff one of the following contingencies holds:

(i) $i \in S(m)$, $m = m'$ and $B = B'$,
(ii) $i \in R(m) \setminus S(m)$ and $m = m'$,
(iii) $i \in B \cap B'$, and $m = m'$,
(iv) $i \notin S(m) \cup R(m) \cup B$ and $i \notin S(m') \cup R(m') \cup B'$.

Condition (i) states that the sender of an email confuses it only with the email itself. In turn, condition (ii) states that each regular recipient of an email who is not a sender confuses it with any email with the same message but possibly sent to a different BCC group. Next, condition (iii) states that each BCC recipient of an email confuses it with any email with the same message but sent to a possibly different BCC group of which he is also a member. Finally, condition (iv) states that each agent confuses any two emails in which he is not involved.

As an example consider the emails $e := s(i, l, j)_\emptyset$ and $e' := s(i, l, j)_{\{k\}}$. We have then $e \not\sim_i e'$, $e \sim_j e'$ and $e \not\sim_k e'$. Intuitively, agent j cannot distinguish between these two emails because he cannot see whether k is a BCC recipient. In contrast, agents i and k can distinguish between these two emails.

Next, we extend the indistinguishability relation to legal states by defining

$$(E, L) \sim_i (E', L')$$

iff the following holds:

- $L_i = L'_i$,
- for any $m_B \in E$ such that $i \in S(m) \cup R(m) \cup B$ there is $m_{B'} \in E'$ such that $m_B \sim_i m_{B'}$,

– for any $m_{B'} \in E'$ such that $i \in S(m) \cup R(m) \cup B'$ there is $m_B \in E$ such that $m_B \sim_i m_{B'}$.

So two states cannot be distinguished by agent i if they agree on his notes and their email sets look the same to him. Since we assume that the agents do not know anything about the other notes, we do not refer to the sets of notes of the other agents. Note that $\sim_i$ is an equivalence relation.

As an example consider the legal states s_1 and s_2 which are identical apart from their sets of emails:

$$E_{s_1} := \{s(i,l,j)_\emptyset, f(j, s(i,l,j), o)_\emptyset\},$$
$$E_{s_2} := \{s(i,l,j)_{\{k\}}, f(j, s(i,l,j), o)_\emptyset, f(k, s(i,l,j), o)_\emptyset\}.$$

We assume here that $l \in L_i$. The corresponding causality orderings clarify that in the first state agent i sends a message with note l to agent j and then j forwards this message to agent o. Further, in the second state agent i sends the same message but with a BCC to agent k, and then both agent j and agent k forward the message to agent o.

From the above definition it follows that $s_1 \not\sim_i s_2$, $s_1 \sim_j s_2$, $s_1 \not\sim_k s_2$ and $s_1 \not\sim_o s_2$. For example, the first claim holds because, as noticed above, $s(i,l,j)_\emptyset \not\sim_i s(i,l,j)_{\{k\}}$. Intuitively, in state s_1 agent i is aware that he sent a BCC to nobody, while in state s_2 he is aware that he sent a BCC to agent k.

In order to express common knowledge, we define for a group of agents G the relation $\sim_G$ as the reflexive, transitive closure of $\bigcup_{i \in G} \sim_i$. Then we define the truth of a formula from our language in a state inductively as follows, where $s = (E, L)$:

$$\begin{array}{ll}
s \models m & \text{iff } \exists B : m_B \in E \\
s \models i \blacktriangleleft m & \text{iff } \exists B : m_B \in E \text{ and } i \in S(m) \cup R(m) \cup B \\
s \models \neg\varphi & \text{iff } s \not\models \varphi \\
s \models \varphi \wedge \psi & \text{iff } s \models \varphi \text{ and } s \models \psi \\
s \models C_G\varphi & \text{iff } s' \models \varphi \text{ for any legal state } s' \text{ such that } s \sim_G s'
\end{array}$$

We say that φ is ***valid*** (and often just write 'φ' instead of 'φ is valid') if for all legal states s, $s \models \varphi$.

The following lemma clarifies when specific formulas are valid. In the sequel we shall use these observations implicitly.

Lemma 1.

(i) $m \to m'$ *is valid iff* $m = m'$ *or* m' *is part of the message* m.
(ii) $m \to i \blacktriangleleft m'$ *is valid iff* $i \in S(m') \cup R(m')$ *or for some note* l *and group* G, $f(i, l.m', G)$ *is part of the message* m.

The second item states that $m \to i \blacktriangleleft m'$ is valid either if i is a sender or a receiver of m' (in that case actually $i \blacktriangleleft m'$ is valid) or i forwarded the message m'. The latter is also possible if i was a BCC receiver of m'. The claimed equivalence holds thanks to condition L.1.

To illustrate this definition let us return to the above example. In state s_2 agent j does not know that agent k received the message $s(i,l,j)$ since he cannot distinguish s_2 from the state s_1 in which agent k did not receive this message. So $s_2 \models \neg K_j k \blacktriangleleft s(i,l,j)$ holds.

On the other hand, in every legal state s_3 such that $s_2 \sim_o s_3$ both an email $f(k, s(i,l,j), o)_C$ and a 'justifying' email $s(i,l,j)_B$ have to exist such that $s(i,l,j)_B \prec f(k, s(i,l,j), o)_C$ and $k \in B$. Consequently $s_3 \models k \blacktriangleleft s(i,l,j)$, so $s_2 \models K_o k \blacktriangleleft s(i,l,j)$ holds, so by sending the forward agent k revealed himself to o as a BCC recipient.

We leave to the reader checking that both $s_2 \models C_{\{k,o\}} k \blacktriangleleft s(i,l,j)$ and $s_2 \models \neg C_{\{j,o\}} k \blacktriangleleft s(i,l,j)$ holds. In words, agents k and o have common knowledge that agent k was involved in a full version of the message $s(i,l,j)$, while the agents j and o don't.

4 Common knowledge

We now clarify when a group of agents acquires common knowledge of the formula expressing that an email was sent. This shows how we can use our framework to investigate epistemic consequences of email exchanges.

Given a set of emails E and a group of agents A, let

$$E_A := \{m_B \in E \mid A \subseteq S(m) \cup R(m) \text{ or } \exists j \in B : (A \subseteq S(m) \cup \{j\})\}.$$

When $e \in E_A$ we shall say that the email e is ***shared by the group*** A. Note that when $|A| \geq 3$, then $e \in E_A$ iff $A \subseteq S(m) \cup R(m)$. When $|A| = 2$, then $e \in E_A$ also when $\exists j \in B : A = S(m) \cup \{j\}$, and when $|A| = 1$, then $e \in E_A$ also when $A = S(m)$ or $\exists j \in B : A = \{j\}$.

The following theorem summarizes our results.

Main Theorem Consider a legal state $s = (E, L)$ and a group of agents A.

(i) $s \models C_A m$ iff there is $m'_{B'} \in E_A$ such that $m' \to m$ is valid.
(ii) Suppose that $|A| \geq 3$. Then $s \models C_A m_B$ iff the following hold:
C1 $Ag = S(m) \cup R(m) \cup B$,
C2 for each $i \in B$ there is $m'_{B'} \in E_A$ such that $m' \to i \blacktriangleleft m$ is valid,
C3 there is $m'_{B'} \in E_A$ such that $m' \to m$ is valid.

In words, $s \models C_A m_B$ iff

- the email m_B involves all agents,
- there is an email shared by the group A that proves the existence of the message m,
- for every agent i that is on the BCC list of m_B there is an email shared by the group A that proves that i forwarded message m.

As an aside let us mention that there is a corresponding result for the case when $|A| < 3$, as well. However, it involves a tedious case analysis concerning the relation between $A, S(m), R(m)$ and B, so we do not present it here.

5 Proof of the Main Theorem

We establish first a number of auxiliary lemmas. We shall use a new strict partial ordering on emails. We define

$$m_B < m'_{B'} \text{ iff } m \neq m' \text{ and } m' \to m.$$

Note that $m' \to m$ precisely if m' is a forward, or a forward of a forward, etc, of m. Then for two emails m_B and $m_{B'}$ from a legal state s with the causality ordering $\prec$, $m_B < m_{B'}$ implies $m_B \prec m_{B'}$ on the account of condition L.1. However, the converse does not need to hold since $m_B \prec m_{B'}$ can hold on the account of L.2 or L.3. Further, note that the $<$-maximal elements of E are precisely the emails in E that are not forwarded.

Given a set of emails E and $E' \subseteq E$ we then define the ***downward closure*** of E' by

$$E'_{\leq} := E' \cup \{e \in E \mid \exists e' \in E' : e < e'\}.$$

The set of emails E on which the downward closure of E' depends will always be clear from the context.

Next, we introduce two operations on states. Assume a state (E, L) and an email $m_B \in E$.

We define the state

$$s \setminus m_B := (E \setminus \{m_B\}, L'),$$

with

$$L'_i := \begin{cases} L_i \cup FI(m) \text{ if } i \in R(m) \cup B \\ L_i \qquad\qquad \text{otherwise} \end{cases}$$

Intuitively, $s \setminus m_B$ is the result of removing the email m_B from the state s, followed by augmenting the sets of notes of its recipients in such a way that they initially already had the knowledge they would have acquired from m_B. Note that $s \setminus m_B$ is a legal state if m_B is an $<$-maximal element of E.

Next, given $C \subseteq B$ we define the state

$$s[m_{B \mapsto C}] := (E \setminus \{m_B\} \cup \{m_C\}, L'),$$

with

$$L'_i := \begin{cases} L_i \cup FI(m) \text{ if } i \in B \setminus C \\ L_i \qquad\qquad \text{otherwise} \end{cases}$$

Intuitively, $s[m_{B \mapsto C}]$ is the result of shrinking the set of BCC recipients of m from B to C, followed by an appropriate augmenting of the sets of notes of the agents that no longer receive m.

Note that $s[m_{B \mapsto C}]$ is a legal state if there is no forward of m by an agent $i \in B \setminus C$, i.e., no email of the form $f(i, l.m, G)_D$ exists in E such that $i \in B \setminus C$.

We shall need the following lemma that clarifies the importance of the set E_A of emails.

Lemma 2. *Consider a legal state $s = (E, L)$ and a group of agents A. Then for some L' the state $s' := ((E_A)_{\leq}, L')$ is legal and $s \sim_A s'$.*

Proof. We prove that for all $<$-maximal emails $m_B \in E$ such that $m_B \notin E_A$ (so neither $A \subseteq S(m) \cup R(m)$ nor $\exists i \in B : (A \subseteq S(m) \cup \{i\})$) we have $s \sim_A s \setminus m_B$. Iterating this process we get the desired conclusion.

Suppose m_B is a $<$-maximal email in E such that $m_B \notin E_A$. Take some $j \in A \setminus (S(m) \cup R(m))$. Suppose first $j \notin B$. Then $s \sim_j s \setminus m_B$ so $s \sim_A s \setminus m_B$.

Suppose now $j \in B$. Define

$$s_1 := s[m_{B \mapsto \{j\}}].$$

Then s_1 is a legal state and $s \sim_j s_1$. Next, define

$$s_2 := s[m_{B \mapsto \emptyset}].$$

Now take some $k \in A \setminus (S(m) \cup \{j\})$. Then $s_1 \sim_k s_2 \sim_j s \setminus m_B$ so $s \sim_A s \setminus m_B$. Note that both s_1 and s_2 are legal states since m_B is $<$-maximal. □

Using the above lemma we now establish two auxiliary results concerning common knowledge of the formula $i \blacktriangleleft m$ or of its negation.

Lemma 3.

(i) $s \models C_A i \blacktriangleleft m$ *iff* $\exists m'_B \in E_A : (m' \to i \blacktriangleleft m)$
or $(A \subseteq S(m) \cup \{i\}$ *and* $\exists m_B \in E_A : (i \in B))$.
(ii) $s \models C_A \neg i \blacktriangleleft m$ *iff* $s \models \neg\, i \blacktriangleleft m$ *and* $(A \subseteq S(m) \cup \{i\}$ *or* $s \models C_A \neg m)$.

To illustrate various alternatives listed in (i) note that each of the following emails in E ensures that $s \models C_{\{j\}} i \blacktriangleleft m$, where in each case m is the corresponding send message:

$$s(i, l, G)_{\{j\}}, f(k, q.s(i, l, G), H)_{\{j\}},$$
$$s(k, l, i)_{\{j\}},\ f(i, q.s(k, l, G), H)_{\{j\}},\ s(j, l, G)_{\{i\}}.$$

The first four of these emails imply $s \models C_{\{j\}} i \blacktriangleleft m$ by the first clause of (i), the last one by the second clause.

Proof. (i) $(\Rightarrow)$ Suppose $s \models C_A i \blacktriangleleft m$. Take the legal state s' constructed in Lemma 2. Then $s \sim_A s'$, so $s' \models i \blacktriangleleft m$.

Hence for some group B we have $m_B \in (E_A)_{\leq}$ and $i \in S(m) \cup R(m) \cup B$. Three cases arise.

Case 1. $i \in S(m) \cup R(m)$.

Then $m \to i \blacktriangleleft m$. So if $m_B \in E_A$, then the claim holds. Otherwise some email $m'_{B'} \in E_A$ exists such that $m_B < m'_{B'}$. Consequently $m' \to m$ and hence $m' \to i \blacktriangleleft m$. So the claim holds as well.

Case 2. $i \notin S(m) \cup R(m)$ and $A \subseteq S(m) \cup \{i\}$.

Then $i \in B$ since $i \in S(m) \cup R(m) \cup B$. Then by the definition of E_A, $m_B \in E_A$ so the claim holds.

Case 3. $i \notin S(m) \cup R(m)$ and $\neg(A \subseteq S(m) \cup \{i\})$.

If for some note l and groups G and C we have $f(i, l.m, G)_C \in (E_A)_{\leq}$, then either $f(i, l.m, G)_C \in E_A$ or for some $m'_{B'} \in E_A$ we have $f(i, l.m, G)_C < m'_{B'}$. In the former case we use the fact that the implication $f(i, l.m, G) \rightarrow i \blacktriangleleft m$ is valid. In the latter case $m' \rightarrow f(i, l.m, G)$ and hence $m' \rightarrow i \blacktriangleleft m$. So in both cases the claim holds.

Otherwise let $s'' = s'[m_{B \mapsto B \setminus \{i\}}]$. Note that s'' is a legal state because i does not forward m in s'. Take some $j \in A \setminus (S(m) \cup \{i\})$. Then $s' \sim_j s''$, so $s \sim_A s''$. Moreover, $s'' \models \neg i \blacktriangleleft m$, which yields a contradiction. So this case cannot arise.

($\Leftarrow$) The claim follows directly by the definition of semantics. We provide a proof for one representative case. Suppose that for some email $m'_B \in E_A$ both $A \subseteq S(m') \cup R(m')$ and $m' \rightarrow i \blacktriangleleft m$. Take some legal state s' such that $s \sim_A s'$. Then for some group B' we have $m'_{B'} \in E_{s'}$. So $s' \models m'$ and hence $s' \models i \blacktriangleleft m$. Consequently $s \models C_A i \blacktriangleleft m$.

(*ii*) Let $s = (E, L)$.
($\Rightarrow$) Suppose $s \models C_A \neg i \blacktriangleleft m$. Then $s \models \neg i \blacktriangleleft m$. Assume $A \not\subseteq S(m) \cup \{i\}$ and $s \not\models C_A \neg m$. Then there is some legal state $s' = (E', L')$ such that $s \sim_A s'$ and $s' \models m$. Then there is some group B such that $m_B \in E'$. Let $j \in A \setminus (S(m) \cup \{i\})$ and let $s'' = (E' \setminus \{m_B\} \cup \{m_{B \cup \{i\}}\}, L')$. Then $s' \sim_j s''$ so $s \sim_A s''$. But $s'' \models i \blacktriangleleft m$ which contradicts our assumption.

($\Leftarrow$) Suppose that $s \models \neg i \blacktriangleleft m$ and either $A \subseteq S(m) \cup \{i\}$ or $s \models C_A \neg m$. We first consider the case that $A \subseteq S(m) \cup \{i\}$. Let s' be any legal state such that $s \sim_A s'$. Assume $s' \models i \blacktriangleleft m$. Then $m_B \in E_{s'}$ for some group B such that $i \in B$. Since $A \subseteq S(m) \cup \{i\}$, any legal state s'' such that $s' \sim_A s''$ contains an email $m_C \in E_{s''}$ for some group C such that $i \in C$. So $s'' \models i \blacktriangleleft m$. In particular, this holds for the state s, which contradicts our assumption. So $s' \models \neg s(i, n, G)$ and hence $s \models C_A \neg s(i, n, G)$.

Now we consider the case that $s \models C_A \neg m$. Let s' be such that $s \sim_A s'$. Then $s' \models \neg m$. Since $i \blacktriangleleft m \rightarrow m$ is valid, we get $s' \models \neg i \blacktriangleleft m$. So $s \models C_A \neg i \blacktriangleleft m$. $\square$

We are now ready to prove the Main Theorem.

Proof

(*i*) ($\Rightarrow$) Suppose $s \models C_A m$. Take the legal state s' constructed in Lemma 2. Then $s \sim_A s'$, so $s' \models m$. So for some group B we have $m_B \in (E_A)_{\leq}$.

Hence either $m_B \in E_A$ or some email $m'_{B'} \in E_A$ exists such that $m_B < m'_{B'}$. In both cases the claim holds.

($\Leftarrow$) Suppose that for some email $m'_B \in E_A$ we have $m' \rightarrow m$. Take some legal state s' such that $s \sim_A s'$. Then by the form of E_A and the definition of semantics for some group B' we have $m'_{B'} \in E_{s'}$. So $s' \models m'$ and hence $s' \models m$. Consequently $s \models C_A m$.

(ii) By the definition of m_B, the fact that the C_A operator distributes over the conjunction, part (i) of the Main Theorem and Lemma 3 we have

$$s \models C_A m_B \text{ iff } \mathbf{C3}\text{-}\mathbf{C6},$$

where

C4 $\bigwedge_{i \in S(m) \cup R(m) \cup B}$ $((A \subseteq S(m) \cup \{i\}$ and $\exists B' : (m_{B'} \in E_A$ and $i \in B'))$ or $\exists m'_{B'} \in E_A : (m' \to i \blacktriangleleft m))$,
C5 $\bigwedge_{i \notin S(m) \cup R(m) \cup B}$ $(A \subseteq S(m) \cup \{i\}$ or $s \models C_A \neg m)$,
C6 $s \models \bigwedge_{i \notin S(m) \cup R(m) \cup B} \neg i \blacktriangleleft m$.

$(\Rightarrow)$ Suppose $s \models C_A m_B$. Then properties **C3**-**C6** hold. But $|A| \geq 3$ and $s \models C_A m$ imply that no conjunct of **C5** holds. Hence property **C1** holds.

Further, since $|A| \geq 3$ the first disjunct of each conjunct in **C4** does not hold. So the second disjunct of each conjunct in **C4** holds, which implies property **C2**.

$(\Leftarrow)$ Suppose properties **C1**-**C3** hold. It suffices to establish properties **C4**-**C6**.

For $i \in S(m) \cup R(m)$ we have $m \to i \blacktriangleleft m$. So **C2** implies property **C4**. Further, since **C1** holds, properties **C5** and **C6** hold vacuously. □

6 Analysis of BCC

In our framework we built emails out of messages using the BCC feature. So it is natural to analyze whether and in what sense the emails can be reduced to messages without BCC recipients.

Given a send email $s(i, l, G)_B$, where $B = \{j_1, \ldots, j_k\}$, we can simulate it by the following sequence of messages:

$$s(i, l, G), f(i, s(i, l, G), j_1), \ldots, f(i, s(i, l, G), j_k).$$

Analogous simulation can be formed for the forward email $f(i, l.m, G)_B$.

In what follows we clarify in what sense this simulation is correct. Below, given a message m we write $f(S(m), m, j)$ for $f(i, m, \{j\})$, where $S(m) = \{i\}$.

Definition 1. *Given a state $s = (E, L)$ such that there is no forward of the message m by the agent j in E, we define $rem_j^m(s)$ as follows:*

- *if $m_B \in E$ for some group B and $j \in B$ and $f(S(m), m, j)_C \notin E$ for any group C then*

$$rem_j^m(s) := (E \backslash m_B \cup \{m_{B \backslash \{j\}}, f(S(m), m, j)_\emptyset\}, L),$$

- *if $m_B \in E$ for some group B and $j \in B$ and $f(S(m), m, j)_C \in E$ for some group C then*

$$rem_j^m(s) := (E \backslash m_B \cup \{m_{B \backslash \{j\}}\}, L),$$

– *otherwise* $rem_j^m(s) := s$.

So, assuming $m_B \in s$ and $j \in B$, we form $rem_j^m(s)$ by replacing the email m_B by $m_{B\setminus\{j\}}$ when for some group C the forward $f(S(m), m, j)_C$ is present in E, or by $m_{B\setminus\{j\}}, f(S(m), m, j)_\emptyset$ when no such forward is present in E.

We assume that in E there is no forward of m by agent j, as otherwise the removal of j from the list of the BCC recipients would yield an illegal state. Indeed, for such a forward of the message m condition L.1 would not hold. In the remainder of this section we assume that such forwards by former BCC recipients are not present.

We are currently working on a formal analysis of a simulation of BCC that does allow such forwards. It is obtained by replacing each such forward $f(j, m, G)$ by $f(j, f(i, m, j), G)$.

We now show that using the above operation $rem_j^m(s)$ we obtain a legal state that is almost equivalent to the original one. We establish first two lemmas concerning the relation between $rem_j^m(s)$ and the knowledge relation of some agent k.

Lemma 4. *For any two legal states s and t, message m and agent j, if $s \sim_k t$ then $rem_j^m(s) \sim_k rem_k^m(t)$.*

Proof. Omitted for the reasons of space. □

Lemma 5. *For any legal state s, message m and agent j, if there is some t' such that $rem_j^m(s) \sim_k t'$ then either $s \sim_k rem_j^m(s)$ or there is some t such that $s \sim_k t$ and $t' = rem_j^m(t)$.*

Proof. Let $s' = rem_j^m(s)$ and suppose $s' \sim_k t'$. If $s = s'$ then $s \sim_k s'$. Suppose otherwise. Then by the definition of $rem_j^m(s)$ we know that there is some group B such that $j \in B$, $m_B \in E_s$ and $m_{B\setminus\{j\}} \in E_{s'}$. Define $B' = B\setminus\{j\}$.

Suppose there is no full version of $f(S(m), m, j)$ in $E_{t'}$. By the definition of $rem_j^m(s)$, there is a full version of $f(S(m), m, j)$ in $E_{s'}$ so then we know that $k \notin S(m) \cup \{j\}$ because $s' \sim_k t'$. Clearly then $s \sim_k s'$.

Suppose there is a full version of $f(S(m), m, j)$ in $E_{t'}$. Then there is some group C such that $m_C \in E_{t'}$. Suppose $j \in C$. Since $m_{B'} \in E_{s'}$, $j \notin B'$ and $s' \sim_k t'$ this means $k \notin S(m) \cup \{j\}$. So $s \sim_k s'$.

Finally, suppose that $m_C \in E_{t'}$, $j \notin C$ and $f(S(m), m, j)_{C'} \in E_{t'}$. Suppose there is no full version of $f(S(m), m, j)$ in E_s. Define t as the state which is like t' but with $E_t = E_{t'}\setminus\{m_C, f(S(m), m, j)_{C'}\} \cup \{m_{C\cup\{j\}}\}$. Clearly, $rem_j^m(t) = t'$. We claim $s \sim_k t$. The condition on the notes is satisfied since the sets of notes in s and s' and in t and t' are identical, and $s' \sim_k t'$. We will to show that for any $m'_D \in E_s$ such that $k \in S(m') \cup R(m') \cup D$ there is some $m'_{D'} \in E_t$ such that $m'_D \sim_k m'_{D'}$. The proof in the other direction is very similar. Take such an m'_D.

Suppose $m'_D = m_B$. We know $m_{B'} \in E_{s'}$ so $m_{B'} \sim_k m_C$. Since $B = B' \cup \{j\}$ then clearly $m_B \sim_k m_{C\cup\{j\}}$ and we know $m_{C\cup\{j\}} \in E_t$ so let $m'_{D'} = m_{C\cup\{j\}}$.

Suppose otherwise. Then $m'_D \in E_{s'}$ so there is $m'_{D'} \in E_{t'}$ such that $m'_D \sim_k m'_{D'}$. We know that $m' \neq m$ and $m' \neq f(S(m), m, j)$ because no full version of $f(S(m), m, j))$ is in E_s so then $m'_{D'} \in E_t$.

Finally, suppose that for some group E, $f(S(m), m, j)_E \in E_s$. Let $E_t = E_{t'} \backslash \{m_C\} \cup \{m_{C \cup \{j\}}\}$. The proof is very similar.
For the case that $m'_D = f(S(m), m, j)_E$, note that $f(S(m), m, j)_E \in E_{s'}$ so $f(S(m), m, j)_E \sim_k f(S(m), m, j)_{C'}$, so let $m'_{D'} = f(S(m), m, j)_{C'}$. □

The theorem below shows that our operation of removing a BCC recipient results in a state that is equivalent for all formulas that do not explicitly mention the newly added forward or the fact that this BCC recipient received the original message.

Theorem 1. *For any state s, message m, agent j and formula φ that does not mention $j \blacktriangleleft m$ or $f(i, m, j)$, $s \models \varphi$ iff $rem^m_j(s) \models \varphi$.*

Proof. We proceed by induction on the structure of φ. The only interesting case is when $\varphi = C_G\psi$.

Suppose $rem^m_j(s) \models C_G\psi$. Let $s \sim_G t$ for some group of agents G. Then there must be a path $s \sim_{j_1} s_1 \sim_{j_2} ... \sim_{j_n} t$, with $j_1, ..., j_n \in G$. Then by Lemma 4, $rem^m_j(s) \sim_{j_1} rem^m_j(s_1) \sim_{j_2} ... \sim_{j_n} rem^m_j(t)$. Hence $rem^m_j(s) \models C_G\psi$ implies that $rem^m_j(t) \models \psi$. By the induction hypothesis, $t \models \psi$. So $s \models C_G\psi$.

Suppose $s \models C_G\psi$. If $s \sim_G rem^m_j(s)$ then clearly $rem^m_j(s) \models C_G\psi$. Suppose otherwise. Let $rem^m_j(s) \sim_G t'$ for some state t'. Then there is a path $rem^m_j(s) = s'_0 \sim_{k_1} s'_1 \sim_{k_2} ... \sim_{k_n} s'_n = t'$, with $k_1, ..., k_n \in G$. We claim that for any s'_i there is a state s_i such that $s \sim_G s_i$ and $rem^m_j(s_i) = s'_i$. We will proceed by induction. Clearly the claim holds for $s'_0 = rem^m_j(s)$. Suppose it holds for s'_{i-1}, so $s \sim_G s_{i-1}$ and $rem^m_j(s_{i-1}) = s'_{i-1}$ for some state s_{i-1}. By Lemma 5 either $s_{i-1} \sim_{k_i} s'_{i-1}$ or there is s_i such that $s_{i-1} \sim_{k_i} s_i$ and $rem^m_j(s_i) = s'_i$. In the first case, since $s \sim_G s_{i-1}$ and $k_i \in G$ we have $s \sim_G s'_{i-1}$ and since $rem^m_j(s) \sim_G s'_{i-1}$ we have $s \sim_G rem^m_j(s)$ which contradicts our assumption. In the second case, $s \sim_G s_i$ so our claim holds. So then it also holds for $s'_n = t'$, and there is some t such that $s \sim_G t$ and $rem^m_j(t) = t'$. But then by assumption $t \models \psi$ and by the induction hypothesis $t' \models \psi$. So $rem^m_j(s) \models C_G\psi$.

Clearly, by repeatedly applying above construction we obtain the simulation of BCC given above. The corollary below shows that in the original and the resulting state the status of the statement that there is common knowledge of the underlying message is the same.

Definition 2. *For a state s, a message m and a group of agents $B = \{j_1, ..., j_n\}$ such that $m_B \in E_s$, we define*

$$rem^m_B(s) := rem^m_{j_1}(rem^m_{j_2}(...rem^m_{j_n}(s))).$$

Corollary 1. *For any legal state s, a group of agents A and an email $m_B \in E_s$ such that $rem^m_B(s)$ is a legal state*

$$s \models C_A m \text{ iff } rem^m_B(s) \models C_A m.$$

7 Conclusions and future work

Email is by now one of the most common forms of group communication. This motivates the study here presented. The language we introduced allowed us to discuss various fine points of email communication, notably forwarding and the use of BCC. The epistemic semantics we proposed aimed at clarifying the knowledge-theoretic consequences of this form of communication. Our presentation focused on the issue of common knowledge aimed at clarifying when a group of agents has a common knowledge of an email.

This framework also leads to natural questions concerning axiomatization of the language and decidability of the semantics. Currently we work on

- a sound and complete axiomatization of the epistemic language $\mathfrak{L}$ of Section 3; at this stage we have such an axiomatization for the epistemic free formulas,
- the problem of decidability of the truth definition given in Section 3; at this stage we have a decidability result for positive formulas.

Acknowledgements

We acknowledge helpful discussions with Jan van Eijck and Rohit Parikh and useful referee comments.

References

1. K. R. Apt, A. Witzel, and J. A. Zvesper. Common knowledge in interaction structures. In *Proceedings of TARK XII*, pages 4–13. The ACM Digital Library, 2009.
2. E-mail leak of degree inflation. BBC News, 2008. Available at `http://news.bbc.co.uk/2/hi/uk_news/education/7483330.stm`.
3. J. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.
4. K. M. Chandy and J. Misra. How processes learn. *Distributed Computing*, 1(1):40–52, March 1986.
5. R. Fagin, J. Halpern, M. Vardi, and Y. Moses. *Reasoning about knowledge*. MIT Press, Cambridge, MA, USA, 1995.
6. L. Lamport. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM*, 21(7):558–565, 1978.
7. Wouter Bos e-mailt per ongeluk zijn netwerk rond. NRC Handelsblad, 7th October 2010, 2010. In Dutch.
8. E. Pacuit. Logics of informational attitudes and informative actions. Manuscript, University of Tilburg, 2010.
9. E. Pacuit and R. Parikh. Reasoning about communication graphs. *Interactive Logic. Proceedings of the 7th Augustus de Morgan Workshop*, pages 135–157, 2007.
10. R. Parikh and R. Ramanujam. A knowledge based semantics of messages. *Journal of Logic, Language and Information*, 12(4):453–467, 2003.
11. J. van Eijck and F. Sietsma. Message passing in a dynamic epistemic logic setting. In *Proceedings of TARK XIII*. The ACM Digital Library, 2011. To appear.
12. Y. Wang, F. Sietsma, and J. van Eijck. Logic of information flow on communication channels. In *Proceedings of AAMAS-10*, pages 1447–1448, 2010.

An Argument for an Analogical Perspective on Rationality & Decision-Making

Tarek R. Besold, Helmar Gust, Ulf Krumnack, Ahmed Abdel-Fattah, Martin Schmidt, and Kai-Uwe Kühnberger

Institute of Cognitive Science
University of Osnabrück
49069 Osnabrück, Germany

Abstract. Humans are without any doubts the prototypical example of agents that can hold rational beliefs and can show rational behaviour. When modeling human decision-making, it seems reasonable to take the remarkable abilities of humans into account with respect to rational behaviour, but also the apparent deficiencies of humans shining up in certain rationality tasks. Based on well-known challenges for human rationality, together with results from psychological studies on decision-making and from the field of computational modeling of analogy-making, we argue that analysis and modeling of rational belief and behaviour should also consider cognitive mechanisms like analogy-making and coherence maximization of the background theory.

1 Introduction

At times, human behaviour seems erratic and irrational. Still, it is widely undoubted that humans can act rational and, in fact, appear to act rational most of the time. In explaining behaviour, we use terms like beliefs and desires. If an agent's behaviour makes the most sense to us, then we interpret it as a reasonable way to achieve the agent's goals given his beliefs. We take this as indication that some concept of rationality does play a crucial role when describing and explaining humans' behaviour in a large variety of situations.

Based on ideas from vernacular psychology, in many cases rational beliefs are interpreted as a foundation of rational behavior. In this extended position paper, we will be mostly concerned with beliefs and knowledge, i.e. the epistemic aspects of rationality.

In the following, we want to shed light on some aspects of rationality from a mostly computational cognitive science point of view. Although, even in psychology or economics there is no generally accepted formal framework for rationality, we will argue for a model that links rationality to the ability of humans to establish analogical relations. This is an attempt for proposing a new perspective and framework for rationality. Furthermore, in the course of a mostly overview-like presentation, we want to give some hints at how already existing frameworks for computational analogy-making integrate some aspects considered characteristic for human decision making.

2 Rationality Concepts and Challenges

2.1 Rationality

Many quite distinct frameworks for modeling rationality have been proposed, and an attempt at clustering these frameworks to the best of our knowledge results in at least four classes: logic-based models (cf. e.g. [1]), probability-based models (cf. e.g. [2]), heuristic-based models (cf. e.g. [3]), and game-theoretically based models (cf. e.g. [4]).

Several of these models have been considered for establishing a normative theory of rationality, not only tryinig to model "rational behaviour", but also to offer predictive power for determining whether a certain belief, action, or behaviour may be considered rational or not. Also, every of these theories specifies some sort of *definition* of rationality. Unfortunately, when comparing the distinct frameworks, it shows that these definitions are in many cases almost orthogonal to each other (as are the frameworks). Therefore, in this paper, we will propose certain cognitive mechanisms for explaining and specifying rationality in an integrated, more homogeneous way.

2.2 Well-Known Challenges

Although the aforementioned frameworks have gained merit in modeling certain aspects of human intelligence, the generality of each such class of frameworks has at the same time been challenged by psychological experiments. For example, in the famous Wason-selection task [5] human subjects fail at a seemingly simple logical task (cf. Table 1). Also, experiments by Byrne on human reasoning with conditionals [6] indicated severe deviations from classical logic (cf. Table 1). Similarly, Tversky and Kahneman's Linda problem [7] illustrates a striking violation of the rules of probability theory (cf. Table 1). Heuristic approaches to judgment and reasoning [8] are often seen as approximations to a rational ideal and in some cases could work in practice, but often lack formal transparency and explanatory power. Game-based frameworks are questioned due to the lack of a unique concept of optimality in game-theory that can support different "rational behaviors" for one and the same situations (e.g. Pareto optimality vs. Nash equilibrium vs. Hick's optimality etc., [9]).

Wason Selection Task: This task shows that a large majority of subjects are seemingly unable to verify or to falsify a simple logical implication: "If on one side of the card there is a D, then on the other there is the number 3". In order to check this rule, subjects need to turn D and 7, i.e. subjects need to check the direct rule application and the contrapositive implication. After a slight modification of the content of the rule (content-change), while keeping the structure of the problem isomorphic, subjects perform significantly better: In [11], the authors show that a change of the abstract rule "$p \rightarrow q$" to a well-known problem significantly increases correct answers of subjects. The authors use the rule "If a person is drinking beer, then he must be over 20 years old." The cards used in the task were "drinking beer", "drinking coke", "25 years old", and "16 years old". Solving this task according to the rules of classical logic comes down to turning "drinking beer" and "16 years old".

Inferences and Conditionals: Also Byrne's observations question whether human reasoning can be covered by a classical logic-based framework. Presented with the information given in Table 1, from 1. 46% of subjects conclude that Marian will not study

Wason-Selection Task [10]: Subjects are given the rule "Every card which has a D on one side has a 3 on the other side." and are told that each card has a letter on one side and a number on the other side. Then they are presented with four cards showing respectively D, K, 3, 7, and asked to turn the minimal number of cards to determine the truth of the sentence.
Inferences and Conditionals [6]: 1. If Marian has an essay to write, she will study late in the library. She does not have an essay to write. 2. If Marian has an essay to write, she will study late in the library. She has an essay to write. 3. If Marian has an essay to write, she will study late in the library. She has an essay to write. If the library stays open, she will study late in the library.
Linda-Problem [7]: Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Linda is a teacher in elementary school. Linda works in a bookstore and takes Yoga classes. Linda is active in the feminist movement. (F) Linda is a psychiatric social worker. Linda is a member of the League of Women Voters. Linda is a bank teller. (T) Linda is an insurance salesperson. Linda is a bank teller and is active in the feminist movement. (T&F)

Table 1. The Wason-selection task questions whether humans reason in such situations according to the laws of classical logic. Byrne's experiments on how humans handle conditionals also shed doubt on a logic-based model. Tversky and Kahneman's Linda problem questions the ability of humans to reason according to the laws of probability theory.

late in the library, erring with respect to classical logic (as denial of the antecedent does not validate a negation of the consequent). Also, from 2. 96% of subjects conclude that Marian will study late in the library, whilst only 38% of subjects reach the same conclusion from 3.. Thus an introduction of another antecedent (without any indication that the antecedent should not hold) dramatically reduced the number of subjects applying a simple modus ponens in their process of forming a conclusion.

Linda Problem: With respect to the Linda problem it seems to be the case that subjects are amenable to the so-called conjunction fallacy: subjects are told a story specifying a particular profile about the bank teller Linda. Then, eight statements about Linda are shown and subjects are asked to order them according to their probability (cf. Table 1). 85% of subjects decide to rank the eighth statements "Linda is a bank teller and active in the feminist movement" (T & F) as more probable than the sixth statement "Linda is a bank teller" (T). This ranking contradicts basic laws of probability theory, as the joint probability of two events (T & F) is less or at most equal to the probability of each individual event.

Classical Resolution Strategies: Strategies that have been proposed to address the mentioned challenges include non-classical logics for modeling subjects' behavior in the Wason-Selection task [12], or a switch from (syntactic) deductions to reasoning in semantic models [13]. Still, these are only individual case-based solutions, which do not (or only hardly) generalize, and thus don't provide a basis for a unified theory or the genesis of a generally accepted broad concept of rationality.

3 Non-Standard Interpretations of Challenges for Rationality

An immediate reaction to the challenges for rationality depicted above may be to deny that humans are always able to correctly reason according to the laws of classical logic or the laws of probability theory. Still, concluding that human behaviour therefore is irrational in general does not seem convincing. The most that can be concluded from the experiments is that human agents are neither deduction machines nor probability estimators, but perform their undisputable reasoning capabilities with other means. From our point of view, subjects' behavior in the described tasks is connected to certain cognitive mechanisms that are used by humans in such reasoning tasks, giving rise to the emergence of behavior commonly described as rational.

3.1 Interlude: Analogy and Analogical Reasoning

Analogies can basically be described as claims of similarity, which are often used in argumentation or when explaining complex situations. Putting it more formally, analogy-making refers to the human ability of perceiving dissimilar domains as similar with respect to certain aspects based on shared commonalities in relational structure or appearance. Analogy and analogy-making research has received growing attention during the last decades, changing the perception of analogy from interpreting it as a special and rarely applied case of reasoning to placing it in the center of human cognition itself [14]. The literature on analogies knows a distinction between two subcategories of analogical mapping: attribute mappings (surface mappings) and relational mappings [15]. Whilst both mapping types are standardly assumed to be one-to-one, attribute mappings are based on attributes or surface properties, such as shape or color (i.e., two objects can be said to be similar with respect to a particular attribute or set of attributes), whilst relational mappings are based on relations between objects, such as having the same role or the same effect (i.e., two objects can then be said to be similar with respect to some relation to one or more other objects). Once such an analogical bridge has been established between two domains, analogical reasoning now allows for carrying over inferences from the base to the target domain in order to extend knowledge about the latter, i.e., an inference which holds between elements in the base domain is also assumed to analogically hold between the corresponding elements of the target domain.

3.2 How Analogy-Making Enters the Picture

In a short reply to Colman's article "*Cooperation, psychological game theory, and limitations of rationality in social interaction*" [16], Kokinov challenges traditional views

on rationality [17]. Taking an initial stance similar to Colman's, agreeing on that rationality fails as both, descriptive theory of human-decision making and normative theory for good decision-making, Kokinov reaches a different, more radical conclusion as Colman did before. Instead of trying to fix the concept of rationality by redefining it, adding formerly unconsidered criteria for optimization of some kind, he proposes to replace the concept of rationality as a theory in its own right by a multilevel theory based on cognitive processes involved in decision-making. Where Colman proposes a collection of ad-hoc strategies for explaining the deviations from rationality which people exhibit in their behaviour, Kokinov proposes analogy as means of unifying the different, formerly unconnected parts of Colman's attempt at describing the mechanisms of decision-making. In Kokinov's view, the classical concept of utility making has to be rendered as an emergent property, which will emerge in most, but not all, cases, converting rationality itself in an emergent phenomenon, assigning rational rules the status of approximate explanations of human behavior.

But evidence for a crucial role of analogy in decision-making cannot only be found in conceptual cognitive science, but also in psychological studies on decision-making and choice processes. An overview by Markman and Moreau [18], based on experiments and observations from psychological studies, amongst others on consumer behaviour and political decision-making, reaches the conclusion that there are at least two central ways how analogy-making influences choice processes. Analogies to other domains can provide means of representation for a choice situation, as generally speaking the making of a decision relies on a certain degree of familiarity with the choice setting. In many cases of this kind, analogy plays a crucial role in structuring the representation of the choice situation, and thus may strongly influence the outcome of a decision. Also, structural alignment (a key process of analogy-making) plays a role when comparing the different possible options offered by a decision situation, with new options being learned by comparison to already known ones. An experimental study by Kokinov [19] demonstrated that people actually do use analogies in the process of decision-making, with significant benefit already if only one case is found to be analogous to the choice situation under consideration. Furthermore, evidence has been found that there is no significant difference between close and remote analogies in this process, and that people are not limited to rely only on analogous cases from their own experience, but that also cases which were only witnessed passively (e.g., by being a bystander, or learning about a situation from reports in the media) may have beneficial influence.

Taking all this together, we strongly argue in favor of taking into account cognitive mechanisms centered around the concept of analogy when analysing and modeling rational belief and behaviour in humans. In the following, we want to provide an analogy-inspired point of view on the aforementioned well-known challenges for rationality.

3.3 Resolving the Wason-Selection Task by Cognitive Mechanisms

As mentioned above, according to [11] subjects perform better (in the sense of more according to the laws of classical logic) in the Wason-Selection task, if content-change makes the task easier to access for subjects. In our reading, subjects' performance is tightly connected to establishing appropriate analogies. Subjects perform badly in the

classical version of the Wason-Selection task, simply because they fail to establish a fitting analogy with an already known situation. In the "beer drinking" version mentioned above, i.e. the content-change version of the task, the situation changes substantially, because subjects can do what they would do in an everyday analogous situation: they need to check whether someone younger than 20 years is drinking beer in a bar. This is to check the age of someone who is drinking beer and conversely to check someone who is younger that 20 years whether he is drinking beer or not. In short, the success or failure of managing the task is crucially dependent on the possibility to establish a meaningful analogy.

3.4 Resolving the Inferences and Conditionals Problem by Cognitive Mechanisms

The results concerning conclusions drawn by the subjects in Byrne's experiments can also be explained through analogy-making. People faced with the information given in 1. will recall similar conversations they had before, using these known situations as basis for their decision on what to conclude. According to Grice [20], in conversations speakers are supposed to provide the hearer with as much information as is needed for exchanging the necessary information, a rule which goes in accordance with our everyday observation. Thus, when being given the additional information that "Marian does not have to write an essay.", the set of candidate situations for establishing an analogy will be biased towards situations in which this information had an impact on the outcome, resulting in the conclusion that Marian would not study late in the library either. Regarding 2. and 3., a similar conjecture seems likely to hold: By additionally mentioning the library, similar situations in which the library might actually have played a crucial role (e.g., by being closed) will be taken into account as possible base domains of the analogy, causing the change in conclusions made.

3.5 Resolving the Linda Problem by Cognitive Mechanisms

In case of Tversky and Kahneman's Linda problem, a natural explanation of subjects' behavior is that people find a lower degree of coherence between Linda's profile and the mere statement "Linda is a bank teller", than they do with the expanded statement "Linda is a bank teller and is active in the feminist movement". In the latter one, at least one conjunct of the statement fits quite well to Linda's profile. In short, subjects prefer situations that seem to have a stronger inner coherence. Coherence is important for the successful establishment of an analogical relation, as it facilitates the finding of a source domain for an analogy. We conjecture that in order to make sense of the task, humans rate statements with a higher probability where facts are arranged in a theory with a higher degree of coherence. Thus, seeing coherence in the first place as a means for facilitating analogy-making, and taking into account that analogy has been identified as a core element of human cognition, the decision for the coherence-maximizing option is not surprising anymore, but fits neatly into the conceptual analogy-based framework, and could even have been predicted (providing inductive support for our general claim).

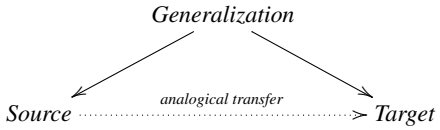


Fig. 1. HDTP's overall approach to creating analogies

4 Rationality, Decision-Making and Analogy-Making Systems

In this section we want to give an overview-like sketch of how computational analogy-making systems can be related to some of the discussed challenges for rationality, as well as to decision-making and choice in general, demonstrating their value as models also in this domain.

4.1 Heuristic-Driven Theory Projection

Heuristic-Driven Theory Projection is a symbolic framework for computing analogical relations between two domains that are axiomatized in first order logic [21]. HDTP, after being given the logic representations of the two domains, by means of anti-unification [22] computes a common generalization of both, and uses this resulting theory as basis for establishing an analogy, also involving analogical transfer of knowledge between the domains (i.e., the system provides an explicit generalization of the two domains as a by-product of the analogy-making process). Thus, conceptually, HDTP proceeds in two phases: in the *mapping phase*, the formal representations of source and target domain are compared to find structural commonalities, and a generalized description is created, which subsumes the matching parts of both domains. In the *transfer phase*, unmatched knowledge in the source domain can be transferred to the target domain to establish new hypotheses in an analogical way, cf. Figure 1.

Think about Rutherford's model of the atom [23] in analogy to a model of the solar system: HDTP, after finding commonalities in the logical representation of the solar system as base domain, and the atom model as target domain (for example, that in both cases less massive objects are somehow related to a more massive central object, or that always a positive distance and a positive force between these lighter objects and the heavier core can be found), a generalization is computed, via which known laws from the base can be reinstantiated in the target (e.g., that a lighter object revolves around a heavier one when there is negative centrifugal force between the lighter and the heavier one, yielding the revolution of the electrons around the nucleus, or that the centrifugal force between two spatially separated objects with positive gravitational force between both is equal to the negative value of that gravity, resulting in stable orbits of the electrons in the model).

HDTP implements a principle (by using heuristics) that maximizes the coverage of the involved domains [21]. Intuitively, this means that the sub-theory of the source (or the target) that can be generated by re-instantiating the generalization is maximized. Putting it the other way round, the original domain-specific information and structure

shall implicitly be preserved as far as possible. The higher the coverage the better, because more support for the analogy is provided by the generalization (in a way, the higher the achieved degree of coverage, the more firmly the analogy is rooted in the underlying domains, used for creating the generalization). A further heuristics in HDTP is the minimization of substitution lengths in the analogical relation, i.e. the simpler the analogy the better [24]. The motivation for this heuristics is to prevent arbitrary associations. Clearly there is a trade-off between high coverage and simplicity of substitutions: An appropriate analogy should intuitively be as simple as possible, but also as general and broad as necessary in order to be non-trivial. Unfortunately, high coverage normally comes with higher complexity of substitutions (as a more complex generalization allows for a higher degree of re-representation of domain-specific structures and information), whilst the simplicity constraint is trying to steer the analogy-making process in exactly the opposite direction. This kind of trade-off is similar to the kind of trade-off that is usually the topic of model selection in machine learning and statistics.

4.2 The Wason-Selection Task Revisited

A modeling of the Wason-Selection task with HDTP is quite simple as long as appropriate background knowledge is available, in case an analogy should be established, or the lack of appropriate background knowledge prevents analogy-making, in case no analogy should be established: On the one hand, if background knowledge for an analogous case is missing (i.e., in the case of HDTP, no domain representation which offers sufficient structural commonalities to the target domain as to serve as a base for the analogy process can be retrieved from memory), then there is no chance to establish an analogical relation. Hence, subjects have to apply other auxiliary strategies, possibly deviating from the expected "right" answer. If there is a source theory with sufficient structural commonalities on the other hand, then the establishment of an analogical relation is straightforward, resulting in a smooth solution process of the task.

4.3 Analogy in Choice

Coming back to Markman and Moreau's meta-study of the role analogy and analogical comparison play in the process of human choice, presented in [18], we want to show some connections of their findings to computational systems for analogy-making.

It is without doubt that the choice of options taken into account when making a decision is of crucial importance for the entire process of decision-making. Markman and Moreau present the formation of consideration sets (i.e., the set of options taken into account by a decision maker) as one of the places at which the influence of analogy on decision-making clearly shines up. An analogical reasoning process is involved when deciding on which scenarios are likely to happen, and thus have to be considered (see, e.g., also [25] for related results). According to their findings, there are different factors influencing which analogies will be used in a choice situation, resulting in a set of analogies which are considered similar or familiar to the current situation. Close analogs have the advantage of probably allowing the transfer of more lower-order relations than distant analogs would, i.e., closer concepts are more likely to be considered as an option due to an easier and more fruitful analogy-making process. This goes in accordance with

characteristics exhibited by many computational models of analogy-making, where we just want to mention HDTP. As pointed out in [21], although HDTP basically aligns any entity, function or predicate, it clearly prefers literally-matching alignments over non-literally ones, and equivalent structures to structural mismatches, thus reconstructing a preference and behaviour also shown by humans.

Also, experiments indicate that commonly shared surface elements of domains are more useful as retrieval cues than are connected relational systems. Also this carries over to the principles underlying HDTP, with HDTP trying to minimize the complexity of analogical relations whilst maximizing the degree of coverage: Connected relational systems have the strong tendency of reaching higher-order stages, whilst direct surface correspondences stay on a low level, allowing for a direct matching of features. Thus, handling common surface elements allows for a certain degree of coverage without having to escalate complexity, probably also making HDTP prefer surface elements for supporting an analogy over relational ones (if both types are equally available).

Finally, it shows that elements related to a person's individual experience do influence the way deicsions are taken. These elements have the advantage of being (mostly) highly accessible, with base domains which form part of someone's past being more likely to have richly connected relational structures, providing good ground for eventual analogical inference. When searching for a way of computationally modeling this phenomenon, it comes to mind that a similar effect can already be found in AMBR, Kokinov's well-documented hybrid analogy-making system [26]. This system exhibits signs of priming effects in the retrieval process of a fitting base domain for an analogy's given target domain, together with a general influence of earlier memory states on later ones.

4.4 Modeling Judgement and Choice

In [27], Petkov and Kokinov present JUDGEMAP, a computational model of judgement and choice based on the general-purpose cognitive architecture DUAL [28], and the aforementioned corresponding AMBR analogy-making system. JUDGEMAP is capable of performing both tasks, giving a judgement on a scale and deciding a choice situation, by means of a process of making forced analogies, exclusively using mapping principles inherited from the underlying AMBR system. JUDGEMAP has been demonstrated to replicate phenomena known from observations of human judgement as, for example, range and frequency effects, or sequential assimilation effects.

Furthermore, several simulations are described, in which it is demonstrated that mechanisms designed for modeling analogy can have influence on judgement and choice, possibly reproducing contextual effects in tasks which don't seem to be related to analogy-making. Among others, it is shown that the pressure for one-to-one mapping, which has been introduced to AMBR for the purpose of analogy-making, can in the model cause phenomena similar to the frequency effect in judgement (i.e., people using all available ratings almost equally often in their judgements), and for the concave form of the functional relation between subjective value (i.e., utility) and money. Also, the effect that humans when judging tend to use middle ratings more often than extreme ones can be explained in terms of a dynamic mechanism used for hypothesis creation in AMBR. Also, the occurence of the preference reversal effect in choice can be explained

by a feature originating from the analogy-making system. The most remarkable part of all this is, that not a single one of the mechanisms used in the JUDGEMAP model had been created for this purpose, but were all obtained from the AMBR model, which supports our claim that structural mapping and analogy play a fundamental role also in judgement and choice, and therefore ultimately also in decision-making.

5 Concluding Remarks

The evidence for a crucial role of analogy-making presented over the last pages falls far from being complete. Yet another example can be given in form of well-known studies on human decision-making under time pressure, which show a change in the applied inference procedure. In [29], the authors report that, whilst the best predicting model of human inference for decision making in an unstressed conditions was a weighted linear model integrating all available information, when time pressure was induced, best predictions were obtained by using a simple lexicographic heuristic [30]. This presumed change from a more complex strategy using complex relational structures to a simple single-attribute-based procedure also can be found in research on analogy-making: In [31], it is reported that anxiety made participants of an analogical-reasoning experiment switch from a preference for complex relational mappings to simple attribute-based mappings.
Still, whilst not claiming completeness of our overview of evidence, we are convinced that even the already given examples and indications are sufficient as not to allow for leaving analogy and cognitive processes out of consideration.

A criticism with respect to the analogy-making approach might be a seeming lack of normativity as a theory. Although work on this topic is still in a very early stage, we are confident that this objection grasps at nothing: Normativity can be introduced in a very natural way by considering the reasonableness (or unreasonableness) of made analogies. Roughly speaking, it is obvious that different analogies may have different degrees of reasonableness, e.g., based on the level to which they result in coherent beliefs and to which they encompass both, the source and the target domain of the analogy.

In this paper, we argue in favor of an introduction of the concept of analogy into conceptual research on rationality and decision-making on a foundational level. Based on a review of some basic concepts and existing work within the fields of analogy research and research on decision-making and choice, together with an exemplifying proposal of new resolution strategies for classical rationality puzzles, we think that the usage of frameworks for establishing analogical relations and the usage of frameworks that can maximize the coherence of a theory necessarily have to be taken into account when modeling (and possibly implementing) what is commonly considered rational belief in a not overly simplified manner.

Of course, this paper is just a very first conceptual step in constructing and establishing the promoted new view, still a great amount of substantial fundamental work has to be done, and numerous open questions have to be answered. Nevertheless, considering the evidence indicating a connection between decision making and analogy originating from psychology, together with characteristics shown by already existing models

of analogy-making (which were designed without any consideration of rationality or an application in decision making), we are strongly confident that an undertaking as argued for in this paper merits the effort, and can lead to important results and insights.

References

1. Evans, J.: Logic and human reasoning: An assessment of the deduction paradigm. Psychological Bulletin **128** (2002) 978–996
2. Griffiths, T., Kemp, C., Tenenbaum, J.: Bayesian Models of Cognition. In: The Cambridge Handbook of Computational Cognitive Modeling. Cambridge University Press (2008)
3. Gigerenzer, G., Hertwig, R., Pachur, T., eds.: Heuristics: The Foundation of Adaptive Behavior. Oxford University Press (2011)
4. Osborne, M., Rubinstein, A.: A Course in Game Theory. MIT Press (1994)
5. Wason, P.C.: Reasoning. In: New Horizons in psychology. Penguin (1966)
6. Byrne, R.: Suppressing valid inferences with conditionals. Cognition **31**(1) (1989) 61–83
7. Tversky, A., Kahneman, D.: Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. Psychological Review **90**(4) (1983) 293–315
8. Gigerenzer, G.: Rationality for Mortals: How People Cope with Uncertainty. Oxford University Press (2008)
9. Chinchuluun, A., Pardalos, P., Migdalas, A., Pitsoulis, L., eds.: Pareto Optimality, Game Theory and Equilibria. Springer (2008)
10. Wason, P.C., Shapiro, D.: Natural and contrived experience in a reasoning problem. The Quarterly Journal of Experimental Psychology **23**(1) (1971) 63–71
11. Cosmides, L., Tooby, J. In: Cognitive Adaptions for Social Exchange. Oxford University Press (1992)
12. Stenning, K., van Lambalgen, M.: Human Reasoning and Cognitive Science. MIT Press (2008)
13. Johnson-Laird, P.: Mental Models. Harvard University Press (1983)
14. Holyoak, K., Gentner, D., Kokinov, B.: Introduction: The Place of Analogy in Cognition. In: The Analogical Mind: Perspectives from Cognitive Science. MIT Press (2001) 1–19
15. Gentner, D.: Structure mapping: A theoretical framework for analogy. Cognitive Science **7** (1983) 155–170
16. Colman, A.M.: Cooperation, psychological game theory, and limitations of rationality in social interaction. Behavioral and Brain Sciences **26**(2) (2003) 139–198
17. Kokinov, B.: Analogy in decision-making, social interaction, and emergent rationality. Behavioral and Braind Sciences **26**(2) (2003) 167–169
18. Markman, A., Moreau, C.: Analogy and analogical comparison in choice. In: The Analogical Mind: Perspectives from Cognitive Science. MIT Press (2001) 363–399
19. Kokinov, B.: Can a Single Episode or a Single Story Change our Willingness to Risk? The Role of Analogies in Decision-Making. In: Advances in Cognitive Economics. NBU Press (2005)
20. Grice, H.P.: Logic and Conversations. In: Syntax and Semantics, Vol. 3: Speech Acts. Academic Press (1975) 41–58
21. Schwering, A., Krumnack, U., Kühnberger, K., Gust, H.: Syntactic principles of heuristic-driven theory projection. Cognitive Systems Research **10**(3) (2009) 251–269
22. Plotkin, G.D.: A note on inductive generalization. Machine Intelligence **5** (1970) 153–163
23. Rutherford, E.: The scattering of α and β particles by matter and the structure of the atom. Philosophical Magazine **21** (1911) 669–688

24. Gust, H., Kühnberger, K., Schmid, U.: Metaphors and heuristic-driven theory projection (hdtp). Theoretical Computer Science **354** (2006) 98–117
25. Schwenk, C.: Cognitive simplification processes in strategic decision-making. Strategic Management Journal **5**(2) (1984) 111–128
26. Kokinov, B., Petrov, A.: Integrating Memory and Reasoning in Analogy-Making: The AMBR Model. In: The Analogical Mind: Perspectives from Cognitive Science. MIT Press (2001) 59–124
27. Petkov, G., Kokinov, B.: JUDGEMAP - integration of analogy-making, judgement, and choice. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci). (2006) 1950–1955
28. Kokinov, B.: The DUAL cognitive architecture: A hybrid multi-agent approach. In Cohn, A., ed.: Proceedings of the Eleventh European Conference of Artificial Intelligence, John Wiley & Sons, Ltd. (1994) 203–207
29. Rieskamp, J., Hoffrage, U.: Inferences under time pressure: How opportunity costs affect strategy selection. Acta Psychologica **127** (2008) 258–276
30. Fishburn, P.: Lexicographics orders, utilities and decision rules: A survey. Management Science **20** (1974) 1442–1471
31. Tohill, J., Holyoak, K.: The impact of anxiety on analogical reasoning. Thinking & Reasoning **6**(1) (2000) 27–40

The developmental paradox of false belief understanding: a dual-system approach

Leon de Bruin[1] and Albert Newen[1]

[1] Ruhr-University of Bochum, Bochum, Germany
lcdebruin@gmail.com
albert.newen@rub.de

Abstract. In our article we explore the developmental paradox of false belief understanding. This paradox follows from the claim that young infants already have an 'implicit' notion of false belief, despite the fact that they consistently fail tests assessing 'explicit' forms of false belief understanding. First, we argue that recent dual-system proposals to solve this paradox are unsatisfactory because they either lack the conceptual resources to deal with the differences between implicit and explicit false belief understanding, or ignore questions about system interaction. Second, we discuss a number of problems dual-system approaches have to address in order to account for the development of false belief understanding, and propose a model that combines a layered model of perspective taking with an inhibition-selection-representation mechanism operating on different levels.

Keywords: Theory of Mind, False Belief Task, implicit false belief understanding, shared representations, decoupling

1. Introduction

In our everyday social interactions we frequently attribute mental states (e.g., beliefs, desires) to others in order to predict or explain their behavior. For example, if Mini knows that Maxi wants the chocolate, and she also knows that Maxi believes the chocolate is in the blue cupboard, she can predict that Maxi will search for the chocolate in the blue cupboard. This works equally well if Mini wants to explain Maxi's behavior: Maxi's searches in the blue cupboard, because he wants the chocolate and believes the chocolate is in the blue cupboard. It is generally accepted that this capacity is enabled by a Theory of Mind (e.g., Perner 1991, Baron-Cohen 1995, Leslie 2000).

Over the last couple of decades, most research in Theory of Mind has focused on the development of false belief understanding. In order to test whether Mini truly understand Maxi's behavior from *his* point of view, we need to make sure that she does not simply attribute to Maxi her *own* belief about the location of the chocolate. The False Belief Test (FBT) has been specifically designed to solve this problem by introducing a condition in which the protagonist has a *false* belief about some state of affairs in the world (usually the location of an object).

In 'explicit' (i.e., verbal) versions of this test, children are asked a direct question about the protagonist's false belief. In the 'unexpected location' FBT (Wimmer & Perner 1983, Baron-Cohen et al. 1985), for example, children observe a protagonist who sees an object being placed in a certain location. Then the protagonist leaves, and the object is moved to another location. When the protagonist returns, he mistakenly believes the object is still in its initial location. At this point, the children are asked to predict where the protagonist will look for the object.

Test results show that 3-year-olds typically give a wrong answer to this question, while four-year-olds answer correctly. Findings on other explicit FBTs, such as the 'unexpected identity' task (Perner et al. 1987; Moses & Flavell 1990; Wellman 1990), confirm this picture. Many researchers have therefore concluded that false belief understanding does not emerge until four years (Flavell 2004, Sodian 2005; see Wellman 2002 for a review, and Wellman et al. 2001 for a meta-analysis).

Recently, however, this conclusion has been challenged on the basis of findings on so-called 'implicit' false belief understanding. Studies based on 'violation of expectation' and 'anticipatory looking' paradigms have claimed that implicit false belief understanding is already present at a considerably earlier age, in 25-month-olds (Southgate et al. 2007), 15-month-olds (Onishi & Baillargeon 2005), and even 13-month-olds (Surian et al. 2007). These findings give rise to a very interesting and widely debated 'developmental paradox': if young infants already understand false belief, as the above studies seem to suggest, then why do they fail the explicit FBT?

Our aim in this article is to propose a model that allows us to account for both implicit and explicit forms of false belief understanding and to shed new light on the developmental paradox. We start by discussing a number of important studies on implicit false belief understanding (section two). Next, we evaluate two recent dual-system accounts of the developmental paradox (Apperly & Butterfill 2009; Baillargeon et al. 2010) in section three. We argue that both accounts fail to solve the developmental paradox because they either lack the conceptual resources to deal with the differences between implicit and explicit forms of false belief understanding, or ignore questions about the interaction between the two systems. In section four, we discuss a number of problems dual-system approaches have to address in order to account for the development of false belief understanding. Section five puts forward a model combining a layered model of perspective taking with an inhibition-selection-representation mechanism that operates on different levels.

2. Implicit False Belief Understanding

The explicit FBT places rather strong demands on children's other cognitive capacities (Bloom & German 2000, Carlson & Moses 2001). Studies on implicit false belief understanding are designed in a way as to reduce these demands in order to see whether children might be capable of false belief understanding at an earlier age. In the implicit FBT, infants no longer have to give an explicit answer to a question about the protagonist's belief. Instead, their understanding of false belief is inferred from the behavior they spontaneously produce (cf. Baillargeon et al. 2010).

Clements & Perner (1994) already showed that linguistic competence is responsible for at least some of the difficulties infants have with the explicit FBT vis-à-vis its implicit counterpart. They adapted an early version of the 'anticipatory looking' paradigm, which is used to test whether children are able to visually anticipate where another agent will search for an object, given his false belief about its location. In their experiment, Clements & Perner (1994) asked children to watch how the protagonist, a mouse called Sam, stored an object in a box in front of one of two mouse holes. While Sam is asleep, the object is moved to a different box in front of the other mouse hole. The experimenters monitored where the children were looking in anticipation of Sam's reappearance, given his false belief about the object's location. They contrasted this spontaneous behavior with the explicit answers children gave to the question where Sam would look for the object.

Clements & Perner (1994) found that 3-year-olds looked at the initial (correct) location when anticipating Sam's return, even when they explicitly made the incorrect claim that he would go to the second location. They labeled this early manifestation of false belief understanding 'implicit', because the participating children were not explicitly aware of the knowledge conveyed in their correct eye gaze. Importantly, the researchers found no sign of implicit false belief understanding in 2-year-old children (cf. Perner & Clements 2000, Clements & Perner 2001, Ruffman et al. 2001, Garnham & Perner 2001).

However, recent studies on implicit false belief understanding challenge the assumption that infants under three do not understand false belief. Onishi & Baillargeon (2005), for example, used a 'violation-of-expectation' FBT to investigate whether children would look reliably longer when agents act in a manner that is inconsistent with their false beliefs. In the experiment, 15-month-old infants were familiarized with a protagonist hiding a toy in one of two locations. The protagonist left, and the toy was moved without his knowledge. Then the infants were shown scenes of the protagonist searching for the hidden toy, either where he falsely believes it to be, or where it was actually located. Onishi & Baillargeon (2005) found that infants looked significantly longer at those scenes in which the protagonist searched at the correct location despite their false belief about where the toy was hidden. Follow-up experiments have shown similar results in even younger infants of thirteen months (Surian et al. 2007).

These findings contradict the results of Clements & Perner (1994). According to Southgate et al. (2007), these different results are due to the fact that Clements & Perner (1994)'s experiment still included a verbal element: in order to maximize the frequency of anticipatory looking at one of the mouse holes, the investigator said aloud, 'I wonder where Sam is going to look?' before asking the question. Southgate et al. (2007) argue that this primed 2-year-old infants to look at the incorrect location. In their own study, they removed the verbal element from the design and used an eye-tracker to measure anticipatory looking in 25-month-olds. The infants observed how a protagonist witnesses a puppet bear that hides a ball in one of two boxes. Then the protagonist becomes distracted and turns away from the scene. Meanwhile, the bear removes the ball from its original hiding location. Southgate et al. (2007) found that most 25-month-olds correctly anticipated the protagonist's behavior and looked at the location where she falsely believed the ball to be hidden.

Other implicit FBTs indicate that infants do not only understand false beliefs about locations, but also about number, identity, and other properties (He & Baillargeon 2007, Scott & Baillargeon 2009, Scott et al., 2007, Song & Baillargeon 2008). These findings suggest that, contrary to what the results of explicit FBTs suggest, false belief understanding may already be present by the age of 13-18 months (see Baillargeon et al. 2010, Poulin-Dubois et al. 2009 for reviews) and perhaps even earlier (e.g., Kovács et al. 2010).

3. The Developmental Paradox of False Belief Understanding: Two Recent Solutions

Whether or not these findings should be interpreted under the heading of *false belief* understanding has been a weighty topic of discussion (Perner & Ruffman 2005; Ruffman & Perner 2005; Csibra & Southgate 2006; Sirois & Jackson 2007; Herschbach 2008). The issue is important because an early onset of false belief understanding during the first year suggests that the core principles of our Theory of Mind are largely part of our biological inheritance, whereas an onset at four years makes it more plausible that Theory of Mind is influenced by cultural processes and closely tied to language acquisition.

If false belief understanding already emerges during the second year of life, then the crucial question is why 3-year-old children consistently fail the explicit false belief task (Wellman et al. 2001), even when paradigms are used that reduce response selection and inhibition demands (e.g., Call & Tomasello 1999; Sodian et al. 2006). This is what we call the 'developmental paradox' of false belief understanding.

Most participants in the debate agree that there is no simple solution to the developmental paradox. Some of them have argued that we need two different systems to account for the intricacies of false belief development (e.g., Call & Tomasello 2008, Penn et al. 2008). Apperly & Butterfill (2009), for example, propose that the findings on implicit and explicit false belief understanding are best explained by respectively an early 'Minimal' Theory of Mind, which is cognitively efficient but limited and inflexible, and a flexible but cognitively very demanding full-blown Theory of Mind. Moreover, they offer an interpretation of infants' performance on the implicit FBT in terms of 'belief-like' states rather than full-blown beliefs. On Apperly & Butterfill's interpretation, infants are sensitive to the agent's belief only insofar as they *register*

the object. First, however, Apperly & Butterfill explain the simpler notion of *encountering*. Encountering is defined as 'a relation between an individual, an object and a location, such that the relation obtains when the object is in the individual's field' (p. 962). A field is defined, simply, as a certain region of space around the individual. Building on this, registering is defined as a slightly more complex psychological relation that obtains between an individual, an object and a location. An individual is said to register an object at a location when (a) she encounters the object at the location and (b) has not since encountered it somewhere else. A registering is off target when the object registered is not located where it is registered to be. The importance of the concept lies in the connections to actions: 'One can understand registration as an enabling condition for action, so that registering an object and location enables one to act on it later […] Further, registration also can be understood as determining which location an individual will direct their actions to when attempting to act on that object' (962). Tracking an agent's registration of something does not require sensitivity to her mental states as propositional attitudes – which only comes with a full-blown Theory of Mind.

Apperly & Butterfill's (2009) proposal seems to provide us with a promising way to explain the different forms of false belief understanding. However, it also raises an important question: how are the two Theory of Mind systems (functionally) related? One possibility is that there is no direct interaction between the early-developing and later-developing Theory of Mind systems. Apperly & Butterfill (2009) think this assumption is supported by the afore-mentioned findings of Clements & Perner (1994) and Southgate et al. (2007). For they show that infants correctly anticipate the action of another agent in terms of looking behavior, even when they respond incorrectly when asked to make an explicit prediction about the agent's future action. Apperly & Butterfill (2009) claim that these results are consistent with the possibility that an early-developing Theory of Mind system for tracking belief-like states is guiding children's eye movements and a later-developing Theory of Mind system underlies children's explicit judgments about beliefs.

The claim that there is no interaction between the two systems *at all* strikes us as implausible, however. For this would imply that our full-blown Theory of Mind system remains invisible during the first years of development and suddenly becomes fully operational at age four. At least we would expect to see *precursors* to such a Theory of Mind system that contribute to the system's functioning. For example, although Baron-Cohen's (1995) Theory of Mind Mechanism can be understood as a separate, late-developing mechanism, it depends on and receives input from the earlier developing Intentionality Detector, Eye-Direction Detector, and Shared Attention Mechanism.

Moreover, recent studies (e.g., Aschersleben et al. 2008, Sodian et al. forthcoming) indicate that the set of specific competencies required for the implicit FBT (e.g., looking-time patterns, eye-direction detecting, etc.) actually predicts performance on the explicit FBT. While this does not imply that there has to be interaction 'all the way down' or 'all the way up' between the two ToM systems, it seems to us that a complete theoretical account has to explain not only the dissociations but also the continuations in the development of false belief understanding. Therefore, we shall assume that the later-developing, cognitively more demanding ToM system depends, at least to some extent, on the early minimal ToM system for its operation (cf. Csibra & Gergely, 1998; Russell, 2007; Surian et al., 2007).

Baillargeon et al. (2010) have recently proposed a different solution to the developmental paradox. According to their proposal, infants come equipped with a psychological reasoning system that consists of two subsystems: sub-system 1 and sub-system 2. Sub-system 1 enables infants to attribute both motivational states and reality-congruent informational states to other agents, and is well in place by the end of the first year. Motivational states are defined as states that specify the agent's motivation in the scene and include goals and dispositions. Reality-congruent informational states, by contrast, specify what knowledge or accurate information the agent possesses about the scene. Sub-system 2 deals with

reality-incongruent informational states, i.e. false beliefs, and becomes operational in the second year of life.[1]

Baillargeon et al. (2010) argue that the developmental paradox of false belief understanding can be solved by means of a careful analysis of the task-requirements of explicit and implicit FBTs. They claim that, whereas the implicit FBT only involves (i) a process of false belief representation, the explicit FBT also requires (ii) a response selection process (when asked the test question, children must access their representation of the agent's false belief to select a response) and (iii) a response-inhibition process (when selecting a response, children must inhibit any prepotent tendency to answer the test question based on their own knowledge (see also Scott and Baillargeon 2009).

What is problematic about Baillargeon et al. (2010)'s solution is that it seems arbitrary to argue that the implicit FBT only involves false belief representation, whereas its explicit counterpart also require selection processing and response-inhibition. If we accept that the implicit FBT involves false belief representation, then it is not clear why it does not require selection processing and response-inhibition as well. For infants still have to select a false belief among other beliefs, and in order to do so they have to inhibit their default attribution of true beliefs.[2] And if this is correct, then it does not work to argue that the failure on the explicit FBT is due to the joint activation of false-belief-representation processes and response-selection processes, which 'overwhelms' the child's limited information-processing resources, and/or the fact that the neural connections between the brain regions that serve these two processes are still immature and inefficient in early childhood.

In other words, whereas Apperly & Butterfill (2009) ignore questions about the interaction between the two systems involved in false belief understanding, Baillargeon et al. (2010) seem to lack the conceptual resources to adequately explain the differences between implicit and explicit forms of false belief understanding.

4. Some Requirements for a Dual-system Account of False Belief Understanding

Most accounts of explicit false belief understanding assume that infants have a default tendency to attribute their own (true) beliefs to other agents (e.g. Leslie et al. 2004) or to respond on the basis of their own knowledge (Birch & Bloom 2007; Carlson & Moses 2001). In order to pass the explicit FBT, infants have to be capable of taking 'offline' (i.e. inhibiting) their own reality-congruent perspective.

According to Baillargeon et al. (2010), as we saw in the previous section, the ability for offline processing is precisely what constitutes the difference in task demands between the implicit and explicit FBT. However, this seems to ignore the fact that the implicit FBT involves offline processing as well, albeit it of a less demanding kind. Take the 'violation-of-expectation' study by Onishi & Baillargeon (2005), for example (see section 2). Although this experiment does not require infants to deal *explicitly* with differences in belief, it does require them to process differences between the visual information available to themselves and the visual information available to the other agent. This can only be accomplished offline, since the other's visual information is not directly available to the infant and needs to be represented by her. Therefore, already the implicit FBT involves a capacity for decoupling from one's own online processing of visual information and processing offline a representation of the visual information accessible to another agent. Yet, the role of decoupling and offline processing in this study is still limited. The infant largely relies on online visual information, and only has to process offline the other

[1] This proposal is an extension of Leslie's (1994) view that the ToMM (Theory of Mind Mechanism) consists of two subsystems: sub-system 1, which is available from 6 to 8 months of age and allows for the processing goal-directed actions, and sub-system 2, which is available from approximately 18 months of age and generates propositional attitude representations.

[2] Baillargeon et al. (2010) assume that infants by default attribute true beliefs to other agents.

agent's representation of the location of a *single* object. More difficult versions of the implicit FBT place stronger demands on offline processing. For instance, Song & Baillargeon (2008) conducted an experiment in which infants had to represent the visual representation of another agent with respect to *both* the location and the identity of *two* objects. Among implicit FBTs, we can thus distinguish between more or less demanding versions requiring more or less decoupling.

The above analysis shows that postulating a single sub-system, such as Baillargeon et al. (2010)'s sub-system 2, fails to explain the differences between implicit and explicit forms of false belief understanding.[3] This is because both implicit and explicit false belief understanding seem to require a three-step inhibition-selection-representation mechanism. Thus, an additional system is needed to explain why young infants already have an implicit understanding of false belief, but consistently fail tests assessing explicit false belief understanding. At the same time, contrary to what Apperly & Butterfill (2009) claim, this alternative dual-system account also has to say something about how the two systems interact throughout the development of false belief understanding.

In order to determine what is required of such a dual-system account it pays to consider an early dual-system view proposed by Alan Leslie. According to this view, explicit false belief understanding depends on (i) a Theory of Mind Mechanism, and (ii) a Selection Processing System. The Theory of Mind Mechanism contains the basic meta-representational concept of belief (as well as the concept of desire and pretense) and provides the infants with an early intentional insight into the behavior of others. Leslie assumes that infants have a default strategy of attributing their own true belief to others.[4] In order to pass the explicit FBT, they must learn to inhibit or override this default strategy, and select the content of the other agent's false belief. Leslie argues that both processes are handled by the Selection Processing System. The job of this system is essentially to 'inhibit competing possible contents for the belief' (Scholl & Leslie 1999, p. 147). Moreover, 'to infer the content of somebody's belief when that content is false, SP is required to select among the possible contents that ToMM makes available' (ibid.). However, since the Selection Processing system matures later than the Theory of Mind Mechanism (which is innate), the infant fails the false belief task until the Selection Processing system is in place. 'The developing performance on false-belief tasks, on this view, reflects not a developing ToMM, but a developing SP' (ibid., p.149).

Is such a ToMM/SP dual-system view able to account for implicit forms of false belief understanding? Interestingly, Leslie (2005) does mention the findings by Onishi & Baillargeon (2005). He claims that they 'underline the early role of ToMM as a core mechanism of attention, identifying learning opportunities as expectations are violated and directing attention to relevant sources of information' (p.532). According to Leslie, the experiment by Onishi & Baillargeon (2005) and others reflect a very early understanding of false belief, which means that SP is already operational in infants of 15 months. However, if this is true, then the whole idea that infants use a default attribution strategy before they pass the explicit FBT becomes problematic. For it means that SP initially provides infants with an early implicit understanding of false belief, then, at a later point of development, exchanges this for a default attribution strategy (despite the fact that the infant already has an understanding of the difference between true and false belief), only to reinstall an explicit form of false belief understanding around the age of four. Nevertheless, Leslie (2005) does seem to follow this line of thought, for he speculates that 'sometime between 15 and 30 months, SP learns to make the true-belief attribution the default' (p.532).

---

[3] Although Baillargeon et al. (2010) postulate two sub-systems, only *one* of them (sub-system 2) is used to explain the difference between implicit and explicit false belief understanding.

[4] Leslie (2000) explains the importance of default belief attribution as follows: 'it is useful to understand why belief attribution has a default bias. If desires set an agent's goals, beliefs inform the agent about the state of the world. A belief that misinforms an agent is a useless, even a dangerous thing: beliefs ought to be true. Therefore, the optimal default strategy for the belief attributer is to assume that an agent's beliefs are true.' (p.1242)

This is puzzling. Why would SP make true-belief attribution the default when infants already have an understanding of false belief, especially given the fact that it has to be overridden again at a later stage of development? The problem here seems to be that Leslie tries to explain the development of false belief understanding completely in terms of the development of a single Selection Processing System, much like Baillargeon et al. (2010) try to explain it completely in terms of the inhibition-selection-representation processes enabled by sub-system 2.

We think that what is needed instead is a more dynamic view of the *concepts* our Theory of Mind Mechanism makes available and how they unfold throughout ontogeny (in close interaction with an inhibition-selection-representation mechanism). This is not per se incompatible with the idea of innate mental state concepts. Spaulding (2010), for example, maintains that innate mental state concepts are not necessarily 'robust' (p.127). She contrasts robust mental states, typified by propositional attitudes such as belief, with 'sub-doxastic' mental states (p.123) that do not possess truth-evaluable, propositional content.[5]

This suggestion is very much in line with recent proposals that emphasize the need to get away from the standard folk psychological concepts of belief and desire when it comes to explaining early socio-cognitive capacities such as those recruited for implicit false belief understanding. These proposals range from the postulation of a Naïve, Weak, or Minimal Theory of Mind, i.e., one lacking paradigm folk psychological concepts (Bogdan 2009, Tomasello et al. 2003, Apperly & Butterfill 2009), Perceptual Mindreading (Bermúdez 2009), or an Early Mindreading System (Nichols & Stich 2003).

We already mentioned Apperly & Butterfill's (2009) proposal of an early-developing Minimal Theory of Mind, which enables infants to track belief-like states and guides their eye movements. In a similar vein Bogdan (2009) postulates the existence of a Naïve Theory of Mind which has 'the primary function of registering and representing another mind's *relations* to the world' (p.63). Naïve Theory of Mind is best understood as an assembled cluster of abilities that enables the grasping and representing of the mental states of others, specifically - gazing, seeing, and emoting. Bogdan (2009) proposes that infants slowly move from initially noticing such things as another's direction of gaze, bodily posture or movement in purely *behavioral* ways to being able to track, register, or represent the target of the *purposed aboutness* of another's attending (2009, p.71).

Like Apperly & Butterfill's (2009) 'encountering' and 'registering', however, it is not clear whether this notion is sufficiently fine-grained to properly explain the different abilities involved in various forms of false belief understanding. Bogdan (2009) explains 'purposed aboutness' as a kind of goal-directed intentionality that is recognizably expressed in the way that organisms respond to aspects of their immediate environment - showing (1) *relatedness* to a target; (2) the *direction* of this relatedness; and (3) the *target* itself. But such an explanation is too general and fails to capture important differences between the abilities required for false belief understanding. Consider Southgate et al. (2007), for example, who employed an eye-tracker to measure anticipatory looking in 25-month-olds. In the experiment, infants observed how a protagonist witnesses a puppet bear that hides a ball in one of two boxes. Then the protagonist becomes distracted and turns away from the scene. Meanwhile, the bear removes the ball from its original hiding location. Southgate et al. (2007) found that most 25-month-olds correctly anticipated the protagonist's behavior and looked at the location where she falsely believed the ball to be hidden. Like the 'violation-of-expectation' study by Onishi & Baillargeon (2005), this experiment requires infants to process differences between the visual information available to themselves and the visual information available to the protagonist. Since the latter is not directly available, it has to be represented by the infant. In this scenario, an understanding of the agent's *purposed aboutness* comes down to being able to anticipate her behavior on the basis of her *perception* of a given object in the previous scene.

Now take a different experiment by Luo & Baillargeon (2005). This study showed that infants of 5 months old, after watching an agent repeatedly reach for object A as opposed to object B, registered the

[5] See Stich (1978) for the origins of this distinction.

'purposed aboutness' (i.e. preference) of this agent for object A over object B. When the objects' positions were reversed, infants expected the agent to reach for object A in its new position, and they looked reliably longer if the agent reached for object B instead. What is important is that in this experiment, unlike the one by Southgate et al. (2007), infants did not need to process *perceptual* incongruencies between their visual perspective and that of the other agent. Both infant and agent perceived the same object, and the infant does not need to represent the reality-incongruent informational state of another agent. In this experiment, understanding *purposed aboutness* means being able to anticipate the behavior of another agent on the basis of her *movements* towards the object.

Both behavior-anticipation strategies depend on different abilities, and although both emerge in the first year of life, they have a different developmental onset. Visual habituation studies indicate that infants are already capable of understanding another agent's goal-directed movement towards an object from 5 months onwards (e.g., Biro & Leslie, 2007; Gergely & Csibra, 2003; Woodward, 1998, 2005). Importantly, they do not selectively attend to goals for events involving inanimate objects, such as rods or claws (Woodward, 1998), or for events in which the agent's hand is disguised by a metallic glove (Guajardo & Woodward, 2004).[6] Understanding another's agent goal-directed perception emerges later in development. In an experiment by Woodward (2003), for example, 7- and 9-month-old infants followed an agent's gaze and when they saw the agent look at and grasp a toy, they not only looked at that toy, but also selectively registered the directedness of the agent towards the toy. However, when the infants only saw the agent look at the object but not touch it, they failed to register this directedness. Infants of 12 months, in contrast, were capable of registering the agent's directedness towards the toy solely on the basis of her gaze. Other studies also indicate that it is only towards the end of the first year that infants become capable of registering more abstract (or 'distal') goal-directed behaviors such as looking and pointing (Phillips et al. 2002, Sodian & Thoermer 2004, Woodward 2003, 2005, Woodward & Guajardo 2002).

This shows that the concept of purposed aboutness is not fine-grained enough to capture the different abilities recruited in the various implicit FBTs. The same seems to hold for the notions of 'encountering' and 'registering' proposed by Apperly & Butterfill (2009). Of course, this is not a principled objection against these positions. Rather, it should be seen as an incentive to develop a more sophisticated conceptual vocabulary to do justice to the empirical findings on false belief understanding.

Ideally, this conceptual vocabulary should be sufficiently flexible to account for two important developmental interactions, namely, (i) between the infant's own action towards an object and her subsequent perception of another agent's goal-directed behavior towards the object, and (ii) between the infant's perception of another agent's goal-directed behavior towards an object and her own perception of the object.

With respect to (i): a study by Sommerville et al. (2005) demonstrated that even 3-month-olds focus on the relation between an actor and her goal if they reached (and not just watched) for a toy before observing another agent grasping it. The more they themselves were engaged in object-directed contact with the toys, the more sensitive they were to the agent goal-directed behavior. More recently, Sommerville et al. (2008) also found that 10-month-old infants who received active training in pulling a cane to retrieve a toy subsequently registered another person's cane-pulling actions as goal-directed behavior, while infants who underwent observational training were unable to do this.

[6] Early visual habituation experiments showed that infants did not register the goals of inanimate objects (Woodward 1998, Guajardo & Woodward, 2004). Recent findings, however, suggest that infants do sometimes perceive inanimate entities as goal-directed agents (Biro & Leslie, 2007; Csibra, 2008; Johnson et al. 2001; Kuhlmeier et al. 2003; Mahajan & Woodward 2009; Luo & Baillargeon, 2005; Shimizu & Johnson, 2004). This seems to depend on the availability of additional cues (e.g. self-propelled motion) indicating the animacy of the agent. The current debate is mainly concerned with the range of cues that might contribute to the infants' goal understanding (cf. Biro & Leslie, 2007).

With respect to (ii): it has been shown that behavioral cues (Biro et al. 2007, Biro & Leslie 2007), infant-directed talk and eye-contact induce gaze-following towards specific objects (Senju & Csibra 2008), and the use of specific linguistic labels increases the salience of objects over others (Xu 2002, Xu et al. 2004). Other experiments suggest that reaching behavior promotes the infant's perception of the spatiotemporal properties of the object of interest, whereas pointing behavior promotes the surface properties of the object (e.g., Csibra & Gergely 2006).

This raises important questions about how the infant's own action informs her perception of goal-directed behavior and vice versa. Some have proposed that the actions of infants and their perceptions of the actions of others are by a cross-modal system that translates action and perception into a unified 'language'. Georgieff & Jeannerod (1998) proposed the term 'shared representation' in order to articulate the idea that action and perception might essentially share the same representational space. This possibility is compatible with the findings of mirror neurons – neurons that fire during both action production and action perception (e.g., Rizzolatti & Craighero 2004, Rizzolatti et al. 2006). It has been argued that these mirrors show a 'human bias', in the sense that they resonate stronger with perceived actions of human versus non-human agents (Press et al. 2007, Tsai et al. 2008). This link between perception and action has also been found in early infancy (Kanakogi & Itakura 2010), and research on newborn imitation has been cited as evidence for an inherited mirror neuron system that underlies imitative behavior in human infants (e.g., Iacoboni et al.1999; Decety et al. 2002; Grezes et al. 2003; Iacoboni 2005; Iacoboni & Depretto 2006). However, there are many open questions about the existence of such a system in infants and its role in infant development (Gerson & Woodward in press; Meltzoff 2006). Moreover, there are also methodological doubts (Hickok 2009) and experiments that fail to report mirror neuron activity (Lingnau et al. 2009).

Therefore, it should be emphasized that mirror neurons are just one way of getting at the more general idea that action production and action perception can be understood in terms of shared representations. There are other grounds for supporting this idea as well, e.g. on the basis of proposals about action coding (e.g., Elsner & Hommel 2001, Prinz 2002) or findings from developmental studies (Meltzoff 2004, 2006; Meltzoff & Moore 1977, 1994; Meltzoff & Brooks 2001).

5. False Belief Understanding as Progressive Decoupling

What is attractive about shared representations is that they explain how perception and action are dynamically co-constituted in what Gallese (2001) calls the primordial 'we space'. However, this also gives rise to an important question about the registration of *agency*. How are the mechanisms that facilitate shared representations able to differentiate situations in which infants observe the goal-directed behavior of another agent from those in which they perform the same action themselves – such as those in the studies by Sommerville et al. (2005, 2008)? This is a serious problem for those who appeal to mirror neurons in their explanation of action understanding. Since both conditions activate the same cortical 'mirror' sectors, an additional mechanism is needed to determine whether the infants performs or observes the action. More in general, it can be seen a problem for proponents of a version of Simulation Theory. The question is, as Gordon (1986) puts it, how infants manage to make 'adjustments for the relevant differences' while avoiding 'total projection'.

Although we think this is indeed an important problem, it should not be overstated. To start with, researchers have proposed various solutions to address this issue. According to one proposal, shared representations are neither first- or third-person. The infant's observation of goal-direct behavior triggers the activation of so-called 'naked representations'. The idea is that mirror neurons encode the sensorimotor and perceptual properties of goal-directed behavior (either perceived or produced) in a shared representational format, but do not register the agent behind the action (deVignemont 2004; Jeannerod & Pacherie 2004; Gallese 2005; Hurley 2008). This is done in a second step by an additional

mechanism. Georgieff & Jeannerod (1998), for example, have argued that this process might be taken care of by a 'Who' mechanism. Evidence for this mechanism comes from experiments showing a differential activation in the posterior insula when the subject took the role of agent, and in the right inferior parietal cortex when it took the role of observer (Farrer et al. 2003; Farrer & Frith 2002; Ruby & Decety 2001).

We are not committed to one of these more specific proposals. We merely mention them in order to make clear that the question of self-other differentiation is mainly problematic for those aiming to explain action understanding solely in terms of mirror neuron processes – like Iacoboni or Gallese. But this is certainly not a position that we wish to defend. In fact, we think the more interesting question for those who subscribe to shared representations is precisely what additional mechanisms are required to explain the differentiation between self and other. This is not only necessary to account for instances of imitative behavior, which is central to most experiments on mirror neurons (cf. Iacoboni 2005), but also action emulation (i.e. achieving the same goal by different means), and cases in which agents actually do the *opposite* of the observed behavior.

Given our current focus on the development of false belief understanding, we are primarily interested in those situations in which infants are sensitive to the reality-incongruent informational states of other agents (i.e. their false beliefs). Instead of opting for 'naked representations', however, we shall follow other researchers (e.g., Leslie et al. 2004, Birch & Bloom 2007; Carlson & Moses 2001) in assuming that infants simply have a tendency to attribute their own representational states to other agents. Thus, we agree with Goldman (2006) that infants' default procedure is to project their own basic concepts onto others.

In section 3, we criticized Baillargeon et al. (2010) for explaining the difference between implicit and explicit false belief understanding in terms of a single inhibition-selection-representation mechanism. More in particular, we suggested that implicit false belief understanding may involve selection processing and response-inhibition as well. This is because infants still have to select a reality-incongruent informational state, and in order to do so they have to inhibit their default tendency to attribute a reality-congruent informational state. In what follows we will explain this proposal in more detail.

Our starting point is Baillargeon et al. (2010)'s notion of a reality-congruent informational state, i.e. an informational state represented on the basis of the infant's own perspective on the scene. We think this notion has potential when it comes to explaining the development of false belief understanding, but it needs to be specified in more detail – in particular in relation to the infant's mode of perspective taking. Furthermore, in line with other dual-system accounts, we postulate two systems in order to account for the development of false belief understanding: (a) an inhibition-selection-representation system, or 'ISRS', and (b) a default attribution mechanism that is responsible for the default attribution of reality-congruent informational states, or 'DAM'. Central to our proposal is the idea that ISRS, like Leslie's Selection Processing system, functions as a 'de-coupling' mechanism. In our model, however, ISRS is not responsible for decoupling *false beliefs*. Rather, it decouples different kinds of reality-congruent informational states.

We propose that there are three ways in which ISRS facilitates the decoupling of reality-congruent informational states, thus allowing infants to understand the reality-incongruent information states of other agents instead. In the first place, ISRS can decouple a reality-congruent informational state by inhibiting its *sensorimotor* properties. The basic idea is that infants by default register reality-congruent informational states on the basis of their own sensorimotor perspective, i.e. their own (intended) movements towards the object. It has been hypothesized that this underlies infants' ability to anticipate the consequences of their own behavior: the brain generates motor-simulations of intended movements by sending efference copies through a forward control mechanism in order to compare them with an ongoing movement to predict its success (e.g., Frith et al. 2000, Blakemore et al. 2002, Blakemore & Frith 2003). In order to understand the goal-directed behavior of *another* agent (i.e. her movement towards the object), three ISRS sub-processes are required: (i) a response inhibition process (infants have to inhibit

their own sensorimotor perspective, (ii) a response selection process (infants have to select the sensorimotor perspective of the other agent), and (iii) a representation process (infants have to represent a reality-incongruent informational state that is informed by the other agent's sensorimotor perspective). This first-order mode of decoupling reality-congruent informational states allows us to explain what happens in experiments like the one by Luo & Baillargeon (2005), where infants are able to anticipate the behavior of another agent on the basis of his or her movement towards the object.

One might wonder how the infant is able to select and represent the sensorimotor perspective of another agent. This is precisely where we think the notion of shared representation could play an important role: the infant's perception of the agent's goal-directed behavior in the familiarization trials leads to 'shared resonance', and provides the infant with the sensorimotor information required to execute the observed action. During the test trial, the infant has to select and represent a reality-incongruent informational state on the basis of this sensorimotor information.

Secondly, ISRS can also decouple a reality-congruent informational state by inhibiting the infant's *perceptual* perspective. Our assumption is here that infants by default register reality-congruent informational states on the basis of their *own perception* of the object. In order to anticipate another agents' behavior on the basis of their visual perspective (i.e., what they can or cannot see), the infant has to take offline its own perceptual perspective. This second-order mode of decoupling again involves three ISRS processes: inhibition, selection and representation.

Now we are in the position to explain what happens in implicit false belief experiments, which show that infants are able to anticipate the behavior of another agent on the basis of his or her perception of the object. Consider again the Southgate et al. (2007) experiment, which showed that 25-month-olds correctly anticipated the behavior of a protagonist with a false belief about the location of a ball. According to our model, in order to anticipate the behavior of the agent on the basis of her visual perspective (i.e., what the agent saw in the previous scene) the infant has to inhibit its own perceptual perspective, and represent the perceptual perspective of the other agent instead.

Finally, the ISRS allows infants to decouple a reality-congruent informational state by inhibiting its *cognitive* perspective. Before they acquire linguistic competence, infants are already capable of representing certain proximal goal-directed actions as informational states with a means-end structure. This provides them with a basic understanding of the 'in-order-to' relations that are characteristic for goal-directed behavior, e.g., the agent reaches out in order to grasp the object. What is important about linguistic symbols is that they allow infants to (re)configure informational states in much more complex 'in-order-to' relations, thereby enabling an increasingly sophisticated typing of the distal goal-directed actions of other agents. Our use of the terms 'perceptual perspective' and 'cognitive perspective' is meant to illustrate precisely this difference between perceiving and/or representing proximal versus more distal goal-directed actions.

This last ability is an important requirement for the explicit FBT. For this task requires infants to deal with rather abstract experimental scenarios, i.e. stories or pictures instead of the interacting real-life agents and objects that feature in the implicit FBT. On our view, infants by default register what happens in these scenarios on the basis of their own cognitive perspective. Now in order to verbally predict another agents' behavior on the basis of *their* cognitive perspective, the infant has to take offline its own reality-congruent perspective. This third-order mode of decoupling again involves three ISRS processes: inhibition, selection and representation.

This allows us to explain what happens in the explicit FBT, but it does not yet explain the developmental paradox of false belief understanding, i.e. why infants have more difficulty with the explicit false belief task vis-à-vis its implicit counterpart. Although we cannot offer a concrete solution to this problem here, we do think there a number of options that should be further investigated. In the first place, the explicit false belief task might simply be more difficult because it requires a much stronger form of decoupling. Evidence suggests that explicit false belief understanding indeed places increasing demands on executive functioning. For example, several studies have found robust correlations between explicit

FBT performance and response inhibition (e.g., Perner & Lang 1999, Cole & Mitchell 2000, Carlson & Moses 2001) and working memory (Carlson et al. 2002, Hala et al. 2003, Perner et al. 2002).

However, it seems unlikely that this is the whole story. What studies such as the one by Southgate et al. (2007) show is that especially *verbal interaction* between infant and experimenter crucially contributes to the difficulty of the explicit FBT. Many experiments have found strong correlations between linguistic competence and explicit FBT performance (Dunn et al. 1991, Astington & Jenkins 1999, Gale et al. 1996, De Villiers & De Villiers 2000, Watson et al. 2002, Farrar & Maag 2002). There are several hypotheses about why children have more difficulty with FBTs involving linguistic interaction. Some researchers propose that children need to master its semantics (Moore et al. 1990), whereas others argue that what is required is getting a handle on its syntactic structure (e.g., Hale & Tager-Flusberg 2003, Lohmann & Tomasello 2003). We are not committed to one of these hypotheses in particular, but we like to point out that they are not incompatible with the previous point about the stronger decoupling requirement. It is very well possible that the explicit FBT requires a stronger form of decoupling precisely *because* it involves language. An intriguing possibility is that infants fail the explicit false belief task because there is something that *interferes* with the decoupling process, namely, their verbal interaction with the experimenter. This requires further investigation, however.

6. Closing Comments and Further Research

One of the strengths of the ISRS-DAM model presented in the previous section is that it does justice to the developmental continuity of false belief understanding, and also gives a clear explanation of how the interaction between the two sub-systems provides infants, at each stage of development, with more advanced capacities to understand other agents. In this way, the model is able to avoid two serious problems for dual-system accounts of false belief understanding, namely, (i) how these systems interact and (ii) what this implies for their ontogenetic development (see section 3).

Of course there are several remaining issues that still need to be addressed. One important question has to do with the role of shared representations in the infant's ability to represent and understand another agent's visual and cognitive perspective. A second question concerns the role of inhibition in decoupling reality-congruent informational states. We have argued that ISRS enables infants to inhibit their own reality-congruent perspective in order to represent the reality-incongruent perspective of another agent. But is this decoupling required for all instances in which infants register reality-incongruent informational states? It seems plausible to assume that the amount of inhibition required diminishes as a result of learning, and infants become increasingly skilled at switching between congruent and incongruent perspectives in the long run.

The ISRS-DAM model also offers new directions for future research, for instance with respect to its neurobiological implementation. An interesting idea is to understand DAM, which underlies the default tendency to attribute reality-congruent informational states, as a *simulation* mechanism. Consequently, it could be investigated to which extent DAM recruits the brain areas traditionally associated with the mirror neuron system: the superior temporal sulcus, the inferior frontal cortex, and the rostral part of the inferior parietal lobe (Iacoboni et al. 1999, 2005, Iacoboni & Dapretto 2006, Koski et al. 2002, 2003, Decety et al. 2002, Chaminade et al. 2005). The ISRS processes that enable the representation of reality-incongruent informational states, by contrast, might be facilitated by a 'mentalizing' network, consisting of the anterior cingulated cortex, the temporoparietal junction, the superior temporal sulcus and the temporal poles (Frith & Frith 2003, Amodio & Frith 2006). Further research has to show whether it is possible to establish such a link between DAM and the mirror neuron system, and ISRS and the mentalizing network.

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7,* 268–277.

Apperly, I., & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116,* 953–970

Aschersleben, G., Hofer, T., & Jovanovic, B. (2008). The link between infant attention to goal-directed action and later theory of mind abilities. *Developmental Science, 11*, 862–868.

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relationship between language and theory-of-mind development. *Developmental Psychology*, *35,* 1311–1320.

Baillargeon, R., Scott, R.M and Zijing, H. (2010). False-belief understanding in infants. *Trends in Cognitive Science 14, 3, 110-118*

Baron-Cohen, S. (1995). Mindblindness: an essay on autism and theory of mind. MIT Press/Bradford Books.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21,* 37–46.

Bermúdez, J. (2009). Mindreading in the Animal Kingdom. In: Lurz, R. (ed.) The Philosophy of Animal Minds. Cambridge: Cambridge University Press.

Birch, S.A.J. & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18(5)*, 382-386.

Biro, S., Csibra, G., & Gergely, G. (2007). The role of behavioral cues in understanding goal-directed actions in infancy. *Progress in Brain Research, 164*, 303–322.

Biro, S., & Leslie, A. M. (2007). Infants' perception of goal-directed actions. Development through cues-based bootstrapping. *Developmental Science, 10(3),* 379–398.

Blakemore, S., Wolpert D., & Frith, C. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences, 6,* 237–42.

Blakemore, S. J., & Frith, C. D. (2003). Self-awareness and action. *Current Opinion in Neurobiology, 13(2),* 219–224.

Bloom, P., & German, T. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77,* B25–B31.

Bogdan, R. (2009). *Predicative Minds: The Social Ontogeny of Propositional Thinking*. Cambrdige, MA: MIT Press.

Braver, T.S., Cohen, J.D., & Barch, D.M. (2002). The role of the prefrontal cortex in normal and disordered cognitive control: a cognitive neuroscience perspective*. In: Stuss D.T., Knight R.T., (eds). Principles of frontal lobe function. Oxford: Oxford University Press,* 428–448.

Braver, T.S., & Cohen, J.D. (2001). Working memory, cognitive control, and the prefrontal cortex: computational and empirical studies. *Cognitive Processes, 2,* 25–55.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Science, 12,* 187–192.

Call, J., & Tomasello, M. (1999). A nonverbal theory of mind test. The performance of children and apes. *Child Development, 70,* 381–395.

Carlson, S., & Moses, L. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72,* 1032–1053.

Carlson, S., Moses, L., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development, 11,* 73–92.

Chaminade, T., Schwarzlose, R. F., Baker, C. I., & Kanwisher, N. (2005). Separate face and body selectivity on the fusiform gyrus. *Journal of Neuroscience, 25* (47), 11055–11059.

Clearfield, M. W., Diedrich, F. J., Smith, L. B., & Thelen, E. (2006). Young infants reach correctly in A not-B tasks: On the development of stability and perseveration. *Infant Behavior & Development*, *29,* 435–444.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9,* 377–397.

Clements, W. A. & Perner, J. (2001). When actions really do speak louder than words? but only implicitly: Young children's understanding of false belief in action. *British Journal of Developmental Psychology*, *19*, 413–432.

Cole, K., & Mitchell, P. (2000). Siblings in the development of executive control and a theory-of-mind. *British Journal of Developmental Psychology, 18,* 279–295.
Csibra, G., & Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. Developmental Science, 1(2), 255–259.
Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. *In Y. Munakata & M. H. Johnson (eds.), Processes of change in brain and cognitive development. Attention and performance, xxi*. Oxford, UK: Oxford University Press.
Csibra, G., & Southgate, V. (2006). Evidence for infants' understanding of false beliefs should not be dismissed. Response to Ruffman and Perner. *Trends of Cognitive Science 10,* 4-5.
Daprati, E., Frank, N., Georgieff, N., Proust, J., Pacherie, E., & Dalery Jeannerod, M. (1997). Looking for the agent: an investigation into consciousness of action and self consciousness in schizophrenic patients. *Cognition, 65,* 71–86.
Decety, J., Chaminade, T., Grezes, J., & Meltzoff, A. (2002). A PET exploration of the neural mechanisms involved in reciprocal imitation. *NeuroImage, 15*, 265–272.
De Villiers, J., & De Villiers, P. (2000). Linguistic determinism and the understanding of false beliefs. In Mitchell, P., & Riggs, K. J. (Eds.), *Children's reasoning and the mind* (pp. 191–228). Hove, England: Psychology Press.
De Vignemont F. (2004). The co-consciousness hypothesis. *Phenomenology and the Cognitive Sciences 3,* 97-114.
Dunn, J., Brown, J., Slomkowski, C., Tesla, C., & Younblade, L. (1991). Young children's understanding of other people's feelings and beliefs: Individual differences and their antecendents. *Child development*, *62,* 1352–1366.
Elsner, B., & Hommel, B. (2001). Effect anticipation and action control. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 229–240.
Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs. another person as being the cause of an action: the neural correlates of the experience of agency. *Neuroimage, 15,* 596–603.
Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: a positron emission tomography study. *Neuroimage, 18,* 324–333.
Farrar, M. J., & Maag, L. (202). Early language development and the emergence of a theory of mind. *First language*, *22,* 197–213.
Flavell, J. H. (2004). Theory-of-mind development: Retrospect and prospect. *Merrill-Palmer Quarterly Journal of Developmental Psychology, 50,* 274–290.
Frith, C., Blakemore, S., & Wolpert, D. (2000). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Brain Research Review, 31,* 357–363.
Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalising. *Philosophical Transactions, Series B, 358 (Special issue on Mechanisms of social interaction (Frith, C. D., & Wolpert, D., Eds.)*, 459–473.
Garnham, W. A., & Perner, J. (2001). When actions really do speak louder than words? but only implicitly: Young children's understanding of false belief in action. British Journal of. Developmental Psychology, 19, 413–432.
Gale, E., deVilliers, P., deVilliers, J., & Pyers, J. (1996). Language and theory of mind in oral deaf children. Paper presented at the *Boston University Conference on Language Development*, Boston, MA.
Gallese, V. (2001). The 'shared manifold' hypothesis - From mirror neurons to empathy. *Journal of Consciousness Studies, 8,* 33-50.
Gallese, V. (2005). 'Being like me': Self-other identity, mirror neurons and empathy*. In Hurley, S. &Chater, N. (Eds.), Perspectives on imitation: From cognitive neuroscience to social science: Mechanisms of imitation and imitation in animals (101–118).* Cambridge, MA: MIT Press
Georgieff, N., & Jeannerod, M. (1998). Beyond consciousnes of external reality: A "Who" system for consciousness and action and self-consciousness. *Consciousness & Cognition, 7,* 465–487.
Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences, 7,* 287–292.
Gerson, S., & Woodward, A.L. (in press*). Building intentional action knowledge with one's hands.* In S.P. Johnson (Ed.), Neo-constructivism. New York: Oxford University Press.
Goldman, A. (2006). Simulating minds: The philosophy, psychology and neuroscience of mindreading. New York: Oxford University Press.

Gordon, R.M. (1986). Folk psychology as simulation. *Mind & Language, 1,* 158-171.
Grezes, J., Armony, J., Rowe, J. & Passingham, R. (2003). Activations related to "mirror" and "canonical" neurones in the human brain: an fMRI study. *Neuroimage 18*, 928-937.
Guajardo, J. J., & Woodward, A. L. (2004). Is agency skin-deep? Surface attributes influence infants' sensitivity to goal-directed action. *Infancy, 6,* 361–384.
Hala, S., Hug, S., & Henderson, A. (2003). Executive functioning and false-belief understanding in preschool children: Two tasks are harder than one. *J. Cogn. Dev.*, *4,* 275–298.
Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: a training study. *Developmental science*, *6(3),* 346–359.
He, Z., & Baillargeon, R. (2007). Understanding of false belief in 11-month-old infants. *Paper presented at the biennial meeting of the Society for Research in Child Development, Boston, MA.*
Herschbach, M. (2008). Folk psychological and phenomenological accounts of social perception. *Philosophical Explorations*, *11(3),* 223-235.
Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans 42–58.
Hurley, S. (2008). The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behavioral Brain Sciences, 31*, 1–58.
Iacoboni, M. (2005). Understanding others: Imitation, language, empathy. In Hurley, S., & Chater, N. (Eds.), *Perspectives on imitation: From cognitive neuroscience to social science: Mechanisms of imitation and imitation in animals* (pp. 77–99). Cambridge, MA: MIT Press.
Iacoboni, M., Woods, R. P., Brass, M., Hekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science, 286*, 2526–2528.
Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience,* 7, 942–951.
Jeannerod, M., & Pacherie, E. (2004). Agency, Simulation and Self-Identification. *Mind & Language, 19,* 113-146.
Kanakogi, Y., & Itakura, S. (2010). The link between perception and action in early infancy: From the viewpoint of the direct-matching hypothesis. *Japanese Psychological Research, 52(2),* 121–131.
Kochanska, G., Coy, K. C., & Murray, K. T. (2001). The development of self-regulation in the first four years of life. *Child Development, 72,* 1091–1111
Kochanska, G., Padavich, D. L., & Koenig, A. L. (1996). Children's narratives about hypothetical moral dilemmas and objective measures of their conscience: Mutual relations and socialization antecedents. *Child Development, 67,* 1420–1436.
Koski, L., Wohlschlager, A., Bekkering, H., Woods, R. P., Dubeau, M. C., Mazziotta, J. C., & Iacoboni, M. (2002). Modulation of motor and premotor activity during imitation of target-directed actions. *Cerebral Cortex, 12,* 847–855.
Koski, L., Iacoboni, M., Dubeau, M. C., Woods, R. P., & Mazziotta, J. C., (2003). Modulation of cortical activity during different imitative behaviors. *Journal of Neurophysiology, 89*, 460–471.
Kovács, A., Teglas, E., & Endress, A (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science, 330,* 1830–1834.
Leslie, A. M. (2000). How to acquire a 'representational theory of mind'. In Sperber, D. (Ed.), *Metarepresentations: A Multidisciplinary perspective* (pp. 197–223). New York: Oxford University Press.
Leslie, A.M., Friedman, O. & German, T.P. (2004) Core mechanisms in "theory of mind". *Trends in Cognitive Sciences 8(12),* 528-533.
Lingnau A., Gesierich B. & Caramazza A. (2009). Asymmetric fMRI adaptation reveals no evidence for mirror neurons in humans. *PNAS 106,* 9925-30.
Lohmann, H. & Tomasello, M. (2003). The role of language in the development in false belief understanding: A training study. *Child Development*, *74,* 1130–1144.
Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychol. Sci., 16,* 601–608.
Luria, A.R. The frontal lobes and the regulation of behavior. In: K.H. Pribram and A.R. Luria, Editors, *Psychophysiology of the frontal lobes*, Academic Press, New York (1973), pp. 3–26
McKay, K.E., J.M. Halperin, S.T. Schwartz, and V. Sharma. 1994. Developmental analysis of three aspects of information processing: Sustained attention, selective attention, and response organization. *Developmental Neuropsychology 10,* 121-132.

Meltzoff, A. N., & Moore, M. K. (1977). Immitation of Facial and Manual Gestures by Human Neonates. *Science, 198,* 75–78.
Meltzoff, A. N., & Moore, M. K. (1994). Imitation, memory, and the representation of persons. *Infant Behavior and Development*, *17,* 83–99.
Meltzoff, A. N., & Brooks, R. (2001). 'Like me' as a building block for understanding other minds: Bodily acts, attention, and intention. In Malle, B. F., Moses, L. J., & Baldwin, D. A. (Eds.), *Intentions and intentionality: foundations of social cognition* (pp. 171–191). Cambridge, MA: MIT Press.
Meltzoff, A. N. (2004). Imitation and other minds: The "like me" hypothesis. In Hurley, S., & Chater, N. (Eds.), Perspectives on Imitation: From Neuroscience to Social Science Vol. II (pp. 55–77). Cambridge, MA: MIT Press.
Meltzoff, A.N. (2006). The "like me" framework for recognizing and becoming an intentional agent. *Acta Psychologica, 124*, 26–43.
Moore, C., Pure, K., & Furrow, D. (1990). Children's understanding of the modal expression of speakers certainty and uncertainty and its relation to the development of representational theory of mind. *Child Development*, *61,* 722–730.
Moses, L. J., & Flavell, J. H. (1990). Inferring false beliefs from actions and reactions. *Child Development, 61,* 929–945.
Moses, L. J., Carlson, S.M., & Sabbagh, M.A. (2005). On the specificity of the relation between executive function and children's theory of mind. In Schneider, W., Schumann-Hengsteler, R., & Sodian, B. (Eds.), *Young children's cognitive development: Interrelations among executive working memory, verbal ability and theory of mind* (pp. 131–146). Hillsdale, NJ: Erlbaum.
Nichols, S., & Stich, S. (2003). *Mindreading. An integrated Account of Pretence, Self-Awareness, and Understanding of Other minds.* Oxford: Clarendon Press.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science, 308(8),* 255–258.
Passler, M.A., Isaac, W. & Hynd, G.W. (1985). Neuropsychological development of behavior attributed to frontal lobe functioning in children. *Developmental Neuropsychology,* 7(2), 131-149
Penn, D. C., Holyoak, K. J., Povinelli, D. J. (2008). Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav Brain Sci., 31,* 109–178.
Perner J., Leekam S. & Wimmer H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology, 5,* 125–137.
Perner, J. 1991. *Understanding the representational mind.* Cambridge, MA: Bradford Books/MIT-Press.
Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences, 3(9),* 337–344.
Perner, J. & Ruffman, T. (2005) Infants' insight into the mind: How deep? *Science 308*, 214–216
Perner, J., & Clements, W. A. (2000). From an Implicit to an Explicit "Theory of Mind". In Rossetti, Y., & Revonsuo, A. (Eds.), Beyond Dissociation: Interaction between dissociated implicit and explicit processing (pp. 273–294). Amsterdam/Philadelphia: John Benjamins Publishing Co.
Perner, J., Lang, B., & Kloo, D. (2002). Theory of Mind and Self Control: More than a common problem of inhibition. *Child Development, 73,* 752–767.
Perner, J. Leekham, S. & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal Developmental Psychology, 5,* 125–137.
Phillips, A., Wellman, H., & Spelke, E. (2002). Infants' ability to connect gaze and emotional expression to intentional action. *Cognition, 85,* 53–78.
Prinz, W. (2002). Experimental approaches to imitation. *In Meltzoff, A. N., & Prinz ,W. (Eds.). The imitative mind: Development, evolution, and brain bases* (pp. 143-162). Cambridge: Cambridge University Press.
Poulin-Dubois, D., Brooker, I., & Chow, V. (2009). The developmental origins of naïve psychology in infancy. In Bauer, P. J. (Ed.), *Advances in child development and behavior* (Vol. 37, pp. 55–104). San Diego, CA: Academic Press.
Press, C., Gillmeister, H., & Heyes, C. (2007). Sensorimotor experience enhances automatic imitation of robotic action. *Proceedings of the Royal Society B: Biological Sciences, 274,* 2639-2644.
Rizzolati, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192.
Rizzolatti, G., Fogassi, L., & Gallese, V. (2006). Mirrors in the mind. *Scientific American, 295(5),* 54-61.

Ruby, P., & Decety, J. (2001). Effect of the subjective perspective taking during simulation of action: a PET investigation of agency. *Nature Neuroscience, 4,* 546–50.
Ruffman, T., Garnham, W., Import, A., & Connolly, D. 2001. Does Eye Gaze Indicate Knowledge of False Belief: Charting Transitions in Knowledge. *Journal of Experimental Child Psychology, 80,* 201-224.
Ruffman, T. and Perner, J. (2005) Do infants really understand false belief?: Response to Leslie. *Trends in Cognitive Science, 9*, 462–463
Russell, J. (2007). Controlling core knowledge: conditions for the ascription of intentional states to self and others by children. *Synthese, DOI 10.1007/s11229-007-9203-8.*
Scholl, B. J., & Leslie, A. M. (1999). Modularity, Development and 'Theory of Mind'. *Mind & Language, 14,* 131–153.
Scott, R., Song, H., Baillargeon, R., & Leslie, A. M. (2007). 18.5-month-old infants attribute to others false beliefs about objects' internal properties. *Paper presented at the biennial meeting of the Society for Research in Child Development, Boston, MA.*
Scott, R.M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. Child Development, 80, 1172–1196.
Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology, 18,* 668-671.
Sirois, S., & Jackson, I. (2007). Social cognition in infancy: A critical review of research on higher order abilities. *European Journal of Developmental Psychology, 4,* 46–64.
Sodian, B. (2005). Theory of mind. The case for conceptual development. In Schneider, W., Schumann-Hengsteler, R., & Sodian, B. (Eds.), *Young children's cognitive development. Interrelationships among working memory, theory of mind, and executive functions* (pp. 95–130). Hillsdale, NJ: Erlbaum.
Sodian, B., & Thoermer, C. (2004). Infants' understanding of looking, pointing, and reaching as cues to goal-directed action. *Journal of Cognition & Development, 5(3),* 289-316
Sodian, B., Thoermer, C., & Dietrich, N. (2006). Two- to four-year-old children's differentiation of knowing and guessing in a non-verbal task. *European Journal of Developmental Psychology, 3*, 222– 237.
Sodian et al. (forthcoming)
Sommerville, J. A., Woodward, A., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition, 96,* B1–B11.
Sommerville, J. A., Hildebrand, E. A., & Crane, C. C. (2008). Experience matters: The impact of doing versus watching on infants' subsequent perception of tool use events. *Developmental Psychology, 44,* 1249-1256.
Song, H., & Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psycholology, 44(6),* 1789–1795.
Southgate, V., Senju, A., & Csibra, G. (2007). Action Anticipation through Attribution of False Belief By Two-Year-Olds. *Psychological Science, 18,* 587–592.
Stich, S. (1978). Beliefs and Subdoxastic States. *Philosophy of Science 45,*499-518
Spaulding, S. 2010. Embodied cognition and mindreading. *Mind & Language, 25(1),* 119-140.
Surian, L., Caldi S, & Sperber, D. 2007. Attribution of beliefs to 13-month-old infants. *Psychological Science*, 18, 580–586.
Tomasello, M., Call, J., Hare, B. (2003). Chimpanzees Understand Psychological States - The Question is Which Ones and to What Extent. *Trends in Cognitive Sciences 7,* 153-6.
Träuble, B., Marinovic, V., & Pauen, S. (2010). Early theory of mind competencies—Do infants understand others' beliefs? *Infancy, 15,* 434–444.
Tsai, C. C., Kuo, W. J., Hung, D. L., & Tzeng, O. J. (2008). Action co-representation is tuned to other humans. *Journal of Cognitive Neuroscience, 20(11),* 2015–2024.
Watson, A., Painter, J., & Bornstein, M. (2002). Longitudinal relations between 2-year-olds' language and 4-year-olds' theory of mind. *Journal of Cognition and Development*, *2,* 449–457.
Wellman, H.M. 1990. *The Child's Theory of Mind*. Cambridge, MA: MIT Press.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of Theory of Mind Development: The Truth about False-Belief. *Child Development, 72(3),* 655–684.
Wellman, H. M. (2002). Understanding the psychological world: Developing a theory of mind. In Goswami, U. (Ed.), *Handbook of Childhood Cognitive Development* (pp. 167–187). Oxford: Blackwell.
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128.

Woodward, A. L. (1998). Infants Selectively Encode the Goal Object of an Actor's Reach. *Cognition, 69,* 1–34.
Woodward, A. L. (2003). Infants' developing understanding of the link between looker and object. *Developmental Science, 6*, 297–311.
Woodward, A. L., & Guajardo, J. J. (2002). Infants' understanding of the point gesture as an object-directed action. *Cognitive Development, 17,* 1061–1084.
Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition, 85,* 223–250
Xu, F., Carey, S., & Quint, N. (2004). The emergence of kind-based object individuation in infancy. *Cognitive Psychology, 49*, 155–190.

On the Tractability of Comparing Informational Structures*

Cédric Dégremont[1], Lena Kurzen[2], and Jakub Szymanik[3]

[1, 3]Institute of Artificial Intelligence, University of Groningen
[2]Institute for Logic, Language and Computation, University of Amsterdam

{cedric.uva | lena.kurzen | jakub.szymanik}@gmail.com

1 Introduction

Epistemic modal logics and their extensions are concerned with global and abstract problems in reasoning about information. One of the features of that approach is its struggle for flexibility: it aims at designing logical systems that can model a large variety of epistemic scenarios [9, 4]. Hence, it is not surprising that the trade-off between expressivity and complexity has been one of the central problems in the epistemic logic literature. Logics need to be quite complex to account for a wide range of problems and it is not a surprise that there are many intractability results in the literature (see e.g., [13] and [5] for a survey).

One of the aims of this paper is to initiate the mapping of the tractability border among the epistemic tasks rather than epistemic logics. As a result, we can identify a theoretical threshold in the difficulty of reasoning about information, as was already done in the context of reasoning with quantifiers (see [19, 20]). In order to do this, we shift our perspective: Instead of investigating the complexity of a given logic that may be used to describe a problem, we turn towards a complexity study of that concrete problem itself, determining what computational resources are needed in order to perform the reasoning. Focusing on specific problems, things may be much easier since concrete problems involved in the study of multi-agent interaction are rarely as general as e.g. satisfiability. In most cases, checking whether a given property is satisfied in a given (minimal) epistemic scenario is sufficient. Hence, many problems turn out to be tractable. Still, we will see that even in this perspective there are some intractable problems. This feasibility border in epistemic tasks seems to be an interesting new topic for a formal study. Moreover, in principle the cognitive plausibility of the border could be empirically assessed by checking whether it correlates with the difficulties faced by human subjects (cf. [23, 21]). So in a sense, we aim to initiate a search for an appropriate perspective and complexity measures that describe in plausible ways the cognitive difficulties agents face while interacting. Certain experimental results in the economics literature [24, 10] explore similar directions. In general, the approach we have described in this paper focuses exclusively on the abstract information structure leaving out any concept of preferences and strategic reasoning.

In this paper we investigate the computational complexity of various decision problems that are relevant for interactive reasoning in epistemic modal logic frameworks. In particular, we explore the complexity of manipulating and comparing information structures possessed by different agents. For instance, we are interested in how difficult it is to answer the following questions.

*Cédric Dégremont and Jakub Szymanik gratefully acknowledge the support of Vici grant NWO-277-80-001.

- Is one agent's information strictly less refined than another agents' information?
- Do two agents have the same knowledge/belief about each other's knowledge/belief?
- Given two agents, is it possible to give some information to one of them such that as a result
  - both agents have similar information structures? (cf. [22].)
  - one of them has more refined information than the other?

For determining the computational complexity of the different problems, we use complexity results from graph theory (see e.g. [12]). Thus, we also clarify the computational impact of assuming S5 accessibility relations in epistemic models, i.e., the impact of assuming partition-based information structures on the complexity of various problems.

After giving the preliminaries in Section 2, we discuss four types of epistemic tasks and their computational complexity: informational similarity (Section 3.1), informational symmetry (Section 3.2) and two kinds of informational manipulation (Section 3.3 and 3.4). Omitted proofs can be found in the appendix. Section 4 concludes.

2 Preliminaries

2.1 Modeling information

We use relational structures from epistemic logic for modeling information (cf. [6, 9]). Kripke models can compactly represent the information agents have about the world and about the information possessed by the other agents. In what follows, $N = \{1, \ldots, n\}$ is a fixed finite set of agents and PROP is a countable set of propositional variables.

Definition 2.1 (Kripke Models)**.** A *Kripke model* $\mathcal{M}$ based on a set of agents N is of the form $(W, (R_i)_{i\in N}, V)$, where $W \neq \emptyset$, for each $i \in N$, R_i is a binary relation on W, and $V : \text{PROP} \to \wp(W)$.

It is frequently assumed that information structures are partition-based [1, 9, 16]:

Definition 2.2 (Epistemic Models)**.** An *epistemic model* is a Kripke model such that for all $i \in N$, R_i is an equivalence relation. (We usually write $\sim_i$ instead of R_i).

We write $|\mathcal{M}|$ to refer to the size of the model $\mathcal{M}$, and $Dom(\mathcal{M})$ to refer to the domain of $\mathcal{M}$. We refer to a pair $(\mathcal{M}, w)$ with $w \in Dom(\mathcal{M})$ as a pointed model. Intuitively R_i encodes i's uncertainty: if sR_it, then if the actual world were s then i would consider it possible that the actual world is t. For any non-empty set $G \subseteq N$, we write R^*_G for the reflexive transitive closure of $\bigcup_{i\in G} R_i$.

2.2 Comparing models and reasoning about submodels

In what follows, we need a reasonable notion of two models being similar. In addition to the notion of isomorphism, we make use of the notions of simulation, simulation equivalence and bisimulation.

Definition 2.3 (Simulation)**.** We say that a pointed Kripke model $(\mathcal{M}, s)$, where $\mathcal{M} = (W, (R_i)_{i\in N}, V)$ and $s \in W$, is simulated by another pointed model $(\mathcal{M}', s')$ (which we denote by $(\mathcal{M}, s) \sqsubseteq (\mathcal{M}', s')$) such that $\mathcal{M}' = (W', (R'_i)_{i\in N}, V')$ with $s' \in W'$ if there exists a binary relation $Z \subseteq W \times W'$ such that sZs' and for any two states x, x' whenever xZx' then for all $i \in N$:

1. x, x' verify the same proposition letters.

2. if xR_iz in $\mathcal{M}$ then there exists some $z' \in W'$ with $x'R'_iz'$ and zZz'.

We say that $\mathcal{M} = (W, (R_i)_{i\in N}, V)$ is simulated by $\mathcal{M}' = (W', (R'_i)_{i\in N}, V')$ (denoted by $\mathcal{M} \sqsubseteq \mathcal{M}'$) if there are $s \in W$ and $s' \in W'$ such that $(\mathcal{M}, s) \sqsubseteq (\mathcal{M}', s')$. We say that a simulation $Z \subseteq W \times W'$ is *total* if for every $s \in W$, there is some $t \in W'$ such that sZt, and for every $t \in W'$, there is some $s \in W$ such that sZt. If $\mathcal{M}$ is simulated by $\mathcal{M}'$ by means of a total simulation, we say $\mathcal{M} \sqsubseteq_{total} \mathcal{M}'$. Moreover, we say that $\mathcal{M} = (W, (R_i)_{i\in N}, V)$ and $\mathcal{M}' = (W', (R'_i)_{i\in N}, V')$ are simulation equivalent if $\mathcal{M}$ simulates $\mathcal{M}'$ and $\mathcal{M}'$ simulates $\mathcal{M}$. The following notion is stronger than simulation equivalence.

Definition 2.4 (Bisimulation)**.** A local bisimulation between two pointed Kripke models with set of agents N, $(\mathcal{M}, s)$ with $\mathcal{M} = (W, (R_i)_{i\in N}, V)$ and $(\mathcal{M}', t)$ with $\mathcal{M}' = (W', (R'_i)_{i\in N}, V')$ is a binary relation $Z \subseteq W \times W'$ such that sZs' and also for any worlds x, x' whenever xZx' then for all $i \in N$:

1. x, x' verify the same proposition letters.

2. if xR_iu in $\mathcal{M}$ then there exists $u' \in W'$ with $x'R'_iu'$ and uZu'.

3. if $x'R'_iu'$ in $\mathcal{M}'$ then there exists $u \in W$ with xR_iu and uZu'.

We say that $\mathcal{M} = (W, (R_i)_{i\in N}, V)$ and $\mathcal{M}' = (W', (R'_i)_{i\in N}, V')$ are bisimilar ($\mathcal{M} \underline{\leftrightarrow} \mathcal{M}'$) if there are $s \in W$ and $s' \in W'$ such that $(\mathcal{M}, s) \underline{\leftrightarrow} (\mathcal{M}', s')$. A bisimulation $Z \subseteq Dom(\mathcal{M}) \times Dom(\mathcal{M}')$ is *total* if for every $s \in Dom(\mathcal{M})$, there is some $t \in Dom(\mathcal{M}')$ such that sZt, and for every $t \in Dom(\mathcal{M}')$, there is some $s \in Dom(\mathcal{M})$ such that sZt. Then we write $\mathcal{M} \underline{\leftrightarrow}_{total} \mathcal{M}'$.

To reason about informational structures that can be obtained by providing agents with new information, we use the notions of submodel and generated submodel.

Definition 2.5 (Submodel)**.** We say that $\mathcal{M}'$ is a submodel of $\mathcal{M}$ iff $W' \subseteq W$, $\forall i \in N$, $R'_i = R_i \cap (W' \times W')$, $\forall p \in \text{PROP}$, $V'(p) = V(p) \cap W'$.

The notion of *induced subgraph* is just like that of a submodel without the condition for the valuations. The notion of *subgraph* is weaker than that of an induced subgraph as it allows that $R'_i \subset R_i \cap W' \times W')$.

Definition 2.6 (Generated submodel)**.** We say that $\mathcal{M}' = (W', (R_i)'_{i\in N}, V')$ is a generated submodel of $\mathcal{M} = (W, (R_i)_{i\in N}, V)$ iff $W' \subseteq W$ and $\forall i \in N$, $R'_i = R_i \cap (W' \times W')$, $\forall p \in \text{PROP}$, $V'(p) = V(p) \cap W'$ and if $w \in W'$ and wR_iv then $v \in W'$. The submodel of $\mathcal{M}$ generated by $X \subseteq W$ is the smallest generated submodel $\mathcal{M}'$ of $\mathcal{M}$ with $X \subseteq Dom(\mathcal{M}')$.

We write $\mathcal{K}_i[w] := \{v \in W \mid wR_iv\}$ to denote i's information set at w and $R^*_G[w] := \{v \in W \mid wR^*_Gv\}$. This notion is generalized by the concept of horizon:

Definition 2.7 (Horizon)**.** The *horizon* of i at $(\mathcal{M}, w)$ (notation: $(\mathcal{M}, w)^i$) is the submodel generated by $\mathcal{K}_i[w]$.

This paper will not use syntactic notions. In terms of intuition, the important definition is that of knowledge K_i: agent i knows ϕ at w if ϕ is true in all states that i considers possible at w. In equivalent semantic terms: i knows E if $E \subseteq \mathcal{K}_i[w]$. E is common knowledge in a group G at w iff $E \subseteq R^*_G[w]$.

2.3 Tractability

Some problems, although computable, nevertheless require too much time or memory to be feasibly solved by a realistic computational device. Computational complexity theory investigates the resources (time, memory, etc.) required for the execution of algorithms and the inherent difficulty of computational

problems [17]. In particular, we want to identify efficiently solvable problems and draw a line between tractability and intractability. In general, the most important distinction is that between problems which can be computed in polynomial time with respect to their size, and those which are believed to have only exponential time algorithmic solutions. The class of problems of the first type is called PTIME (P for short); one can demonstrate that a problem belongs to this class if one can show that it can be computed by a deterministic Turing machine in polynomial time. Problems belonging to the second class are referred to as NP-hard. They are at least as difficult as problems belonging to the NPTIME (NP) class; this is the class of problems which can be computed by nondeterministic Turing machines in polynomial time. NP-complete problems are NP-hard problems belonging to NPTIME, hence they are intuitively the most difficult problems among the NPTIME problems.

3 Complexity of comparing and manipulating information

3.1 Information similarity

The first natural question we would like to address is whether an agent in a given situation has similar information to the one possessed by some other agent (in a possibly different situation). One very strict way to understand such similarity is through the use of isomorphism.

For the general problem of checking whether two Kripke models are isomorphic, we can give tight complexity bounds, as this problem is polynomially equivalent to graph isomorphism. The graph isomorphism problem is neither known to be NP-complete nor to be tractable and the set of problems with a polynomial-time reduction to the graph isomorphism problem is called GI.

Decision Problem 3.1 (Kripke model isomorphism)**.**
Input: *Pointed Kripke models* $(\mathcal{M}_1, w_1)$, $(\mathcal{M}_2, w_2)$.
Question: *Are* $(\mathcal{M}_1, w_1)$ *and* $(\mathcal{M}_2, w_2)$ *isomorphic, i.e. is it the case that* $(\mathcal{M}_1, w_1) \cong (\mathcal{M}_2, w_2)$*?*

Fact 3.2. *Kripke model isomorphism is GI-complete.*

However, isomorphism is arguably a too restrictive notion of similarity. Bisimilarity is a weaker but still a very natural concept of similarity for relational structures. Here the question arises as to whether working with $S5$ models – a common assumption in the epistemic logic and interactive epistemology literature – rather than arbitrary Kripke structures has an influence on the complexity of the task.

Decision Problem 3.3 (Epistemic model bisimilarity)**.**
Input: *Two pointed multi-agent epistemic S5 models* $(\mathcal{M}_1, w_1), (\mathcal{M}_2, w_2)$.
Question: *Are the two models bisimilar, i.e.* $(\mathcal{M}_1, w_1) \underline{\leftrightarrow} (\mathcal{M}_2, w_2)$*?*

In [3], it has been shown that deciding bisimilarity is P-complete for finite labelled transition systems. It follows that epistemic models bisimilarity is also in P.

Fact 3.4. *Multi-agent epistemic S5 model bisimulation can be done in polynomial time with respect to the size of the input* $(|\mathcal{M}_1| + |\mathcal{M}_2|)$.

Thus, multi-agent epistemic S5 model bisimilarity is in P. Now, of course the question arises if it is also P-hard.[1]
Open problem *Is multi-agent epistemic model (S5) bisimulation P-hard?*

[1] We conjecture that we can show P hardness using methods of simulating an arbitrary relation R by a combination of two equivalence relations $\sim_1$ and $\sim_2$ as follows: we replace each wRv by $w \sim_1 z \sim_2 v$, for a new state z.

Without any assumptions on the accessibility relations for the agents, we immediately get P-completeness for multi-agent Kripke models as the problem is equivalent to bisimilarity for finite labelled transition systems.

The picture. Deciding whether two models are bisimilar is tractable for S5 epistemic models, and in case of the arbitrary Kripke structures it is among the hardest tractable problems. Kripke model isomorphism lives on the tractability border. It is open whether isomorphism for (partition-based) epistemic models is tractable and whether epistemic S5 model bisimilarity is P-complete, which we indeed conjecture to be the case.

3.2 Informational symmetry: knowing what others know

The preceding notions of similarity are very strong. In the context of analyzing epistemic interactions between agents, weaker notions of similarity are of interest. In general, the information that agents have about each other's information state plays a crucial role. We will now analyze the problem of deciding whether two agents' views about the interactive epistemic structure, and in particular about the knowledge of other agents, are equivalent. A first reading is simply to fix some fact $E \subseteq W$ and ask whether E is common knowledge in a group G. Clearly this problem is tractable.

Fact 3.5. *Given a pointed model $(\mathcal{M}, w)$, some $E \subseteq Dom(\mathcal{M})$ and $G \subseteq N$, deciding whether E is common knowledge in the group G at w can be done in polynomial time.*

Proof. From reachability for R^*_G. □

However, instead of fixing some specific fact of interest, the question might be whether a situation is symmetric with respect to two given agents, say Alice and Bob. In other words, is the interactive informational structure from Alice's perspective similar to how it is from Bob's perspective?

Definition 3.6. We write $\mathcal{M}[i/j]$ to be the model obtained by switching labels between i and j.

Definition 3.7. We say that two pointed multi-agent epistemic models $(\mathcal{M}, s)$ and $(\mathcal{M}', s')$ (with set of agents N) are flipped bisimiliar for agents $i, j \in N$, $(\mathcal{M}, s) \underline{\leftrightarrow}_f^{(i,j)} (\mathcal{M}', s')$, iff $(\mathcal{M}, s) \underline{\leftrightarrow} (\mathcal{M}'[i/j], s')$.

A natural question is the relation of flipped bisimulation to the fact that all knowledge of both agents is common knowledge. The following is immediate:

Observation 3.8. *If in $\mathcal{M}$, $\sim^*_{\{1,2\}} \subseteq \sim_j$ for $j \in \{1, 2\}$, then for all $w \in Dom(\mathcal{M})$, $(\mathcal{M}, w) \underline{\leftrightarrow}_f^{(1,2)} (\mathcal{M}, w)$.*

Is other direction true? Locally, even on S5 models, flipped self-bisimulation is a much weaker requirement: it does not even imply that (shared) knowledge of facts is common knowledge:

Fact 3.9. *There exists a pointed S5 epistemic model which is a, b-flipped bisimilar to itself, where the two agents know that p (with $p \in$ PROP), and p is not common knowledge between a and b.*

But required globally of every state, we do have the following converse:

Fact 3.10. *Let $\mathcal{M}$ be a transitive model with $w \in Dom(\mathcal{M})$. Whenever the submodel $\mathcal{M}'$ of $\mathcal{M}$ generated by $\{w\}$ is such that every state is $0, 1$-flipped bisimilar to itself, then for any $p \in$ PROP, if j knows p at w (i.e., $V(p) \subseteq K_j[w]$) for some $j \in \{0, 1\}$ then p is common knowledge between 0 and 1 at w.*

Let us recall the notion of horizon (see Definition 2.7). It is the submodel generated by the information set of the agent: the *horizon* of i at (M, w) (notation: $(M, w)^i$) is the submodel generated by $\mathcal{K}_i[w]$.

Decision Problem 3.11 (Flipped multi-agent epistemic model horizon bisimilarity)**.**
Input: *Two pointed multi-agent epistemic models* $(\mathcal{M}, w)$*,* $(\mathcal{M}', w')$*, two agents* i*,* j*.*

Question: *Are the horizons of agents* i *and* j *in* $(\mathcal{M}, w)$ *and* $(\mathcal{M}', w')$ *respectively flipped bisimilar for* i, j*, namely is it the case that:* $(\mathcal{M}, w)^i \underline{\leftrightarrow} (\mathcal{M}', w')^j [i/j]$*?*

Fact 3.12. *Flipped multi-agent epistemic S5 model horizon bisimilarity is in P. Given a multi-agent epistemic model* $(\mathcal{M}, w)$*, it is trivial to decide if for two agents* i, j *it holds that* $(\mathcal{M}, w)^i \underline{\leftrightarrow} (\mathcal{M}, w)^j [i/j]$*.*

Proof. We can use a polynomial algorithm for Kripke model bisimilarity. Horizons of two agents at the same point in a model are always equal in S5 because of reflexivity of the accessibility relations. □

Fact 3.13. *Without any assumptions on the accessibility relations, the computational complexity of multi-agent Kripke model flipped horizon bisimilarity is P-complete.*

Proof. Follows from [3] and the fact that in general the horizons of two agents can be disjoint. □

The picture. Deciding horizon flipped bisimilarity in Kripke models is among the hardest tractable problems. It is trivial for partition-based models. Deciding whether a fact is commonly known is tractable.

3.3 Can we reshape an agent's mind into some desired informational state?

So far, we have been comparing agents' informational states within models. The next interesting problem is to decide whether new informational states (satisfying desired properties) can be achieved in certain ways. One immediate question is whether one can give some information to an agent (i.e. to restrict her horizon) such that after the update her horizon is bisimilar to the horizon of some other agent. Concretely, we would like to know if there is any type of information that could reshape some agent's information to fit some desired new informational state or at least be similar to it. We will thus investigate the task of checking whether there is a submodel that has certain properties. This means that we determine if it is possible to purposely refine a model in a certain way. This question is in line with problems addressed by arbitrary public announcement logic and arbitrary event modal logic [2, 11, 22].[2]

We start with the problem of checking whether there is a submodel of one model that is bisimilar to another one. On graphs, this is related to the problem of deciding if one contains a subgraph bisimilar to another. Note that in the problem referred to in the literature as "subgraph bisimulation" [8], the subgraph can be any graph whose vertices are a subset of the vertices of the original graph, and the edges can be any subset of the edges of the original graph restricted to the subset of vertices. To be more specific, the problem investigated in [8] is the following:

> Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, is there a graph $G_2' = (V_2', E_2')$ with $V_2' \subseteq V_2$ and $E_2' \subseteq E_2$ such that there is a total bisimulation between G_2' and G_1?

Since we want to investigate the complexity of reasoning about epistemic interaction using modal logic, we are interested in subgraphs that correspond to *relativization* in modal logic: induced subgraphs. This leads us to an investigation of *induced* subgraph bisimulation.

Decision Problem 3.14 (Induced subgraph bisimulation)**.**
Input: *Two finite graphs* $G_1 = (V_1, E_1)$*,* $G_2 = (V_2, E_2)$*,* $k \in \mathbb{N}$*.*

Question: *Is there an induced subgraph of* G_2 *with at least* k *vertices that is bisimilar to* G_1*, i.e. is there some* $V' \subseteq V_2$ *with* $|V'| \geq k$ *and* $(V', E_2 \cap (V' \times V')) \underline{\leftrightarrow}_{total} G_1$*?*

[2] Note that in the current work, we focus on the semantic structures only and do not require that the submodel can be characterized by some formula in a certain epistemic modal language.

Even though the above problem looks very similar to the original subgraph bisimulation problem (NP-hardness of which is shown by reduction from Hamiltonian Path), NP-hardness does not follow immediately.[3] Nevertheless, we can show NP-hardness by reduction from Independent Set.

Proposition 3.15. *Induced subgraph bisimulation is NP-complete.*

Now, an analogous result for Kripke models follows. The intuitive interpretation here (with an epistemic/doxastic interpretation of the accessibility relation) is whether it is possible to 'gently' restrict one model without letting its domain get smaller than k such that afterwards it is bisimilar to another model. The intuition is that we would like the new information to change as minimally as possible the informational state of the target agent.

Decision Problem 3.16 (Submodel bisimulation for Kripke models)**.**
Input: *Kripke models $\mathcal{M}_1$, $\mathcal{M}_2$ with set of agents N, $k \in \mathbb{N}$.*
Question: *Is there a submodel $\mathcal{M}_2'$ of $\mathcal{M}_2$ with $|Dom(\mathcal{M}_2')| \geq k$ such that $\mathcal{M}_1$ and $\mathcal{M}_2'$ are totally bisimilar i.e. $\mathcal{M}_1 \underline{\leftrightarrow}_{total} \mathcal{M}_2'$?*

Corollary 3.17. *Submodel bisimulation for Kripke models is NP-complete.*

As we are interested in the complexity of reasoning about the interaction of epistemic agents as it is modeled in (dynamic) epistemic logic, let us now see how the complexity of induced subgraph bisimulation changes when we make the assumption that models are partitional, i.e. that the relation is an equivalence relation, as it is frequently assumed in the AI or interactive epistemology literature. We will see that this assumption makes the problem significantly easier.

Proposition 3.18. *If for graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, E_1 and E_2 are reflexive, transitive and symmetric, then induced subgraph bisimulation for G_1 and G_2 can be solved in linear time.*

Assuming the edge relation in a graph to be an equivalence makes induced subgraph bisimulation a trivial problem because, unless its set of vertices is empty, every such graph is bisimilar to the graph $(\{v\}, \{(v, v)\})$. But for S5 models this is of course not the case, as the bisimulation takes into account the valuation. Nevertheless, we will now show that also for single agent S5 models, the problem of submodel bisimulation is significantly easier than in the case of arbitrary single agent Kripke models. To be more precise, we will distinguish between two problems:

The first problem is *local* single agent S5 submodel bisimulation. Here we take as input two pointed S5 models. Then we ask whether there is a submodel of the second model that is bisimilar to the first one. Thus, the question is whether it is possible to restrict one of the models in such a way that there is a state in which the agent has exactly the same information as in the situation modeled in the other model.

Decision Problem 3.19 (Local S5 submodel bisimulation for single agent epistemic models)**.**
Input: *A pointed S5 epistemic model $(\mathcal{M}_1, w)$ with $\mathcal{M}_1 = (W_1, \sim_1, V_1)$ and $w \in W_1$, and an S5 epistemic model $\mathcal{M}_2 = (W_2, \sim_2, V_2)$.*
Question: *Is there a submodel $\mathcal{M}_2' = (W_2', \sim_2', V_2')$ of $\mathcal{M}_2$ such that $(\mathcal{M}_1, w) \underline{\leftrightarrow} (\mathcal{M}_2', w')$ for some $w' \in Dom(\mathcal{M}_2')$?*

Proposition 3.20. *Local submodel bisimulation for single agent pointed epistemic models is in P.*

The second problem we consider is *global* S5 submodel bisimulation, where the input are two models $\mathcal{M}_1$ and $\mathcal{M}_2$ and we ask whether there exists a submodel of $\mathcal{M}_2$ such that it is totally bisimilar to $\mathcal{M}_1$.

[3] For Induced Subgraph Bisimulation, a reduction from Hamiltonian Path seems to be more difficult, as does a direct reduction from the original subgraph bisimulation problem.

Decision Problem 3.21 (Total S5 submodel bisimulation for single agent epistemic models)**.**
Input: *Two S5 epistemic models* $\mathcal{M}_1 = (W_1, \sim_1, V_1), \mathcal{M}_2 = (W_2, \sim_2, V_2)$.
Question: *Is there a submodel* $\mathcal{M}'_2 = (W'_2, \sim'_2, V'_2,)$ *of* $\mathcal{M}_2$ *such that* $\mathcal{M}_1 \underline{\leftrightarrow}_{total} \mathcal{M}'_2$*?*

We can show that even though the above problem seems more complicated than local submodel bisimulation, it can still be solved in polynomial time. The proof uses the fact that finding a maximum matching in a bipartite graph can be done in polynomial time (see e.g. [18]).

Theorem 3.22. *Total submodel bisimulation for single agent epistemic models is in P.*

Now, the question arises whether the above results also hold for the multi-agent case.

Decision Problem 3.23 (Global submodel bisimulation for multi-agent pointed epistemic models)**.**
Input: *Two epistemic models* $\mathcal{M}_1 = (W_1, (\sim_{1i})_{i \in N}, V_1), \mathcal{M}_2 = (W_2, (\sim_{2i})_{i \in N}, V_2)$*, for* N *being a finite set (of agents), and* $k \in \mathbb{N}$.
Question: *Is there a submodel* $\mathcal{M}'_2 = (W'_2, (\sim'_{2i})_{i \in N}, V'_2)$ *of* $\mathcal{M}_2$ *such that* $\mathcal{M}_1 \underline{\leftrightarrow}_{total} \mathcal{M}'_2$*?*

We conjecture that using similar ideas to those outlined in footnote 1 to show that the above problem is NP-complete for models with at least two agents.

The picture. Induced subgraph bisimulation is intractable (NP-complete) and so is submodel bisimulation for arbitrary Kripke models. For S5 models, induced subgraph bisimulation is tractable, and so are local submodel bisimulation and total submodel bisimulation in the single agent case. We think that NP-completeness can be shown for the case of at least two agents.

3.4 Simulation vs Bisimulation

In dynamic systems with diverse agents, an interesting question is whether it is possible to give some information to one agent such that afterwards she knows at least as much as some other agent. This is captured by an asymmetric notion, that of simulation. With this difference, the question can be raised of the effect on tractability and intractability of requiring simulation versus requiring bisimulation. With this motivation, we would like to explore the problem of induced subgraph simulation.

Decision Problem 3.24 (Induced subgraph simulation)**.**
Input: *Two finite graphs* $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$, $k \in \mathbb{N}$.
Question: *Is there an induced subgraph of* G_2 *with at least* k *vertices that is simulated by* G_1*, i.e., is there some* $V' \subseteq V_2$ *with* $|V'| \geq k$ *and* $(V', E_2 \cap (V' \times V')) \sqsubseteq_{total} G_1$*?*

Proposition 3.25. *Induced subgraph simulation is NP-complete.*

In [7], it has been shown that given two graphs it is also NP-complete to decide if there is a subgraph (not necessarily an induced one) of one such that it is simulation equivalent to the other graph. Here, we show that this also holds if the subgraph is required to be an induced subgraph.

Decision Problem 3.26 (Induced subgraph simulation equivalence)**.**
Input: *Two finite graphs* $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$, $k \in \mathbb{N}$.
Question: *Is there an induced subgraph of* G_2 *with at least* k *vertices that is similar to* G_1*, i.e. is there some* $V' \subseteq V_2$ *with* $|V'| \geq k$ *and* $(V', E_2 \cap (V' \times V')) \sqsubseteq_{total} G_1$ *and* $G_1 \sqsubseteq_{total} (V', E_2 \cap (V' \times V'))$*?*

Proposition 3.27. *Induced subgraph simulation equivalence is NP-complete.*

As a corollary of the two previous propositions, we get that for arbitrary Kripke models both submodel simulation and submodel equivalence are NP-complete. We conjecture that for single agent S5, we can use similar methods as used in the proof of Theorem 3.22. Let us conclude with an interesting open question, as to whether the results from [15] also hold for epistemic models.

Open problem *Is deciding simulation (equivalence) of epistemic models at least as hard as deciding bisimilarity?*

The picture. Induced subgraph simulation and equivalence are both intractable (NP-complete). The same holds for Kripke model simulation (equivalence). It remains to be investigated for epistemic models.

4 Conclusions and Further Work

In this work, we have identified concrete epistemic tasks related to the comparison and manipulation of informational states of agents in possibly different situations. Interestingly, our complexity analysis shows that the preceding problems live on both sides of the border between tractability and intractability:

Problem	**Tractable?**	**Comments**
Kripke model isomorphism	unknown	in GI
Epistemic model bisimilarity	Yes	Conjecture: P-hard for ≥ 2 agents
Flipped horizon bisimilarity	Yes	P-complete for arbitrary models
Kripke submodel bisimulation	No	NP-complete for arbitrary models; in linear time for S5
Local S5 submodel bisimulation	Single agent: Yes	unknown
Total S5 submodel bisimulation	Single agent: Yes	Conjecture: NP-complete for ≥ 2 agents
Kripke submod. simulation (equiv.)	No	Conjecture: in P for single agent S5

Table 1: Summary of the results and open questions.

As such, this work is a first step towards mapping out the complexity of concrete epistemic problems based on epistemic modeling. It would be interesting to systematize this approach to a larger class of problems. Further work to complete the picture includes the open problems that we mentioned in our analysis in Section 3. Solving them would clarify the border between tractability and intractability in the domain of epistemic reasoning tasks. This would then also shed some light on the more general question as to what is the impact of the assumption of S5 on the complexity of certain problems from graph theory. It would moreover clarify whether for some epistemic tasks, moving from single agent to multi-agent scenarios has the consequence of crossing the border between tractability and intractability.

How would we like to interpret our results? One conclusion, we can draw from our case study is that assuming partition-based information structures simplifies epistemic tasks of comparing and manipulating informational structures. In particular, we saw that comparing agents's informational structures via bisimulation is tractable in the multi-agent case, meaning that it should be relatively easy to say whether Alice's information is strictly less refined than Bob's. Furthermore, deciding whether two agents have symmetric knowledge about each other's knowledge should be also in principle easy (PTIME for S5 models and P-complete for arbitrary models). Finally, we proved that things are getting harder if one wants to know wether a certain manipulation of agents' knowledge is possible. Deciding whether the information structure of Alice is more refined than that of Bob is in general intractable, independently of choosing bisimilarity or isomorphism as our notion of similarity. However, the problem becomes easy if one assumes that agents's knowledge can be modeled by equivalence relations. On the other hand,

substituting bisimulation by simulation gives rise to an interesting open problem whether computing simulation equivalence of epistemic models is a at least as hard as deciding their bisimilarity.

References

[1] Robert J. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28(3):263–300, 1999.

[2] Philippe Balbiani, Alexandru Baltag, Hans van Ditmarsch, Andreas Herzig, Tomohiro Hoshi, and Tiago de Lima. 'knowable' as 'known after an announcement'. Technical Report IRIT/RR-2008-2-FR, IRIT, University of Toulouse 3, 2008. URL: ftp://ftp.irit.fr/IRIT/LILAC/2008_Report_Balbiani_et_al.pdf.

[3] José L. Balcázar, Joaquim Gabarró, and Miklos Santha. Deciding bisimilarity is P-complete. *Formal Aspects of Computing*, 4(6A):638–648, 1992.

[4] Alexandru Baltag and Lawrence S. Moss. Logics for epistemic programs. *Synthese*, 139(2):165–224, 2004.

[5] Johan van Benthem and Eric Pacuit. The tree of knowledge in action: towards a common perspective. In I. Hodkinson G. Governatori and Y. Venema, editors, *Advances in Modal Logic*, volume 6. College Publications, 2006.

[6] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Number 53 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, UK, 2001.

[7] Lorenzo De Nardo, Francesco Ranzato, and Francesco Tapparo. The subgraph similarity problem. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):748–749, 2009.

[8] Agostino Dovier and Carla Piazza. The subgraph bisimulation problem. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):1055–1056, 2003.

[9] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. MIT Press, Cambridge, 1995.

[10] Nick Feltovich. Reinforcement-based vs. beliefs-based learning in experimental asymmetric-information games. *Econometrica*, 68:605–641, 2000.

[11] Tim French and Hans van Ditmarsch. Undecidability for arbitrary public announcement logic. In Carlos Areces and Robert Goldblatt, editors, *Advances in Modal Logic*, pages 23–42. College Publications, 2008.

[12] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

[13] Joseph Y. Halpern and Moshe Y. Vardi. The complexity of reasoning about knowledge and time. I. Lower bounds. *Journal of Computer and Systems Science*, 38(1):195–237, 1989.

[14] Monika R. Henzinger, Thomas A. Henzinger, and Peter W. Kopke. Computing simulations on finite and infinite graphs. In *FOCS '95: Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 453–462. IEEE Computer Society Press, 1995.

[15] Antonín Kučera and Richard Mayr. Why is simulation harder than bisimulation? In *CONCUR '02: Proceedings of the 13th International Conference on Concurrency Theory*, pages 594–610, London, UK, 2002. Springer-Verlag.

[16] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[17] Christos H. Papadimitriou. *Computational Complexity*. Addison Wesley, November 1993.

[18] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1982.

[19] Ian Pratt-Hartmann and Lawrence S. Moss. Logics for the relational syllogistic. *The Review of Symbolic Logic*, 2(04):647–683, 2009.

[20] Jakub Szymanik. Computational complexity of polyadic lifts of generalized quantifiers in natural language. *Linguistics and Philosophy*, 33:215–250, 2010.

[21] Jakub Szymanik and Marcin Zajenkowski. Comprehension of simple quantifiers. Empirical evaluation of a computational model. *Cognitive Science: A Multidisciplinary Journal*, 34(3):521–532, 2010.

[22] Hans van Ditmarsch and Tim French. Simulation and information: Quantifying over epistemic events. In *Knowledge Representation for Agents and Multi-Agent Systems: First International Workshop, KRAMAS 2008, Sydney, Australia, September 17, 2008, Revised Selected Papers*, pages 51–65, Berlin, Heidelberg, 2009. Springer-Verlag.

[23] Rineke Verbrugge. Logic and social cognition. The facts matter, and so do computational models. *Journal of Philosophical Logic*, 38(6):649–680, 2009.

[24] Roberto Weber. Behavior and learning the "dirty faces" game. *Experimental Economics*, 4:229–242, 2001.

A Proofs of selected theorems

Fact: 3.9 There exists a pointed S5 epistemic model which is a,b-flipped bisimilar to itself, where the two agents knows that p (with $p \in$ PROP), and p is not common knowledge between a and b.

Proof. Consider the model $\mathcal{M} = \langle W, \sim_a, \sim_b, V\rangle$ with $W = \{-2,-1,0,1,2\}$, $\sim_a$ is the smallest equivalence relation on W containing $\{(-2,-1),(0,1)\}$, $\sim_b$ is the smallest equivalence relation on W containing $\{(-1,0),(1,2)\}$, and $V(p) = \{-1,0,1\}$. It is easy to check that both Alice and Bob knows p at 0: $\mathcal{K}_a[0] = \{0,1\} \subseteq V(p)$ and $\mathcal{K}_b[0] = \{-1,0\} \subseteq V(p)$. Also p is not common knowledge between Alice and Bob at 0, indeed $0 \sim_a 1 \sim_b 2$ and $2 \notin V(p)$. Now it remains to show that $\mathcal{M}, 0$ is a,b-flipped bisimilar to itself. The flipped bisimulation is defined as $Z = \{(n, 0-n) \mid n \in W\}$. It is easily checked by inspection that Z is indeed a a,b-flipped bisimulation. □

Fact 3.10: Let $\mathcal{M}$ be a transitive model with $w \in Dom(\mathcal{M})$. If the submodel $\mathcal{M}'$ of $\mathcal{M}$ generated by $\{w\}$ is such that every state is $0,1$-flipped bisimilar to itself. Then for any $p \in$ PROP, if j knows p at w, i.e., $(V(p) \subseteq K_j[w]$, for some $j \in \{0,1\}$, then p is common knowledge between 0 and 1) at w.

Proof. We prove the contrapositive. Assume that p is not common knowledge between 0 and 1 at w. It follows that we have a $0,1$-path of length n with $n \in \omega$ of the form $wR_f(1)w_1Rf(2)\dots Rf(n-1)w_{n-1}$ with $w_{n-1} \notin V(p)$ and $f(k) \in \{0,1\}$ for all $k \in n$. Clearly, all the states in the preceding sequence are in $\mathcal{M}'$ so they must be $0,1$-flipped bisimilar to themselves and, in particular, to w. Hence, by definition of a flipped bisimulation we have a sequence of the form $wR^1_f(1)w^1_1Rf^1(2)\dots Rf^1(n-1)w^1_{n-1}$. Especially, $w^1_{n-1} \notin V(p)$ and $f^1(k) = |1 - f(k)|$. Iterating the process we can obtain a sequence of the form $wR^{n-1}_f(1)w^{n-1}_1Rf^{n-1}(2)\dots Rf^{n-1}(n-1)w^{n-1}_{n-1}$ with $w^{n-1}_{n-1} \notin V(p)$ and with f^{n-1} being one of the constant functions of the co-domain $\{0,1\}$. By transitivity it follows that $w^{n-1}_{n-1} \in K_0[w] \cap K_1[w] \cap \overline{V(p)}$, contradicting the assumption that at least of the agents knew p at w. □

Proposition 3.15: Induced subgraph bisimulation is NP-complete.

Proof. Showing that the problem is in NP is straightforward. Hardness is shown by reduction from Independent Set. First of all, let $I_k = (V_{I_k}, E_{I_k} = \emptyset)$ with $|V_{I_k}| = k$ denote a graph with k vertices and no edges. Given the input of Independent Set, i.e. a graph $G = (V, E)$ and some $k \in \mathbb{N}$ we transform it into (I_k, G), k, as input for Induced Subgraph Bisimulation.

Now, we claim that G has an independent set of size at least k iff there is some $V' \subseteq V$ with $|V'| \geq k$ and $(V', E \cap (V' \times V')) \underline{\leftrightarrow}_{total} I_k$.

From left to right, assume that there is some $S \subseteq V$ with $|S| = k$, and for all $v, v' \in S$, $(v, v') \notin E$. Now, any bijection between S and V_{I_k} is a total bisimulation between $G' = (S, E \cap (S \times S))$ and I_k, since $E \cap (S \times S) = \emptyset$ and $|S| = |V_{I_k}|$.

For the other direction, assume that there is some $V' \subseteq V$ with $|V'| = k$ such that for $G' = (V', E' = E \cap (V' \times V'))$ we have that $G' \underline{\leftrightarrow}_{total} I_k$. Thus, there is some total bisimulation Z between G' and I_k. Now, we claim that V' is an independent set of G of size k. Let $v, v' \in V'$. Suppose that $(v, v') \in E$. Then since G' is an induced subgraph, we also have that $(v, v') \in E'$. Since Z is a total bisimulation, there has to be some $w \in I_k$ with $(v, w) \in Z$ and some w' with $(w, w') \in E_{I_k}$ and $(v', w') \in Z$. But this is a contradiction with $E_{I_k} = \emptyset$. Thus, V' is an independent set of size k of G. The reduction can clearly be computed in polynomial time. This concludes the proof. □

Proposition 3.18: If for graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ it holds that E_1 and E_2 are reflexive, transitive and symmetric, then the problem of induced subgraph bisimulation for G_1 and G_2 can be solved in linear time.

Proof. In this proof, we will use the fact that $G_1 = (V_1, E_1) \underline{\leftrightarrow}_{total} G_2 = (V_2, E_2)$ if and only if it is the case that $V_1 = \emptyset$ iff $V_2 = \emptyset$. Let us prove this. From left to right, assume that $G_1 = (V_1, E_1) \underline{\leftrightarrow}_{total} G_2 = (V_2, E_2)$. Then since we have a total bisimulation, it must be the case that either $V_1 = V_2 = \emptyset$ or $V_1 \neq \emptyset \neq V_2$.

For the other direction, assume that $V_1 = \emptyset$ iff $V_2 = \emptyset$. Now, we show that in this case, $V_1 \times V_2$ is a total bisimulation between G_1 and G_2. If $V_1 = V_2 = \emptyset$, we are done. So, consider the case where $V_1 \neq \emptyset \neq V_2$. Let $(v_1, v_2) \in V_1 \times V_2$, and assume that $(v_1, v_1') \in E_1$ for some $v_1' \in V_1$. Since E_2 is reflexive, we know that there is some $v_2' \in V_2$ such that $(v_2, v_2') \in E_2$. Of course $(v_1', v_2') \in V_1 \times V_2$. The back condition is analogous. Since $V_1 \times V_2$ is total, we thus have $G_1 \underline{\leftrightarrow}_{total} G_2$. Hence, $G_1 = (V_1, E_1) \underline{\leftrightarrow}_{total} G_2 = (V_2, E_2)$ if and only if it is the case that $V_1 = \emptyset$ iff $V_2 = \emptyset$.

Therefore, for solving the induced subgraph bisimulation problem for input G_1 and G_2 with E_1 and E_2 being reflexive, transitive and symmetric and $k \in \mathbb{N}$, all we need to do is to go through the input once and check whether $V_1 = \emptyset$ iff $V_2 = \emptyset$, and whether $|V_2| \geq k$. If the answer to both is *yes* then we know that $G_1 \underline{\leftrightarrow}_{total} G_2$ and since $|V_2| \geq k$, we answer *yes*, otherwise *no*. □

Proposition 3.20: Local submodel bisimulation for single agent pointed epistemic models is in P.

Proof. Given the input of the problem, i.e. a pointed epistemic model $\mathcal{M}_1, w$ with $\mathcal{M}_1 = (W_1, \sim_1, V_1)$, and $w \in W_1$ and an epistemic model $\mathcal{M}_2 = (W_2, \sim_2, V_2)$, we run the following procedure.

1. For all $[w_2] \in W_2/\sim_2$ do the following:
   (a) Initialize the set $Z := \emptyset$.
   (b) for all $w' \in [w]$ do the following
      i. For all $w_2' \in [w_2]$ check if for all $p \in$ **Prop** it holds that $w' \in V_1(p)$ iff $w_2' \in V_2(p)$. If this is the case, set $Z := Z \cup (w', w_2')$.
      ii. if there is no such w_2', continue with 1, otherwise we return Z and we stop.
2. In case we didn't stop at 1(b)ii, we can stop now, and return *no*.

In the worst case, this takes $|\mathcal{M}_1| \cdot |\mathcal{M}_2|$ steps.

If the procedure has stopped at 2, there is no bisimulation with the required properties. To see this, note that if we stopped in 2, this means that there was no $[w_2] \in W_2/\sim_2$ such that for every state in $[w]$ there is one in $[w_2]$ in which exactly the same propositional letters are true. Thus, since we were looking for a bisimulation that is also defined for the state w, such a bisimulation cannot exist.

If the algorithm returned a relation Z, this is indeed a bisimulation between $\mathcal{M}_1$ and the submodel $\mathcal{M}_2'$ of $\mathcal{M}_2$ where $\mathcal{M}_2' = (W_2', \sim_2', V_2')$, where

$$W_2' = \{w_2 \in W_2 \mid \text{there is some } w_1 \in [w] \text{ such that } (w_1, w_2) \in Z\}$$

and $\sim_2'$ and V_2' are the usual restrictions of $\sim_2$ and V_2 to W_2. This follows from the following two facts: First, for all pairs in Z it holds that both states satisfy exactly the same proposition letters. Second, since Z is total both on $[w]$ and on W_2' and all the states in $[w]$ are connected to each other by $\sim_1$ and all states in W_2' are connected to each other by $\sim_2'$, both the *forth* and *back* conditions are satisfied. This concludes the proof. □

Theorem 3.22: Total submodel bisimulation for single agent epistemic models is in P.

First we introduce some notation used in the proof.

Notation A.1. *Let $\mathcal{M} = (W, \sim, V)$ be a single agent epistemic model. For the valuation function $V : \mathbf{Prop} \to W$, we define $\hat{V} : W \to 2^{\mathbf{Prop}}$, with $w \mapsto \{p \in \mathbf{Prop} \mid w \in V(p)\}$. Abusing notation, for $X \subseteq W$ we sometimes write $\hat{V}(X)$ to denote $\{\hat{V}(w) \mid w \in X\}$. For $w \in W$, $[w] = \{w' \in W \mid w \sim w'\}$ denotes the equivalence class of w under $\sim$. $W/\sim$ denotes the set of all equivalence classes of W for the relation $\sim$.*

Definition A.2. Given a single agent epistemic model $\mathcal{M} = (W, \sim, V)$, $\mathcal{M}^{min\_cells}$ denote a model obtained from $\mathcal{M}$ by the following procedure:

1. Initialize X with $X := W/\sim$.
2. Go through all the pairs in $X \times X$.
   (a) When you find $([w], [w'])$ with $[w] \neq [w']$ such that $\hat{V}([w]) = \hat{V}([w'])$, continue at 2 with $X := X - [w']$.
   (b) Otherwise, stop and return the model $\mathcal{M}^{min\_cells} := (\bigcup X, \sim', V')$, where $\sim'$ and V' are the usual restrictions of $\sim$ and V to $\bigcup X$.

Fact A.3. *With input $\mathcal{M} = (W, \sim, V)$, the procedure in Definition A.2 runs in time polynomial in $|\mathcal{M} = (W, \sim, V)|$.*

Proof. Follows from the fact that the cardinality of $W/\sim$ is bounded by $|W|$; we only enter step 2 at most $|W|$ times, and each time do at most $|W|^2$ comparisons. □

Fact A.4. *The answer to total submodel bisimulation for single agent epistemic models (Decision Problem 3.21) with input $\mathcal{M}_1 = (W_1, \sim_1, V_1), \mathcal{M}_2 = (W_2, \sim_2, V_2)$ is* yes *iff it is with input ${\mathcal{M}_1}^{min\_cells} = (W_1, \sim_1, V_1), \mathcal{M}_2 = (W_2, \sim_2, V_2)$.*

Proof. From left to right, we just need to restrict the bisimulation to the states of ${\mathcal{M}_1}^{min\_cells}$. For the other direction, we start with the given bisimulation and then extend it as follows. For the states in a cell $[w']$ which was removed during the construction of ${\mathcal{M}_1}^{min\_cells}$, can be mapped to the ones of a cell $[w]$ in ${\mathcal{M}_1}^{min\_cells}$ with the same valuation. □

Proof. By Fact A.3 and Fact A.4, transforming $\mathcal{M}_1$ into ${\mathcal{M}_1}^{min\_cells}$ can be done in polynomial time. Thus, w.l.o.g. we can assume that $\mathcal{M}_1$ is already of the right shape; i.e. $\mathcal{M}_1 = {\mathcal{M}_1}^{min\_cells}$. Given the two models as input, we construct a bipartite graph $G = ((W_1/\sim_1, W_2/\sim_2), E)$ where E is defined as follows.

$$([w_1], [w_2]) \in E \text{ iff } \hat{V}_1([w_1]) \subseteq \hat{V}_2([w_2]).$$

Claim A.5. *There is a submodel $\mathcal{M}_2'$ of $\mathcal{M}_2$ such that $\mathcal{M}_1 \underline{\leftrightarrow}_{total} \mathcal{M}_2'$ iff G has a matching of size $|W_1/\sim_1|$.*

Proof. From left to right, assume that there is a submodel $\mathcal{M}_2' = (W_2', \sim_2', V_2')$ of $\mathcal{M}_2$ such that $\mathcal{M}_1 \underline{\leftrightarrow}_{total} \mathcal{M}_2'$. Let Z be such a total bisimulation.

Note that since we assumed that $\mathcal{M}_1 = \mathcal{M}^{min\_cells}$ the following holds:

1. For all $([w_1],[w_2]) \in W_1/\sim_1 \times W_2/\sim_2$ it is the case that whenever $Z \cap ([w_1] \times [w_2]) \neq \emptyset$, then for all $[w_1'] \in W_1/\sim_1$ such that $[w_1'] \neq [w_1], Z \cap ([w_1'] \times [w_2]) = \emptyset$.

Thus, the members of different equivalence classes in $W_1/\sim_1$ are mapped by Z to into different equivalence classes of $W_2/\sim_2$.

Now, we construct $\dot{E} \subseteq E$ as follows.

$$([w_1],[w_2]) \in \dot{E} \text{ iff } ([w_1],[w_2]) \in E \text{ and } ([w_1] \times [w_2]) \cap Z \neq \emptyset.$$

Then $|\dot{E}| \geq |W_1/\sim_1|$ because of the definitions E and $\dot{E}$ and the fact that Z is a bisimulation that is total on W_1. Now, if $|\dot{E}| = |W_1/\sim_1|$ then we are done since by definition of $\dot{E}$, for each $[w_1] \in W_1/\sim_1$ there is some $[w_2] \in W_2/\sim_2$ such that $([w_1],[w_2]) \in \dot{E}$. Then it follows from 1, that $\dot{E}$ is indeed a matching.

If $|\dot{E} > |W_1/\sim_1|$ then we can transform $\dot{E}$ into a matching E' of size $W_1/\sim_1|$: For each $[w_1] \in W_1/\sim_1$, we pick *one* $[w_2] \in W_2/\sim_2$ such that $([w_1],[w_2]) \in \dot{E}$ and put it into E' (note that such a $[w_2]$ always exists because by definition of $\dot{E}$, for each $[w_1] \in W_1/\sim_1$ there is some $[w_2] \in W_2/\sim_2$ such that $([w_1],[w_2]) \in \dot{E}$; moreover because of 1 all the $[w_2] \in W_2/\sim_2$ that we pick will be different). Then the resulting $E' \subseteq \dot{E} \subseteq E \subseteq (W_1/\sim_1 \times W_2/\sim_2)$ is a matching of G of size $|W_1/\sim_1|$. Thus, we have shown that if there is a submodel $\mathcal{M}_2'$ of $\mathcal{M}_2$ such that $\mathcal{M}_1 \underline{\leftrightarrow}_{total} \mathcal{M}_2'$ then G has a matching of size $|W_1/\sim_1|$.

For the other direction, assume that G has a matching $E' \subseteq E$ with $|E'| = |W_1/\sim_1|$. Then, recalling the definition of E, it follows that for all $[w] \in W_1/\sim$ there is some $[w'] \in W_2/\sim_2$ such that $([w],[w']) \in E'$ and thus $\hat{V}_1([w]) \subseteq \hat{V}_2([w'])$.

Let us define the following submodel $\mathcal{M}_2'$ of $\mathcal{M}_2$. $\mathcal{M}_2' = (W_2', \sim_2', V_2')$, where

$$W_2' = \{w_2 \in W_2 \mid \text{there is some } w \in W_1 \text{ such that } \hat{V}_1(w) = \hat{V}_2(w_2) \text{ and } ([w],[w_2]) \in E'\}$$

and $\sim_2'$ and V_2' are the usual restrictions of $\sim_2$ and V_2 to W_2'.

Now, we define a relation $Z \subseteq W_1 \times W_2'$, which we then show to be a total bisimulation between $\mathcal{M}_1$ and $\mathcal{M}_2'$

$$(w_1, w_2) \in Z \text{ iff } \hat{V}(w_1) = \hat{V}_2(w_2) \text{ and } ([w_1],[w_2]) \in E'.$$

Next, let us show that Z is indeed a bisimulation.

Let $(w_1, w_2) \in Z$. Then, by definition of Z, for every propositional letter p, $w_1 \in V_1(p)$ iff $w_2 \in V_2(p)$. Next, we check the *forth* condition. Let $w_1 \sim_1 w_1'$ for some $w_1' \in W_1$. Then since $(w_1, w_2) \in Z$, and thus $([w_1],[w_2]) \in E'$, there has to be some $w_2' \in [w_2]$ such that $\hat{V}_2(w_2') = \hat{V}_1(w_1')$. Then since $[w_1'] = [w_1]$ and $[w_2'] = [w_2]$, $([w_1'],[w_2']) \in E'$. Then $w_2' \in W_2'$, and $(w_1', w_2') \in Z$.

For the *back* condition, let $w_2 \sim_2 w_2'$, for some $w_2' \in W_2'$. Then by definition of W_2', there is some $w \in W_1$ such that $\hat{V}_1(w) = \hat{V}_2(w_2')$ and $([w],[w_2']) \in E'$. Thus, it follows that $(w, w_2') \in Z$. Now, we still have to show that $w_1 \sim_1 w$. As the following hold: $([w],[w_2']) \in E', [w_2] = [w_2']$, $([w],[w_2]) \in E'$ (because $(w_1, w_2) \in Z$) and E' is a matching, it follows that $[w] = [w_1]$. Thus, $w_1 \sim_1 w$.

Hence, we conclude that Z is a bisimulation. It remains to show that Z is indeed total.

Let $w_1 \in W_1$. Since E' is a matching of size $W_1/\sim_1$, there is some $[w_2] \in W_2/\sim_2$ such that $([w_1],[w_2]) \in E'$. Thus, there is some $w_2' \in [w_2]$ such that $\hat{V}_1(w_1) = \hat{V}_2(w_2')$. This means that $w_2' \in W_2'$ and $(w_1, w_2') \in Z$. So Z is total on W_1.

Let $w_2 \in W_2'$. By definition of W_2', there is some $w \in W_1$ such that $\hat{V}_1(w) = \hat{V}_2(w_2)$ and $([w],[w_2]) \in E'$. Thus, by definition of Z, $(w, w_2) \in Z$. Therefore, Z is indeed a total bisimulation between $\mathcal{M}_1$ and $\mathcal{M}_2'$. This concludes the proof of Claim A.5. □

Hence, given two models, we can transform the first one using the polynomial procedure of Definition A.2 and then we construct the graph G, which can be done in polynomial time as well. Finally, we use a polynomial algorithm to check if G has a matching of size ${M_1}^{min\_cells}$. If the answer is yes, we return *yes*, otherwise *no*. This concludes the proof of Theorem 3.22. □

Proposition 3.25: Induced subgraph simulation is NP-complete.

Proof. Showing that the problem is in NP is straightforward. Hardness is shown by reduction from Independent Set. First of all, let $I_k = (V_{I_k}, E_{I_k} = \emptyset)$ with $|V_{I_k}| = k$ denote a graph with k vertices and no edges. Given the input of Independent Set, i.e. a graph $G = (V, E)$ and some $k \in \mathbb{N}$ we transform it into (I_k, G), k, as input for Induced Subgraph Simulation.

Now, we claim that G has an independent set of size at least k iff there is some $V' \subseteq V$ with $|V'| \geq k$ and $(V', E \cap (V' \times V')) \sqsubseteq_{total} I_k$.

From left to right, assume that there is some $S \subseteq V$ with $|S| = k$, and for all $v, v' \in S$, $(v, v') \notin E$. Now, any bijection between S and V_{I_k} is a total simulation (and in fact an isomorphism) between $G' = (S, E \cap (S \times S))$ and I_k, since $E \cap (S \times S) = \emptyset$ and $|S| = |V_{I_k}|$.

For the other direction, assume that there is some $V' \subseteq V$ with $|V'| = k$ such that for $G' = (V', E' = E \cap (V' \times V'))$ we have that $G' \sqsubseteq_{total} I_k$. Thus, there is some total simulation Z between G' and I_k. Now, we claim that V' is an independent set of G of size k. Let $v, v' \in V'$. Suppose that $(v, v') \in E$. Then since G' is an induced subgraph, we also have that $(v, v') \in E'$. Since Z is a total simulation, there has to be some $w \in I_k$ with $(v, w) \in Z$ and some w' with $(w, w') \in E_{I_k}$ and $(v', w') \in Z$. But this is a contradiction with $E_{I_k} = \emptyset$. Thus, V' is an independent set of size k of G. The reduction can clearly be computed in polynomial time. This concludes the proof. □

Proposition 3.27: Induced subgraph simulation equivalence is NP-complete.

Proof. For showing that the problem is in NP, note that we can use a simulation equivalence algorithm as provided in [14]. Hardness can again be shown by reduction from Independent Set. Given the input for Independent Set, i.e. a graph $G = (V, E)$ and some $k \in \mathbb{N}$, we transform it into two graphs $I_k = (V_{I_k} = \{v_1, \dots v_k\}, E_{I_k} = \emptyset)$ and G, and we keep the $k \in \mathbb{N}$. This can be done in polynomial time.

Now, we claim that G has an independent set of size k iff there is an induced subgraph of G with k vertices that is similar to I_k. From left to right assume that G has such an independent set S with $S \subseteq V$, $|S| = k$ and $E \cap S \times S = \emptyset$. Then $(S, \emptyset)$ is isomorphic to I_k since both have k vertices and no edges. Thus, they are also simulation equivalent.

For the other direction, assume that there is an induced subgraph $G' = (V', E')$ with $V' \subseteq V$, $|V'| = k$ and $E' = (V' \times V') \cap E$ such that G' is simulation equivalent to I_k. Suppose that there are $v, v' \in V'$ such that $(v, v') \in E$. Since G' is an induced subgraph, it must be the case that $(v, v') \in E'$, but since I_k simulates G', this leads to a contradiction since I_k does not have any edges.

This concludes the proof. □

The Ditmarsch Tale of Wonders — Dynamics of Lying

Hans van Ditmarsch

University of Sevilla, Sevilla, Spain, hvd@us.es

1 Introduction

My favourite of Grimm's fairytales is 'Hans im Glück' (Hans in luck). A close second comes 'The Ditmarsch Tale of Wonders'. In German this is called a 'Lügenmärchen', a 'Liar's Tale'. It contains the passage "A crab was chasing a hare which was running away at full speed; and high up on the roof lay a cow which had climbed up there. In that country the flies are as big as the goats are here." These are very obvious lies. Nobody considers it possible that this is true. Crabs are reputedly slow, hares are reputedly fast.

In the real world, if you lie, sometimes other people believe you and sometimes they don't. When can you get away with a lie? Consider the well-known consecutive numbers riddle (see [10], and the Appendix), where Anne has 2 and Bill has 3, and they only know that their natural numbers are one apart. Initially, Anne is uncertain between Bill having 3 or 1, and Bill is uncertain between Anne having 2 or 4. So both Anne and Bill do not initially know their number. Suppose Anne says to Bill: "I know your number." Anne is lying. Bill does not consider it possible that Anne knows his number, so he tells Anne that she is lying. However, Anne did not know that Bill would not believe her. She considered it possible that Bill had 1, in which case Bill would have considered it possible that Anne was telling the truth, and would then have drawn the incorrect conclusion that Anne had 0. I.e., if you are still following us... It seems not so clear how this should be formalized in a logic interpreted on epistemic modal structures, and this is the topic of our paper.

What is a lie? Let p be a Boolean proposition. You lie that p if you believe that $\neg p$ while you say that p and with the intention that the addressee believes p. This definition seems standard since Augustine [12]. A *believed lie* therefore is one that, when told, is believed by the addressee to be truthful. We abstract from the intentional aspect and model the believed lie. (Similarly, in AGM belief revision, we incorporating new information, abstracting from the process that made it acceptable.)

What are the modal preconditions and postconditions of a lie? Let i be the speaker (assumed female) and let j be the addressee (assumed male). Then the precondition of 'i is lying that p to j' is $B_i\neg p$, and the postcondition is $B_j p$. Also, the precondition should be preserved. More refined preconditions are conceivable, e.g., that the addressee consider it possible that the lie is true, or believes that the speaker knows the truth about p. Those are plausible additional conditions

rather than rock-bottom requirements. Concerning the postcondition: the liar does not merely intend the speaker to believe p, but also wants him to believe that the speaker believes p. It is obvious that the postcondition should not be merely $B_j p$, but $B_j C_{ij} p$: after a lie that p, the addressee believes that speaker has shared knowledge with him about p. The modellings we propose satisfy this, but we restrict our discussion to logics without common knowledge.

In a dynamic setting, what we want, so far, is: *Lying that p is the epistemic action transforming information states satisfying $B_i \neg p$ into information states satisfying $B_j p$ and preserving $B_i \neg p$.* We need to make a choice concerning: information state, epistemic action, epistemic modal operator, and, finally, how to generalize lying about Booleans to lying about modal formulae. As *information state* we propose a multi-agent Kripke model. We consider one agent lying to one other agent; or one agent lying to the group of all other agents. A Kripke model transformation calls for a *dynamic modality*. As lying is the opposite of telling the truth, a variation of public announcement logic seems obvious. First, we model lying announcements by an external observer, comparable to truthful announcements by that observer. Then, we model lying of agent i to agent j, where both agents are modelled in the Kripke model.

Clearly, our epistemic modality cannot be knowledge. If the liar correctly believes that p is false and lies that p after which the addressee j believes p, then j holds a false belief. In AI, the next best thing to knowledge is belief, i.e., $KD45$ belief. We will *aim* for that, and therefore have to address the problem that consistency of belief is not necessarily preserved after update.

When generalizing from 'lying that p' to 'lying that φ' for epistemic propositions, we have to change to postcondition. The addressee j believes that φ is true when announced. It may no longer be true after the liar and the addressee have processed the information contained in the lie. We should require that j believes that φ *was* true before the lie, not that it still is true after the lie. This is because of Moorean phenomena: if I am lying to you, agent j, that $p \wedge \neg B_i p$, after the lie you believe p, not that you are ignorant about it. Lying in the consecutive number riddle is of that kind.

We conclude this introduction with an overview of the literature. Lying has been a thriving topic in the philosophical community for a long, long time [15, 5, 11, 12]— indeed, almost any analysis starts with quoting Augustine on lying (check!). The precision of the belief preconditions and postconditions is illuminating. E.g., emphasis that the addressee should not merely believe the lie but believe it to be believed by the speaker. Indeed, ... and even believed to be commonly believed, would the modal logician say. Interesting scenarios involving eavesdroppers (can you lie to an eavesdropper?) clearly are relevant for logic and multi-agent system design, and also claims that you can only lie if you really *say* something: an omission is not a lie [12]. Wrong, says the computer scientist: if the protocol is common knowledge, you can lie by *not* acting when you should; say, by not stepping forward in the muddy children problem although you know that you are muddy. The philosophical literature also clearly distinguishes between false propositions and propositions believed to be false but in fact true, so

that when you lie about them, in fact you tell the truth. Interesting Gettier-like scenarios are discussed. Also, much is said on the morality of lying and on its intentional aspect. As said, we abstract from the intentional aspect of lying. We also abstract from its moral aspect.

In the modal logical community, papers on lying include [2, 16, 4, 19, 14, 9, 20]. They (almost) all model lying as an epistemic action, inducing a transformation of an epistemic model. Lying has been discussed by Baltag *et al.* from the inception of BMS onward [2, 4]; the latter also discusses lying in logics with knowledge and plausible belief (AGM belief revision with lying, so to speak), as does [19]. In [20] (dating from 2007) the conscious update in [7] is applied to model lying by an external observer to the public (of agents). The recent [14] gives a modal logic of lying, bluffing and (after all) intentions—they do not model lying as an epistemic action, and do not seem to realize the trouble this gets you into when you lie about a Moore-sentence. In [16, 9] the unbelievable lie is considered; this is the issue consistency preservation in $KD45$ updates.

2 Logical preliminaries

The logic of *lying* public announcements complements the well-known logic of *truthful* public announcements [13, 3], that is an extension of multi-agent epistemic logic. Its language, structures, and semantics are as follows.

Given a finite set of agents N and a countable set of propositional variables P, the *language* $\mathcal{L}(!)$ *of public announcement logic* is inductively defined as

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid B_i\varphi \mid [!\varphi]\psi$$

where $p \in P$, $i \in N$. For $B_i\varphi$, read 'agent i believes formula φ'. For $[!\varphi]\psi$, read 'after truthful announcement of φ, formula ψ (is true)'.

An *epistemic model* $M = \langle S, R, V\rangle$ consists of a *domain* S of *states* (or 'worlds'), an *accessibility function* $R : N \to \mathcal{P}(S \times S)$, where each R_i is an accessibility relation, and a *valuation* $V : P \to \mathcal{P}(S)$. For $s \in S$, (M, s) is an *epistemic state*, also known as a pointed Kripke model. The class of models where all accessibility relations are serial, transitive and euclidean is called $\mathcal{K}D45$. Without any restrictions we call the model class $\mathcal{K}$.

Assume an epistemic model $M = \langle S, \sim, V\rangle$.

$$\begin{array}{ll}
M, s \models p & \text{iff } s \in V_p \\
M, s \models \neg\varphi & \text{iff } M, s \not\models \varphi \\
M, s \models \varphi \wedge \psi & \text{iff } M, s \models \varphi \text{ and } M, s \models \psi \\
M, s \models B_i\varphi & \text{iff for all } \ t \in S : R_i(s, t) \text{ implies } M, t \models \varphi \\
M, s \models [!\varphi]\psi & \text{iff } M, s \models \varphi \text{ implies } M|\varphi, s \models \psi
\end{array}$$

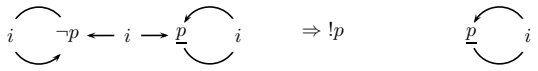
where the model restriction $M|\varphi = \langle S', R', V'\rangle$ is defined as $S' = \{s' \in S \mid M, s' \models \varphi\}$ $(= [\![\varphi]\!]_M)$, $R'_i = R_i \cap (S' \times S')$ and $V'(p) = V(p) \cap S'$. A complete proof system for this logic (for class $\mathcal{S}5$, originally) is presented in [13]. The interaction between announcement and belief is

$$[!\varphi]B_i\psi \leftrightarrow \varphi \to B_i[!\varphi]\psi$$

The interaction between announcement and other operators we assume known. It changes predictably in the other logics we present. The class $\mathcal{K}D45$ is not closed under public announcements: given $\neg p \wedge B_i p$, and new information $!\neg p$, agent i's accessibility relation becomes empty: she believes everything.

In the coming sections, we will only vary the dynamic part of the logic.

For an example of the semantics of public announcement, consider a situation wherein the agent is uncertain about p, and receives the information that p. In view of the continuation, we draw all access. A state has been given the value of the atom there as its name. The actual state is underlined.



3 Logic of truthful and lying public announcements

We expand the language of truthful public announcement logic with another inductive construct $[¡\varphi]\psi$, for 'after lying public announcement of φ, formula ψ (is true)'; in short 'after the lie that φ, ψ'. This is the language $\mathcal{L}(!, ¡)$.

Truthful public announcement logic is the logic to model the revelations of a benevolent god, taken as the truth without questioning. The announcing agent is not modelled in public announcement logic, but only the effect of her announcements on the audience, the set of all agents. Consider a *false* public announcement, made by a malevolent entity, the devil. Everything he says is false. Everything is a lie. Not surprisingly, god and the devil are inseparable and should be modelled simultaneously. This is as in religion.

An alternative for the semantics of public announcements is the semantics of *conscious updates* [7]. (In fact, [7] and [13] were independently proposed.) When announcing φ, instead of eliminating states where φ does not hold, one eliminates *access* to states where φ does not hold. The effect of the announcement of φ is that only states where φ is true are accessible for the agents. It is not a model restricting transformation but an arrow restricting transformation. We see this as the logic of *believed* public announcements. There is no relation between the agent accepting new information and the truth of that information.

In [20], this believed announcement of φ is called manipulative update with φ. The original proposal there is to view this as non-deterministic choice $!\varphi \cup ¡\varphi$ between truthful announcement and lying announcement, with the following semantics

$$M, s \models [!\varphi]\psi \text{ iff } M, s \models \varphi \text{ implies } M^\varphi, s \models \psi$$
$$M, s \models [¡\varphi]\psi \text{ iff } M, s \models \neg\varphi \text{ implies } M^\varphi, s \models \psi$$

where epistemic model M^φ is as M except that (with S the domain of M)

$$R_i^\varphi \;:=\; R_i \cap\; (S \times [\![\varphi]\!]_M).$$

We can keep writing $!\varphi$ for 'arrow elimininting' truthful announcement without risk of ambiguity with 'state eliminating' truthful announcement, because on the

states s where φ is true in M we have that

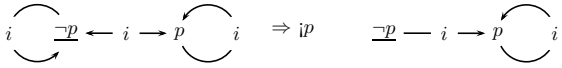
$$(M^{\varphi}, s) \leftrightarrows (M|\varphi, s).$$

The axioms for truthful announcement remain what they were and the axiom for the reduction of belief after lying is

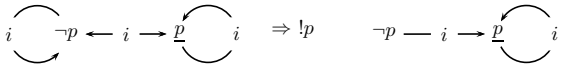
$$[¡\varphi]B_i\psi \leftrightarrow \neg\varphi \to B_i[!\varphi]\psi.$$

After the lying announcement that φ, agent i believes that ψ, if and only if, on condition that φ is false, agent i believes that ψ *after truthful announcement* that φ. To the credulous person who believes the lie, the lie appears to be the truth. This proposal to model lying has been investigated in detail in [20].

For an example, we show the effect of truthful and lying announcement of p in the model with uncertainty about p. The actual state must be different in these models: when lying, p is (believed) false, and when being truthful, p is (believed) true. For lying we get



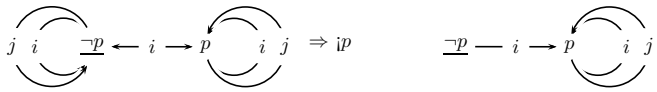
whereas for truthtelling we get



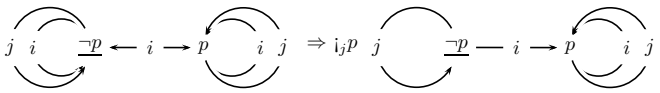
4 Agent announcement logic

In the logic of lying and truthful public announcements, the outside observer is implicit. Therefore, it is also implicit that she believes that the announcement is false or true. In multi-agent epistemic logic, it is common to formalize 'agent i truthfully announces φ' as 'the outside observer truthfully announces $B_i\varphi$'. However, 'agent i lies that φ' cannot be modelled as 'the outside observer lies that $B_i\varphi$'.

For a counterexample, consider an epistemic state where i does not know whether p, j knows whether p, and p is true. Agent j is in the position to tell i the truth about p. The reader can check that a truthful public announcement of B_ip indeed simulates that i truthfully announces p. Now suppose p is false, and that j lies that p. A lying public announcement of B_ip does not result in the desired information state, because this makes agent j believe his own lie. In fact, as he already knew $\neg p$, this makes j's beliefs inconsistent.



Instead, a lie from j to i should have the following effect:



After this lie we have that j still believes that $\neg p$, i believes that p, and i believes that i and j have common belief of p. We satisfied the requirements of a truthful and lying agent announcement.

Apart from lying and telling the truth, another form of announcement is *bluffing*. You are bluffing that φ, if you say that φ but are uncertain about φ. The precondition for bluffing is therefore $\neg(B_i\varphi \vee B_i\neg\varphi)$. If belief is explicit there are always three preconditions for announcing φ: $B_i\varphi$, $B_i\neg\varphi$, and $\neg(B_i\varphi \vee B_i\neg\varphi)$, the preconditions for truthtelling, lying, and bluffing. If belief is implicit there are only two preconditions for announcing φ: φ and $\neg\varphi$, for truthtelling and lying. God and the devil are omniscient, and bluffing is therefore inconceivable for them. More prosaically, they can be considered an agent with an accessibility relation that is the identity on the model.

The logical language $\mathcal{L}(!_j, ¡_j, ¡!_j)$ of *agent announcement logic* is defined by adding inductive constructs

$$[!_j\varphi]\psi \mid [¡_j\varphi]\psi \mid [¡!_j\varphi]\psi$$

to the epistemic language, for, respectively, j truthfully announces φ, j is lying that φ, and j is bluffing that φ; where agent j addresses all other agents i.

The preconditions of these three types of announcement are all different, but their effect on the speaker and on the listeners are the same: States where φ was believed by j, if any (none, if j is lying), remain accessible for j (i); states where $\neg\varphi$ was believed by j, if any (none, if j is truthful), remain accessible for j (ii); states where φ was believed by i, if any (if there are none, i will 'go mad'), remain accessible for i (iii); and states where $\neg\varphi$ was believed by i, if any, are no longer accessible for i (iv). This is embodied by the following semantics.

$$\begin{array}{l} M, s \models [!_j\varphi]\psi \text{ iff } M, s \models B_i\varphi \text{ implies } M_j^\varphi, s \models \psi \\ M, s \models [¡_j\varphi]\psi \text{ iff } M, s \models B_i\neg\varphi \text{ implies } M_j^\varphi, s \models \psi \\ M, s \models [¡!_j\varphi]\psi \text{ iff } M, s \models \neg(B_i\varphi \vee B_i\neg\varphi) \text{ implies } M_j^\varphi, s \models \psi \end{array}$$

where M_j^φ is as M except that a new accessibility relation R' is defined as (S is the domain of M, and $i \neq j$)

$$\begin{array}{l} R'_j := R_j \\ R'_i := R_i \cap (S \times [\![\varphi]\!]_M) \end{array}$$

If φ is believed by j in state s in M we have that

$$(M_j^\varphi, s) \leftrightarrows (M|B_j\varphi, s).$$

This justifies that there is no difference between agent j truthfully announcing that φ and the truthful public announcement of $B_j\varphi$.

The principles for j lying to i are as follows:

$$[¡_j\varphi]B_i\psi \leftrightarrow B_j\neg\varphi \rightarrow B_i[!_j\varphi]\psi$$
$$[¡_j\varphi]B_j\psi \leftrightarrow B_j\neg\varphi \rightarrow B_j[¡_j\varphi]\psi$$

In other words, the liar knows that he is lying, but the dupe he is lying to, believes that the liar is telling the truth. The principles for truthtelling and bluffing are similar, but with (the obvious) different conditions on the right hand side. With these principles, the logic is completely axiomatizated. (This is, because it is a logic for a specific action model. See the next section.)

The Appendix illustrates agent lying in the consecutive numbers riddle. In the continuation we discuss consequences and variations of public lying and agent lying: an action model perspective, how to address the issue of unbelievable lies, lying about beliefs, and lying and plausible beliefs.

5 Action models and lying

Whether I am telling the truth to you, am lying, or am bluffing, to you it all appears as the same annoucement. A familiar way to formalize uncertainty about actions are *action models* [3]. We can view truthful and lying public announcement as the two points of an action model, and we can also view truthful, lying and bluffing agent announcement as the three different points in another action model.

An *action model* $\mathsf{M} = \langle \mathsf{S}, R, \mathsf{pre}\rangle$ consists of a *domain* S of *actions*, an *accessibility function* $R : N \rightarrow \mathcal{P}(\mathsf{S} \times \mathsf{S})$, where each R_i is an accessibility relation, and a *precondition function* $\mathsf{pre} : \mathsf{S} \rightarrow \mathcal{L}$, where $\mathcal{L}$ is a logical language. A pointed action model is an *epistemic action*. Performing an epistemic action in an epistemic state means computing their restricted modal product—restricted to state/action pairs (t, t) such that $M, t \models \mathsf{pre}(\mathsf{t})$. With such an epistemic action (M, s) we can associate a dynamic modal operator $[\mathsf{M}, \mathsf{s}]$ in the usual way.

The action model M' for truthful and lying public announcement consists of two actions suggestively named ! and ¡ with preconditions φ and $\neg\varphi$ in $\mathcal{L}(!, ¡)$, respectively, and for all agents only action ! is accessible. *Truthful public announcement of* φ is the epistemic action $(\mathsf{M}', !)$. Given that $\mathsf{pre}(!) = \varphi$, $[!\varphi]\psi$ corresponds to $[\mathsf{M}', !]\psi$. *Lying that* φ is the epistemic action $(\mathsf{M}', ¡)$.

The action model M'' for agent announcement consists of three actions named $¡!_j$, $!_j$, and $¡_j$ with preconditions $\neg(B_j\varphi \vee B_j\neg\varphi)$, $B_j\varphi$, and $B_j\neg\varphi$, respectively (all in $\mathcal{L}(!_i, ¡_i, ¡!_i)$). The announcing agent j has identity access on the action model and to the other agents only action $!_j$ is accessible. Agent j truthfully announcing φ to all other i is is the epistemic action $(\mathsf{M}'', !_j)$—with precondition $B_j\varphi$, therefore—and similarly lying and bluffing are the action models $(\mathsf{M}'', ¡_j)$ and $(\mathsf{M}'', ¡!_j)$. Action models M' and M'' are depicted in Figure 1.

The action model representations validate the axioms for announcement and belief, for all versions shown; and they justify that these axioms form part of

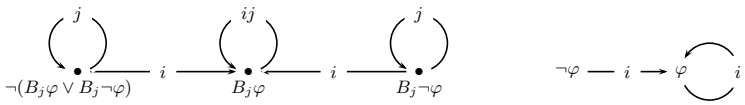


Fig. 1. Action models for lying, truthtelling and bluffing

complete axiomatizations.[1] These axioms are simply instantations of a more general axiom for an epistemic action followed by a belief. Note that M' and M'' are both in class $\mathcal{KD}45$ but nevertheless, as we have seen, executing a $\mathcal{KD}45$ epistemic action in a $\mathcal{KD}45$ epistemic state does not guarantee a $\mathcal{KD}45$ updated epistemic state.

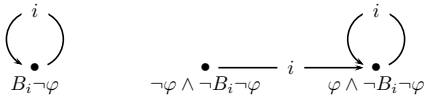
5.1 Unbelievable lies

The class of $\mathcal{S}5$ epistemic models is closed under update with $\mathcal{S}5$ epistemic actions, such as truthful public announcements, but the class of $\mathcal{KD}45$ models is *not* closed under update with $\mathcal{KD}45$ epistemic actions such as a lying public announcement. (It is not even closed under update with correct information.) The problem is that beliefs may be mistaken and that new information may be incorrect. Either way, if you tell me that p but I already believe the opposite, then I 'go mad' if I accept the new information without discarding the old information. My accessibility relation has become empty: I lose the D in $\mathcal{KD}45$.

$\mathcal{KD}45$-preserving updates have been investigated in [16, 1, 9]. Aucher [1] defines a language fragment that makes you go mad ('crazy formulas'). Steiner [16] proposes that the agent does not incorporate the new information if she already believes to the contrary. In that case, nothing happens. Otherwise, access to states where the information is not believed is eliminated, just as for believed public announcements. This solution to model unbelievable lies (and unbelievable truths!) is similarly proposed in the elegant and promising [9], where it is called *cautious update*—a suitable term.

Steiner gives a useful parable for the case where you do not accept new information. Someone is calling you and is telling you something that you don't want to believe. What do you do? You start shouting through the phone: 'What did you say? Is there anyone on the other side? The connection is bad!' And then you hang up, quickly, before the caller can repeat his message. Thus you create common knowledge that the message has been received but its content not accepted.

A three-point action model for cautious update is as follows. The difference with the action model for truthful and lying public announcement is that those alternatives now have an additional precondition $\neg B_i\neg\varphi$, meaning that the announcement is 'believable'.



[1] The logic of believed announcements was originally axiomatized in [7]. The redescription of these operations with an action model, providing the alternative axiomatization, was suggested in [17, 8].

We have explored this modelling of lying in more depth. We consider these mere variations, and move on. Note that for agent announcements, the addressee does not go mad if she already believes $\neg p$ and the speaker is lying that p. The addressee then merely concludes that $\neg p \wedge B_j p$: the speaker must be mistaken in his truthful belief of p. Of course the addressee will *still* go mad if she believed $\neg B_j p$.

For believed announcements we mentioned the problem that the agent believes new information whether it is true or not. For cautious update it still is the case that the agent can process (although maybe not believe) new information whether it is true or not, and even whether she already believed it or not. Going mad is too strong a response, but not changing contradictory beliefs is too weak. The next section presents a solution in between.

5.2 Lying and plausible belief

Suppose that we also have a preference relation, expressing which states are more and less plausible. We then can distinguish degrees of belief. For example, suppose states s and t are indistinguishable for agent i but she considers s more plausible than t; and proposition p is true in s and false in state t. The agent (defeasibly) *believes* φ if φ is true in all preferred states, and the agent *knows* (or, strongly believes) φ if φ is true in all accessible states. We keep writing B for belief and we write K for knowledge. Given that, $B_i p$ is true in t, because p is true in the preferred state s, but $K_i p$ is not true in t. When presented with evidence that $\neg p$, in t, i will eliminate s from consideration; t is now the most preferred state, and $B_i \neg p$ is now true. Such a distinction between epistemic access and preference can also be made in the action models, where agents may consider more and less plausible actions. We will refrain from details, see [18, 17, 4]. How to model lying with plausibility models was summarily discussed in [4, 19].

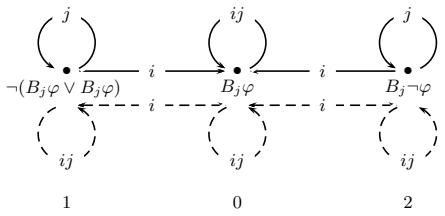


Fig. 2. Belief and preference

The action model of Figure 1 enriched with plausibility is depicted in Figure 2. The adressee i is most inclined to believe that j is telling the truth (0), less inclined to believe that he is bluffing (1), and least inclined to believe that he is lying (2). Agent i's accessibility relation is the dashed relation. (This is the universal relation. We assume transitivity.) She cannot excluse any of the three

types of announcement. From that and her preference the solid accessibility relation is what she considers most likely. This will determine her plausible beliefs. (A third, intermediate degree of belief, is implicit in the figure.) Now consider an epistemic state wherein i has hard evidence that $\neg B_j p$, and let j announces p, thus suggesting $B_j p$. In the first place, i will now not go mad, the problem discussed before. She will merely eliminate truthtelling from the alternatives, and from the two remaining alternatives she considers it more likely that j is bluffing than that he is lying. If she had also hard evidence that j is not bluffing, she will still not go mad, and finally conclude that he is a liar.

5.3 Lying about beliefs

If I lie to you that "you don't know that I will fly to Amsterdam tomorrow", something of the form $p \wedge \neg B_j p$, the lie succeeds if you believe p afterwards, i.e., if $B_j p$ is then true, not if the contradictory sentence $B_j(p \wedge \neg B_j p)$ is true. This is not merely some theoretical boundary case. I can very well lie to you about the knowledge or ignorance of other agents or about my own knowledge. In fact, I do that all the time.

Agents may announce factual propositions but also modal propositions, and thus be lying and bluffing about them. For example, in the consecutive number riddle, both i and j may lie about their knowledge or ignorance of the other's number.

In social interaction, untruthfully announcing modalities is not always considered lying (with the moral connotation). Suppose we work in the same department and one of our colleagues, X, is having a divorce. I know this. I also know that you know this. But we have not discussed the matter between us. I can bring up the matter in conversation by saying 'You know that X is having a divorce!'. But this is unwise. You may not be willing to admit your knowledge, because X's husband is your friend, which I have no reason to know; etc. A better strategy for me is to say 'You may not know that X is having a divorce'. This is a lie. I do not consider it possible that you do not know that. But, unless we are very good friends, you will not laugh in my face to that and respond with 'Liar!'.

It is also strange that I may be *bluffing* if I tell you p, given that in fact I don't know if p, but I would be *lying* if I tell you that I believe that p. This is because I believe that I don't believe p: $\neg B_i p$ entails by negative introspection $B_i \neg B_i p$, where $\neg B_i p$ is now the negation of the announced formula $B_i p$!

6 Conclusions and further research

Lying is an epistemic action inducing a transformation of an epistemic model. We presented logics for public lying and truthtelling, and logics for agent lying, bluffing, and truthtelling. These logics abstract from the moral and intentional aspect of lying, and only consider the effect of lies that are believed by the

addressee. We also presented versions that treat unbelievable lies differently, and lying in the presence of plausible (defeasible) belief.

There are many topics for further research. **1.** Explicit agency is missing in our approach (as so often in dynamic epistemic logics). **2.** We only summarily discussed common knowledge—this seems a straightforward enough generalization, that also allows for more more refined preconditions then merely requiring that lies are believable for the addressee. A good (and possibly strongest?) precondition seems:

$$B_i\neg\varphi \wedge \neg B_j\neg\varphi \wedge C_{ij}((B_i\varphi \vee B_i\neg\varphi) \wedge \neg(B_j\varphi \vee B_j\neg\varphi))$$

3. One problem with lying to some and telling the truth to others is that you have to keep track of who knows the truth and who not, and that you should carefully consider what you can still say and in whose company. In everyday communication, this (logical) computational cost of lying seems a strong incentive against lying. Can this intuition be formalized? We are inspired by results on the computational cost of insincere voting in social choice theory [6]: in well-designed voting procedures this is intractable, so that sincere voting is your best strategy. **4.** In multi-agent systems with several agents on may investigate how robust certain communication procedures are in the presence of few liars; and results might be compared to those for signal analysis with 'intentional' noise. **5.** Finally, we would like to model a liar's paradox in a dynamic epistemic logic.

Acknowledgement

I thank the workshop reviewers for their comments.

References

1. G. Aucher. Consistency preservation and crazy formulas in BMS. In S. Hölldobler, C. Lutz, and H. Wansing, editors, *Logics in Artificial Intelligence, 11th European Conference, JELIA 2008. Proceedings*, pages 21–33. Springer, 2008. LNCS 5293.
2. A. Baltag. A logic for suspicious players: Epistemic actions and belief updates in games. *Bulletin of Economic Research*, 54(1):1–45, 2002.
3. A. Baltag, L.S. Moss, and S. Solecki. The logic of public announcements, common knowledge, and private suspicions. In I. Gilboa, editor, *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, pages 43–56, 1998.
4. A. Baltag and S. Smets. The logic of conditional doxastic actions. In K.R. Apt and R. van Rooij, editors, *New Perspectives on Games and Interaction*, Texts in Logic and Games 4. Amsterdam University Press, 2008.
5. S. Bok. *Lying: Moral Choice in Public and Private Life*. Random House, New York, 1978.
6. V. Conitzer, J. Lang, and L. Xia. How hard is it to control sequential elections via the agenda? In *IJCAI'09: Proceedings of the 21st international joint conference on Artifical intelligence*, pages 103–108. Morgan Kaufmann Publishers Inc., 2009.

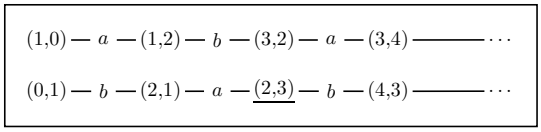
7. J.D. Gerbrandy and W. Groeneveld. Reasoning about information change. *Journal of Logic, Language, and Information*, 6:147–169, 1997.
8. B. Kooi. Expressivity and completeness for public update logics via reduction axioms. *Journal of Applied Non-Classical Logics*, 17(2):231–254, 2007.
9. B. Kooi and B. Renne. Arrow update logic. Manuscript, 2010.
10. J.E. Littlewood. *A Mathematician's Miscellany*. Methuen and company, 1953.
11. J.E. Mahon. Two definitions of lying. *Journal of Applied Philosophy*, 22(2):21–230, 2006.
12. J.E. Mahon. The definition of lying and deception. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, 2008. `http://plato.stanford.edu/archives/fall2008/entries/lying-definition/`.
13. J.A. Plaza. Logics of public communications. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic, and Z.W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, pages 201–216. Oak Ridge National Laboratory, 1989.
14. C. Sakama, M. Caminada, and A. Herzig. A logical account of lying. In *Proceedings of JELIA 2010, LNAI 6341*, pages 286–299, 2010.
15. F.A. Siegler. Lying. *American Philosophical Quarterly*, 3:128–136, 1966.
16. D. Steiner. A system for consistency preserving belief change. In *Proceedings of the ESSLLI Workshop on Rationality and Knowledge*, pages 133–144, 2006.
17. J. van Benthem. Dynamic logic of belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.
18. H. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese (Knowledge, Rationality & Action)*, 147:229–275, 2005.
19. H. van Ditmarsch. Comments on 'the logic of conditional doxastic actions'. In K.R. Apt and R. van Rooij, editors, *New Perspectives on Games and Interaction*, Texts in Logic and Games 4, pages 33–44. Amsterdam University Press, 2008.
20. H. van Ditmarsch, J. van Eijck, F. Sietsma, and Y. Wang. On the logic of lying. In J. van Eijck and R. Verbrugge, editors, *Games, Actions and Social Software*. Springer, 2011. FoLLI-LNCS series 'Texts in Logic and Games'. To appear.

Appendix: Lying about consecutive numbers

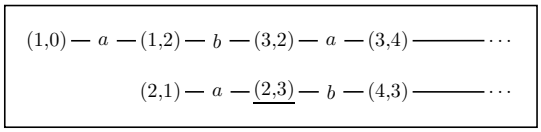
The consecutive numbers riddle is often attributed to Littlewood [10]. It is as follows.

> *Anne and Bill are each going to be told a natural number. Their numbers will be one apart. The numbers are now being whispered in their respective ears. They are aware of this scenario. Suppose Anne is told 2 and Bill is told 3.*
> *The following truthful conversation between Anne and Bill now takes place:*
> – *Anne: "I do not know your number."*
> – *Bill: "I do not know your number."*
> – *Anne: "I know your number."*
> – *Bill: "I know your number."*
> *Explain why is this possible.*

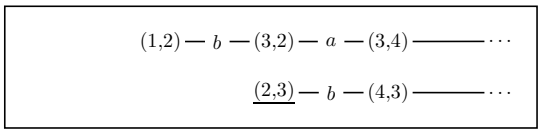
First, the standard analysis of the informative consequences of these four announcements.



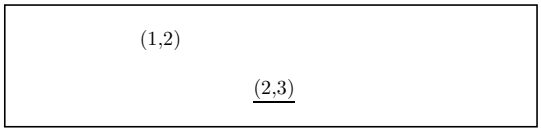
– Anne: "I do not know your number."



– Bill: "I do not know your number."



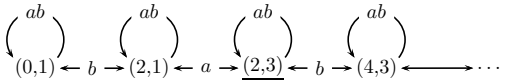
– Anne: "I know your number."



– Bill: "I know your number."

This last announcement does not make a difference anymore, as it is already common knowledge that Anne and Bill know each other's number.

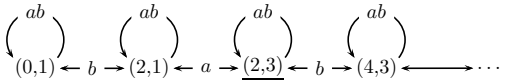
Next, we show two different scenarios for the consecutive number riddle with lying. This is agent lying (and truthtelling), the actions we modelled as $!_i\varphi$ and $¡_i\varphi$. (Bluffing is not an option in this example, because the lying is about ignorance or knowledge, and introspective agents *know* their ignorance and *know* their knowledge.) As we are reasoning from the actual state $(2,3)$, we do not depict the top chain of possibilities any more. And as beliefs may now be incorrect, we show all arrows. Positions in the model where a change took place (i.e., where arrows have been removed) are shown in red. The first scenario consists of Anne lying in her first announcement. We do not model Bill's response that Anne is a liar! After Anne's lie, in the actual state $(2,3)$, Bill does not consider any state possible, and therefore believes everything. (Of course you have Bill say that he has gone mad—by way of truthfully announcing that $B_j(p \wedge \neg p)$.)



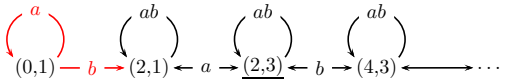
– Anne: “I know your number.” Anne is lying

– Bill: “That’s a lie.”

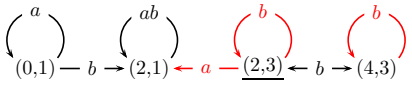
In the second scenario Anne initially tells the truth, after which Bill is lying, resulting in Anne mistakenly concluding (and announcing) that she knows Bill’s number: observe that she believes it to be 1. This mistaken announcement by Anne is informative to Bill: he learns from it (correctly) that Anne’s number is 3.



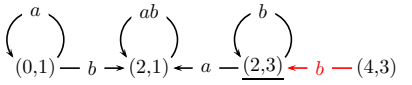
– Anne: “I do not know your number.”



– Bill: “I know your number.” Bill is lying



– Anne: “I know your number.” Anne is mistaken.



On combining cognitive and formal modeling: a case study involving strategic reasoning

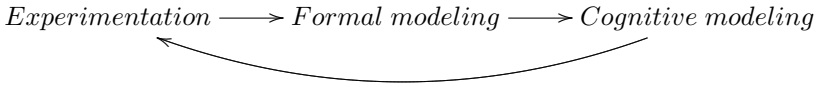
Sujata Ghosh and Ben Meijering

Department of Artificial Intelligence
University of Groningen.
sujata,b.meijering@ai.rug.nl

Abstract. This paper builds up a bridge between the formal and cognitive modeling of human reasoning aspects. To this end, we focus on empirical studies on playing a certain game, namely marble drop, that involves reasoning about other minds, and build up a formal system that can model the different strategic reasoning methods employed by the participants in the empirical study. Finally, we show how the syntactic framework of the formal system can aid in building up a cognitive model of the participants of the marble drop game.

1 Introduction

In recent years, a lot of questions have been raised regarding the idealization that a formal model undergoes while representing social reasoning methods (e.g. see [3]). Do these formal methods represent human reasoning satisfactorily or should we concentrate more on the empirical studies and models based on those empirical data? Without going into this debate here, we combine empirical studies, formal modeling and cognitive modeling to study human strategic reasoning. Our proposal is the following: rather than thinking about them as separate ways of modeling, we can consider them to be complementary and investigate how they can aid each other to bring about a more meaningful model of the real-life scenarios.



In [5], a formal framework has been introduced to model human strategic reasoning as exemplified by certain psychological experiments focusing on a dynamic game scenario, namely the Marble Drop game [10]. In continuation of the work done in [5], this paper builds on the formal framework to give a more realistic reasoning model of the participants. Moreover, we propose to use a cognitive model of these participants based on the formal framework.

For the experimental work, the advantage of using dynamic games to study higher-order social reasoning is that they allow for repeated presentation, which yields more observations than is typical in other paradigms such as, for example,

false-belief story and/or picture tasks. More observations yield more reliable outcome measures such as accuracy of decisions and decision (or reaction) times (RTs). Examples of dynamic games used in empirical studies are the Centipede game [9], the matrix game [6], the road game [4], and Marble Drop [10, 11]. These examples are all game-theoretically equivalent because they share the same extensive form, namely that of the original Centipede game [17].

Previous empirical studies have shown higher-order social reasoning to be far from optimal, and have argued that higher-order social reasoning is complicated and cognitively demanding (e.g., [19]). However, Meijering et al. [10, 11] demonstrated that performance improved to near ceiling if participants (1) were assigned to stepwise instruction and training, (2) were asked to predict the other player's move, and (3) were presented with concrete and realistic games.

Based on empirical findings that show that the participants do not always follow the backward induction method [13], in this paper a formal framework is presented to model forward, backward as well as combined reasoning attempts of the participants. As discussed in [15], in backward induction reasoning, a player, at every stage of the game, only reasons about the opponents future behavior and beliefs. On the other hand, in forward induction reasoning, a player, at every stage, only considers the past choices of the opponents. Based on the formal framework, a cognitive model is proposed as a better alternative to the model proposed in [8]. The new model can represent these different reasoning methods in the Marble Drop game.

Before proceeding into the main sections of this paper, we should mention here that this paper should be considered as a preliminary report of a cognitive model of strategic reasoning that is being constructed with the aid of a formal framework. We still need to do the important tasks of predicting and testing the strategies that have come to our notice based on the empirical findings in [10, 11] and the eye-tracking study reported in [12]. The formal framework is introduced to capture the findings of the eye-tracking experiment, so that it can provide an easy, mechanical representation of the eye-tracking analyses to be used in the construction of the cognitive computational model.

2 Empirical work

We provide here a short discussion of the experimental studies on which this work is based. The first part gives a description of the Marble Drop game and the second part provides an analysis of the eye-tracking experiment.

2.1 Marble drop game

Figure 1 depicts examples of a zeroth-, first-, and second-order Marble Drop game. A white marble is about to drop, and its path can be manipulated by removing trapdoors (i.e., the diagonal lines). In this example, the participant controls the blue trapdoors and the computer controls the orange ones. Each bin contains a pair of payoffs. The participant's payoffs are the blue marbles

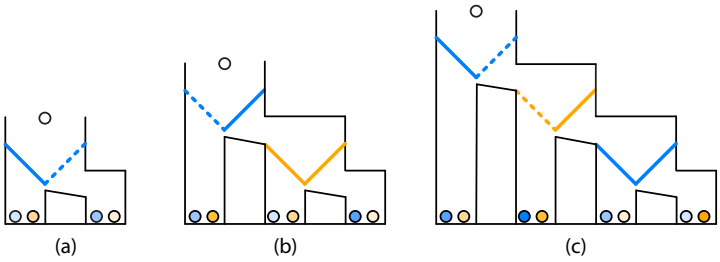


Fig. 1. Examples of a zeroth, first-, and second-order Marble Drop game. The blue marbles are the participant's payoffs and the orange marbles are the computer's payoffs. The marbles can be ranked from light to dark, light less preferred than dark. For each player, the goal is that the white marble drops into the bin with the darkest possible marble of their color. The participant controls the blue trapdoors (i.e., blue diagonal lines) and the computer the orange ones. The dashed lines represent the trapdoors that both players should remove to attain the darkest possible marble of their color.

and the computer's payoffs are the orange marbles. The marbles can be ranked from light to dark, light marbles being less preferred than dark. For each player, the goal is that the white marble ends up in the bin that contains the darkest possible color-graded marble of their color.

For example, at the start of the game in Figure 1c, Player I has to decide whether to remove the left trapdoor (end) or to remove the right trapdoor (continue). Player I's marble in bin 2 is darker than in bin 1, but what will Player II decide if Player I continues? Player II may want to continue the game to the last bin, as Player II's marble in bin 4 is darker than in bin 2, but what will Player I decide at the last set of trapdoors? Player I would stop the game in bin 3, as Player I's marble in bin 3 is darker than in bin 4. Thus, Player II should stop the game in bin 2, as Player II's marble in bin 2 is darker than in bin 3. Consequently, Player I should decide to continue the game from bin 1 to bin 2.

Marble Drop games provide visual cues as to which payoff belongs to whom, who decides where, what consequences decisions have, and how a game concludes. In matrix games [6], participants had to reconstruct this from memory. Meijering et al. [10] hypothesized that the supporting structure of the representation of Marble Drop would facilitate higher-order social reasoning, and, in fact, participants assigned to Marble Drop games performed better than participants assigned to matrix games [11].

2.2 Eye-tracking study

Behavioral measures such as responses and reaction times shed some light on higher-order social reasoning. However, they show the end result of higher-order social reasoning, not the online process. The online process (i.e., the strategies that participants use) may prove valuable in the study of higher-order social reasoning, because strategies determine to a great extent what cognitive resources

are employed. For example, an algorithmic strategy such as backward induction puts a lesser strain on working memory than a strategy that explicitly models mental states. Johnson, Camerer, Sen, and Rymon [7] used a novel approach to measure online higher-order social reasoning. In their sequential bargaining games, information displayed on a computer screen was masked with boxes and participants could uncover parts of that information by clicking on the boxes with the mouse. This approach allowed Johnson et al. to investigate the sequence in which participants uncovered information during reasoning.

A concern with this approach is that participants may have felt disinclined to click on the information repeatedly and would rather adopt an artificial strategy that involves fewer mouse clicks but puts a higher strain on working memory. To avoid that, Meijering, Van Rijn, Taatgen, and Verbrugge [12] conducted a study in which they used eye-tracking, which is not as obtrusive as Johnson et al.'s method of masking information. Participants' eye movements were recorded while they were playing Marble Drop games. The eye movement data yielded insight into the comparisons that participants made and the sequence of those comparisons during each game.

The proportions of fixations at bins 1 to 4 are depicted in Figure 2. The proportions were averaged over games, and plotted against position in the total fixation sequence. In other words, Figure 2 shows the general increase or decrease of fixations at a particular bin over time spent in a game. The results showed that participants did not seem to use backward induction, at least, initially. Figure 2(a) shows that, on the first position, the proportion of fixations at bins 1 and 2 was higher than at bins 3 and 4. In contrast, backward induction would yield a higher proportion of first fixations at bins 3 and 4, as backward reasoning starts with a comparison of the payoffs in bins 3 and 4. However, as of position 4 in Figure 2(a), the fixation patterns seem to correspond with backward induction: the proportion of fixations at bins 3 and 4 was higher than the proportion of fixations at bins 1 and 2, and the way that the proportions change over time (i.e., they decrease for bins 3 and 4, and increase for bins 1 and 2) correspond with eye movements that go from right to left.

The patterns are less obvious in Figure 2(b), because the figure shows fixations that were averaged over another set of games. Where Figure 2(a) (bottom panel) depicts mean proportions for games in which a rational participant should end the game because the computer would continue, Figure 2(b) (bottom panel) depicts mean proportions for games in which a rational participant should continue because the computer would continue. Differential fixation sequences imply that participants did not use pure backward induction, because backward induction works independently of the payoff values.

Instead of backward induction, participants may have applied forward reasoning, or a mix of backward and forward reasoning. Figure 2(a) hints at the latter possibility as participants fixated from left to right during the first four fixations, and from right to left during later fixations. To test what strategies participants may have used, we construct cognitive computational models (cf. Section 4) that implement various strategies, and use these models to predict eye movements that we can test against the observed eye movements. To aid in

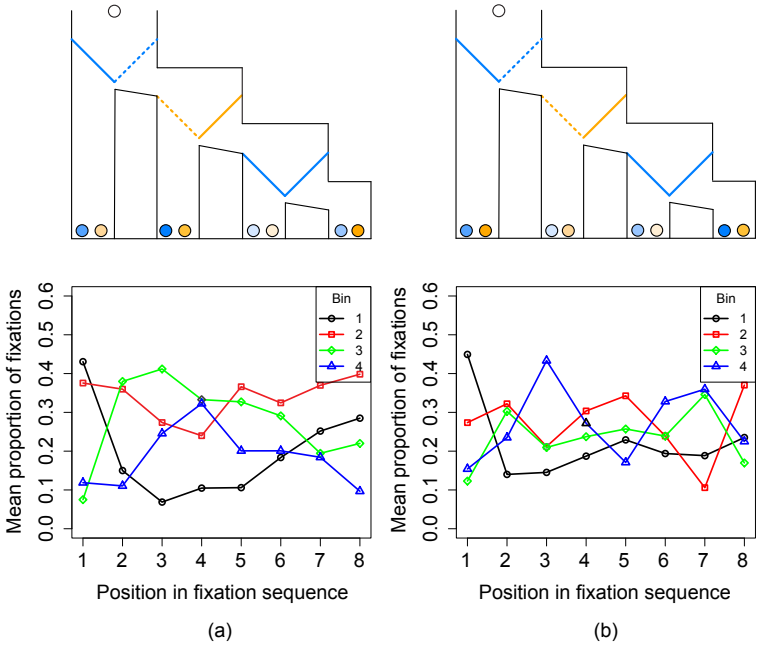


Fig. 2. The bottom panel depicts mean proportions of fixations at bins 1, 2, 3, and 4, calculated separately for position in the total fixation sequence. In (a), Player I should end the game in bin 1, because given the chance, Player II would continue, and the game would end with a lesser payoff for Player I . In (b), Player I should continue the game, because given the chance, Player II would continue, and the game would end with a better payoff for Player I . The games in the top panel are examples of the former and latter type of games. We did not depict standard errors, because we fitted (non-)linear models instead of traditional ANOVAs, which typically include contrasts between (successive) positions of fixations.

the construction we build up a formal framework (cf. Section 3) and show how the formal and cognitive modeling can interplay to provide a better model for strategic reasoning (cf. Section 4.2). As mentioned in the introduction, we are presently at the phase of building up this cognitive model and predicting and testing strategies are our next steps.

3 A formal framework

In this section, we present a formal system to represent the different ways of strategic reasoning that the participants of the Marble Drop game (cf. Section 2.1) undertake, suggested by the eye-tracking study described in Section 2.2. We extend the system developed in [5] by adding special propositional variables representing players' payoffs and comparison of such payoffs, inspired by [2].

3.1 Strategy specifications

Following the lines of work in [16, 14], a syntax for specifying partial strategies and their compositions in a structural manner involving simultaneous recursion has been proposed in [5]. The main case specifies, for a player, what conditions she tests for before making a move. The pre-condition for the move depends on observables that hold at the current game position as well as some simple finite past-time conditions and some finite look-ahead that each player can perform in terms of the structure of the game tree. Both the past-time and future conditions may involve some strategies that were or could be enforced by the players. These pre-conditions are given by the following syntax.

Below, for any countable set X, let $BPF(X)$ (the boolean, past and future combinations of the members of X) be sets of formulas given by the following syntax:

$$BPF(X) := x \in X \mid \neg\psi \mid \psi_1 \vee \psi_2 \mid \langle a^+\rangle\psi \mid \langle a^-\rangle\psi.$$

where $a \in \Sigma$, a finite set of actions.

Formulas in $BPF(X)$ can be read as usual in a dynamic logic framework and are interpreted at game positions. The formula $\langle a^+\rangle\psi$ (respectively, $\langle a^-\rangle\psi$) talks about one step in the future (respectively, past). It asserts the existence of an a edge after (respectively, before) which ψ holds. Note that future (past) time assertions up to any bounded depth can be coded by iteration of the corresponding constructs. The "time free" fragment of $BPF(X)$ is formed by the boolean formulas over X. We denote this fragment by $Bool(X)$.

Syntax Let $P^i = \{p^i_0, p^i_1, \ldots\}$ be a countable set of observables for $i \in N$ and $P = \bigcup_{i\in N} P^i$. To this set of observables we add two new kinds of propositional variables $(u_i = q_i)$ to denote 'player i's utility (or payoff) is q_i' and $(r \leqslant q)$ to denote that 'the rational number r is less than or equal to the rational number q'. The syntax of strategy specifications is given by:

$$Strat^i(P^i) := [\psi \mapsto a]^i \mid \eta_1 + \eta_2 \mid \eta_1 \cdot \eta_2,$$

where $\psi \in BPF(P^i)$. For a detailed explanation see [5]. The basic idea is to use the above constructs to specify properties of strategies as well as to combine them to describe a play of the game. For instance the interpretation of a player i's specification $[p \mapsto a]^i$ where $p \in P^i$, is to choose move "a" at every game position belonging to player i where p holds. At positions where p does not hold, the strategy is allowed to choose any enabled move. The strategy specification $\eta_1 + \eta_2$ says that the strategy of player i conforms to the specification η_1 or η_2. The construct $\eta_1 \cdot \eta_2$ says that the strategy conforms to specifications η_1 and η_2.

Let $\Sigma = \{a_1, \ldots, a_m\}$, we also make use of the following abbreviation.

- $null^i = [\top \mapsto a_1] + \cdots + [\top \mapsto a_m]$.

It will be clear from the semantics (which is defined shortly) that any strategy of player i conforms to $null^i$, or in other words this is an empty specification. The empty specification is particularly useful for assertions of the form "there exists a strategy" where the property of the strategy is not of any relevance.

Semantics We consider perfect information games as models. Let $M = (T, V)$ with $T = (S, \Rightarrow, s_0, \widehat{\lambda}, \mathcal{U})$, where $(S, \Rightarrow, s_0, \widehat{\lambda})$ is an extensive form game tree, $\mathcal{U} : frontier(T) \times N \to \mathbb{Q}$ is a utility function. Here, $frontier(T)$ denotes the leaf nodes of the tree T. Finally, $V : S \to 2^P$ is a valuation function. The truth of a formula $\psi \in BPF(P)$ at the state s, denoted $M, s \models \psi$, is defined as follows:

- $M, s \models p$ iff $p \in V(s)$.
- $M, s \models \neg\psi$ iff $M, s \not\models \psi$.
- $M, s \models \psi_1 \vee \psi_2$ iff $M, s \models \psi_1$ or $M, s \models \psi_2$.
- $M, s \models \langle a^+ \rangle \psi$ iff there exists an s' such that $s \overset{a}{\Rightarrow} s'$ and $M, s' \models \psi$.
- $M, s \models \langle a^- \rangle \psi$ iff there exists an s' such that $s' \overset{a}{\Rightarrow} s$ and $M, s' \models \psi$.

The truth definition for the new propositions are as follows:

- $M, s \models (u_i = q_i)$ iff $\mathcal{U}(s, i) = q_i$.
- $M, s \models (r \leqslant q)$ iff $r \leqslant q$, where r, q are rational numbers.

Strategy specifications are interpreted on strategy trees of T. We also assume the presence of two special propositions $\mathbf{turn}_1$ and $\mathbf{turn}_2$ that specify which player's turn it is to move, i.e. the valuation function satisfies the property

- for all $i \in N$, $\mathbf{turn}_i \in V(s)$ iff $\widehat{\lambda}(s) = i$.

One more special proposition **root** is assumed to indicate the root of the game tree, that is the starting node of the game. The valuation function satisfies the property

- $\mathbf{root} \in V(s)$ iff $s = s_0$.

A partial strategy σ, say of player i, can be viewed as a set of total strategies of the player [14] and each such strategy is a subtree of T.

The semantics of the strategy specifications are given as follows. Given the game $T = (S, \Rightarrow, s_0, \widehat{\lambda}, \mathcal{U})$ and a partial strategy specification $\eta \in Strat^i(P^i)$, we define a semantic function $[\![\cdot]\!]_T : Strat^i(P^i) \to 2^{\Omega^i(T)}$, where each partial strategy specification is associated with a set of total strategy trees.

For any $\eta \in Strat^i(P^i)$, the semantic function $[\![\eta]\!]_T$ is defined inductively as follows:

- $[\![[\psi \mapsto a]^i]\!]_T = \Upsilon \in 2^{\Omega^i(T)}$ satisfying: $\mu \in \Upsilon$ iff μ satisfies the condition that, if $s \in S_\mu$ is a player i node then $M, s \models \psi$ implies $out_\mu(s) = a$.
- $[\![\eta_1 + \eta_2]\!]_T = [\![\eta_1]\!]_T \cup [\![\eta_2]\!]_T$
- $[\![\eta_1 \cdot \eta_2]\!]_T = [\![\eta_1]\!]_T \cap [\![\eta_2]\!]_T$

Above, $out_\mu(s)$ is the unique outgoing edge in μ at s. Recall that s is a player i node and therefore by definition of a strategy for player i there is a unique outgoing edge at s.

To model players' responses, we introduce the formula $\bar{\imath}?\zeta$ in the syntax of $BPF(P^i)$, where $\bar{\imath}$ denotes the opponent of i. The intuitive reading of the formula is "player $\bar{\imath}$ is playing according to a partial strategy conforming to the specification ζ at the current stage of the game", and the semantics is given by,

- $M, s \models \bar{\imath}?\zeta$ iff $\exists T'$ such that $T' \in [\![\zeta]\!]_T$ and $s \in T'$.

3.2 Marble Drop game: a test case

We now express the empirical strategic reasoning performed by the participants of the Marble drop game described in Section 2.1. The game form is structurally equivalent to the Centipede game tree. Figure 3a gives the corresponding tree structure, and figures 3b and 3c correspond to example cases.

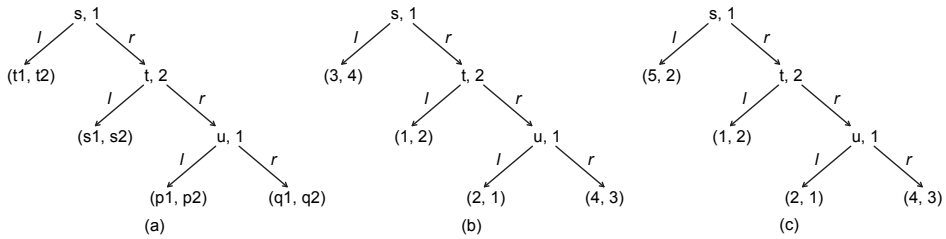


Fig. 3. Example trees.

Using the strategy specification language introduced in Section 3.1, we express the different reasoning methods of participants that have been validated by the experiments described in Section 2. The reasoning is carried out by an outside agent (participant) regarding the question:

> How would the players 1 and 2 play in the game, under the assumptions that both players are rational (thus will try to maximize their utility), and that there is common knowledge of rationality among the players.

We abbreviate some formulas which describe the payoff structure of the game.

$\langle r\rangle\langle r\rangle\langle l\rangle((u_1 = p_1) \wedge (u_2 = p_2)) = \alpha$ (two r moves and one l move lead to (p_1, p_2))
$\langle r\rangle\langle r\rangle\langle r\rangle((u_1 = q_1) \wedge (u_2 = q_2)) = \beta$ (three r moves lead to (q_1, q_2))
$\langle r\rangle\langle l\rangle((u_1 = s_1) \wedge (u_2 = s_2)) = \gamma$ (one r move and one l move lead to (s_1, s_2))
$\langle l\rangle((u_1 = t_1) \wedge (u_2 = t_2)) = \delta$ (one l move leads to (t_1, t_2))

A formula describing backward reasoning giving the correct answer corresponding to the game tree given in Figure 3b is:

$$\varphi_1 : \big([\alpha \wedge \beta \wedge \langle r\rangle\langle r\rangle \mathbf{turn}_1 \wedge (2 \leqslant 4) \wedge \gamma \wedge \langle r\rangle \mathbf{turn}_2 \wedge (2 \leqslant 3) \wedge \mathbf{root} \wedge \mathbf{turn}_1 \wedge \delta \wedge (3 \leqslant 4) \mapsto r]^1, [\alpha \wedge \beta \wedge \langle r\rangle\langle r\rangle \mathbf{turn}_1 \wedge (2 \leqslant 4) \wedge \gamma \wedge \langle r\rangle \mathbf{turn}_2 \wedge (2 \leqslant 3) \mapsto r]^2, [\alpha \wedge \beta \wedge \langle r\rangle\langle r\rangle \mathbf{turn}_1 \wedge (2 \leqslant 4) \mapsto r]^1\big)$$

> 'If the utilities and the turns of players at the respective nodes are as in Figure 3b, then player 1 would play r at the root node, player 2 would continue playing r at his node, after which player 1 can finish off by playing r.'

Another formula describing forward reasoning giving a wrong answer corresponding to the game tree given in Figure 3b is:

$$\varphi_2 : \big([\mathbf{root} \wedge \mathbf{turn}_1 \wedge \delta \wedge \langle r \rangle \mathbf{turn}_2 \wedge \gamma \wedge (1 \leqslant 3) \mapsto l]^1\big)$$

'If the utilities at the first two leaf-nodes of the game are as Figure 3b, and players 1 and 2 move respectively in the first two non-terminal nodes, then player 1 would play l at the root node finishing it off.'

The last formula describes forward reasoning giving a correct answer corresponding to the game tree given in Figure 3c is:

$$\varphi_3 : \big([\mathbf{root} \wedge \mathbf{turn}_1 \wedge \delta \wedge \langle r \rangle \mathbf{turn}_2 \wedge \gamma \wedge (1 \leqslant 5) \mapsto l]^1\big)$$

'If the utilities at the first two leaf-nodes of the game are as Figure 3c, and players 1 and 2 move respectively in the first two non-terminal nodes, then player 1 would play l at the root node finishing it off.'

These are just some examples to show that one can actually list possible ways of reasoning that can be performed by human reasoners in the Marble Drop game. Such a list aids in developing the cognitive models of the reasoners, as we shall see in the next section.

4 Cognitive modeling

Analyses of eye movements are challenging because they have to deal with great variability typically found in eye-movement data. Salvucci and Anderson [18] suggested using a cognitive computational model to predict eye movements, which can be compared with observed eye movements. This method helps to disentangle explained (i.e., hypothesized) variance from unexplained variance (due to e.g. measurement errors).

Van Maanen and Verbrugge [8] suggested a cognitive model that implemented backward induction. However, the eye-tracking study conducted by Meijering et al. [12] suggests that participants did not use pure backward induction. Thus, in this paper we present preliminary ideas about a more generic cognitive model that implements backward and forward reasoning as well as possible mixtures of the two. Before going into the specific details of our construction of the cognitive computational model, we first provide a general description of the model that we are going to develop.

4.1 ACT-R modeling

The model that we propose has been implemented in ACT-R, which is an integrated theory of cognition as well as a cognitive architecture that many cognitive scientists use to model human cognition [1]. ACT-R consists of modules that link with cognitive functions (e.g., vision, motor processing, and declarative processing) and map with specific brain regions. Each module has a buffer associated with it, and the modules communicate among themselves via these buffers.

A very important property of ACT-R is that cognitive resources are bounded, because each buffer can store just one piece of information at a time. Consequently, if a model has to keep track of more than one piece of information, it has to move it back and forth between two important modules: declarative memory and the problem state. Moving information back and forth comes with a time cost, and could cause a so-called cognitive bottleneck.

The declarative memory module represents long-term memory and stores information encoded in so-called chunks (i.e., knowledge structures). For example, a chunk can be represented as some expression with a defined meaning (e.g. formal expressions). Each chunk in declarative memory has an activation value that determines the speed and success of its retrieval. Whenever a chunk is used, the activation value of that chunk increases. As the activation value increases, the probability of retrieval increases and the latency of retrieval decreases. For example, whenever the chunk of a successful formula is used, its activation value increases. As the activation value of a successful formula increases, its probability (and speed) of retrieval increases.

Anderson [1] provided a formalization of the mechanism that produces the relationship between the probability and speed of retrieval. As soon as a chunk is retrieved from declarative memory, it is put into the module buffer. As mentioned earlier, each ACT-R module has a buffer that may contain one chunk at a time. On a functional level of description, the chunks that are stored in the various buffers are the knowledge structures the cognitive architecture is aware of.

The problem state module (sometimes referred to as 'imaginal') slightly alleviates bounds on cognitive resources, as it also contains a buffer that can hold one chunk. Typically, the problem state stores a sub-solution to the problem at hand. In the case of a social reasoning task, this may be the outcome of a reasoning step that will be relevant in subsequent reasoning. Storing information in the problem state buffer is associated with a time cost (typically 200ms). The cognitive model that we present relies on the declarative and problem state modules. More specifically, it retrieves relevant information from declarative memory and moves that information to the problem state buffer whenever it requests the declarative module to retrieve new information, which the declarative module stores in its buffer.

A central procedural system recognizes patterns in the information stored in the buffers, and responds by sending requests to the modules, for example, 'retrieve a fact from declarative memory'. This condition-action mechanism is implemented in production rules. For example, the following production rule represents comparing the last two payoff values in order to decide whether to end or continue a Marble Drop game:

IF the goal is to compare the last two payoff values,
AND the first is greater than the second,
THEN respond end the game.

Here, the first line refers to the goal buffer, the second line to the problem state buffer, and the third line to a manual action. With this brief introduction to ACT-R modeling, we now move on to the specific model construction.

4.2 A cognitive computational model of Marble Drop

The cognitive model that we propose here is based on the model presented previously by [8], but it is more generic because it is not based on a fixed strategy. Instead, the model is based on formulas (cf. Section 3.2) that are selected from a list provided by the logical framework. The formulas can either represent backward reasoning, forward reasoning, or a mix of both (see examples φ_1, φ_2, φ_3 in Section 3.2).

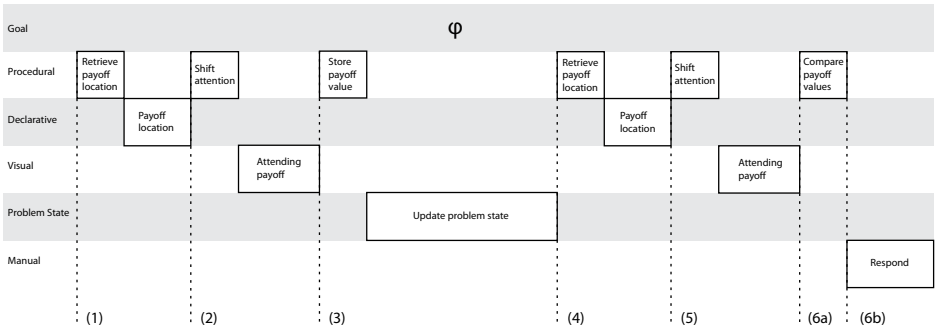


Fig. 4. Flowchart of the ACT-R model.

The flowchart of the model is depicted in Figure 4. Throughout an entire game, the goal buffer stores a chunk that represents which formula (represented by φ) is used. For each pair of payoffs that are compared in the formula, the model iterates through the following steps: The model retrieves the location of the first payoff from declarative memory (1). That location is represented in a chunk, and the more often a location chunk is retrieved the faster that retrieval will be, as the activation value of a chunk increases with each retrieval. As soon as the location of the first payoff is retrieved from declarative memory, the model shifts attention to that location (2). More specifically, the model requests the visual module to shift attention to the location it has just retrieved. After attending the payoff, the model stores the payoff value in the problem state buffer (3). If the model does not store the payoff value in the problem state buffer, it will be lost (i.e., replaced) when the model retrieves a new piece of information. Whenever the payoff value is moved from declarative memory to the problem state, the model retrieves the location of the second payoff from declarative memory (4). After retrieving that location, the model shifts attention to it (5). Now, the model has attended both payoffs, and it compares the payoff value stored in the problem state with the payoff value stored in the visual buffer (6a). After comparing the last pair of payoffs in the formula, the model produces a response (6b).

At the start of each game, a new formula chunk is retrieved from declarative memory. The model tags a formula chunk according to its success, that is, whether the model's response was correct or incorrect, which is indicated by the task feedback presented after each game. The model learns to play Marble Drop games better and faster, as it requests the declarative module to retrieve successful formulas, and the more often those are retrieved and tagged, the higher their activation value becomes. Higher activation value, in turn, increases the probability and speed of retrieving a formula.

The model produces responses and associated reaction times, which we can analyze and compare with the behavioral data. In addition, the model also produces fixations, which we can compare with the human eye movement data. By comparing the model's fixation sequences with the observed fixation sequences in Marble Drop games (Meijering et al. [12]), we can determine what formulas provide a good description of human higher-order social reasoning.

5 Conclusion

The eye-tracking study of Meijering et al. [12] has shown that participants did not use a pure backward induction strategy in the Marble Drop game. We, therefore, constructed a logical model to describe the game, and possible strategies. We use the logical model as a basis for a cognitive computational model, implemented in the cognitive architecture ACT-R.

We want to emphasize that the cognitive model can be considered as a virtual human being. It can do the very same task presented to the participants in Meijering et al.'s [10, 11] studies, and it produces responses and associated response times. The cognitive computational model is useful for a better understanding of higher-order social reasoning, because we can analyze the model output and see which formulas are successful and how quickly the model learns to apply one (set of) formula(s) instead of other formulas.

An advantage of having cognitive models, besides having statistical models, is that cognitive models can be broken down into mechanisms. Our ACT-R model comprises cognitive functions (e.g., a declarative memory and a problem state representation), and we can determine to what extent each cognitive function contributes to the model's behavior (i.e., the responses and response times) in Marble Drop games.

Another advantage of a cognitive model is that we can compare the model's output with Meijering et al.'s human data, and acquire a better understanding of individual differences. Higher-order social reasoning probably consists of multiple serial and concurrent cognitive functions, and thus it may be prone to great individual differences. Our cognitive model may help to determine what formulas fit the responses of a particular (subset of) participant(s). This fit not only concerns patterns in responses and response times, but also patterns in eye-movements. The model's execution of a formula yields eye movements, and we can calculate the explanatory power of eye movement patterns in (subsets of) the human data.

Acknowledgements: The authors gratefully thank Rineke Verbrugge for motivating this work, reading the preliminary drafts a number of times, and always suggesting meaningful changes in both the content and the language of the presentation. The authors would also like to thank the anonymous reviewers for their comments which helped to improve this paper. The first author acknowledges the Netherlands Organisation of Scientific Research grant 600.065.120.08N201, and the second author acknowledges the Netherlands Organisation of Scientific Research grant 227-80-001.

References

1. J. Anderson. *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, New York (NY), 2007.
2. G. Bonanno. Modal logic and game theory: Two alternative approaches. *Risk, Decision and Policy*, 7(03):309–324, December 2002.
3. B. Edmonds. Social intelligence and multi-agent systems. In *Invited talk, MALLOW'10*, 2010.
4. L. Flobbe, R. Verbrugge, P. Hendriks, and I. Krämer. Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17:417–442, 2008. Special issue on formal models for real people, edited by M. Counihan.
5. S. Ghosh, B. Meijering, and R. Verbrugge. Logic meets cognition: empirical reasoning in games. In *Proceedings of the 3rd International Workshop on Logics for Resource Bounded Agents (LRBA 2010), in 3rd Multi-Agent Logics, Languages, and Organisations Federated Workshops, MALLOW'10, CEUR Workshop Proceedings*, volume 627, pages 15–34, 2010.
6. T. Hedden and J. Zhang. What do you think I think you think? Strategic reasoning in matrix games. *Cognition*, 85:1–36, 2002.
7. E. J. Johnson, C. F. Camerer, S. Sen, and T. Rymon. Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory*, 104(1):16–47, 2002.
8. L. v. Maanen and R. Verbrugge. A computational model of second-order social reasoning. In *Proceedings of the 10th International Conference on Cognitive Modeling*, pages 259–264, 2010.
9. R. McKelvey and T. Palfrey. An experimental study of the centipede game. *Econometrica*, 60(4):803–836, 1992.
10. B. Meijering, L. v. Maanen, H. v. Rijn, and R. Verbrugge. The facilitative effect of context on second-order social reasoning. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society, Cognitive Science Society*, pages 1423–1428, 2010.
11. B. Meijering, H. v. Rijn, N. Taatgen, and R. Verbrugge. I do know what you think I think: Second-order theory of mind in strategic games is not that difficult. In *Proceedings of the 33nd Annual Meeting of the Cognitive Science Society, Cognitive Science Society*, 2011.
12. B. Meijering, H. van Rijn, N. Taatgen, and R. Verbrugge. Eye movements during higher-order social reasoning. Technical report, University of Groningen, 2011. (in prep).
13. M. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge, MA, 1994.

14. S. Paul, R. Ramanujam, and S. Simon. Stability under strategy switching. In *Proceedings of the 5th Conference on Computability in Europe (CiE 2009), LNCS 5635*, pages 389–398. Springer, 2009.
15. A. Perea. Backward induction versus forward induction reasoning. *Games*, 1(3):168–188, 2010.
16. R. Ramanujam and S. Simon. A logical structure for strategies. In *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, volume 3 of *Texts in Logic and Games*, pages 183–208. Amsterdam University Press, 2008.
17. R. Rosenthal. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic theory*, 25(1):92–100, 1981.
18. D. D. Salvucci and J. R. Anderson. Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16(1):39–86, 2001.
19. R. Verbrugge and L. Mol. Learning to apply theory of mind. *Journal of Logic, Language and Information*, 17:489–511, 2008. Special issue on formal models for real people, edited by M. Counihan.

First and second-order false-belief reasoning: Does language support reasoning about the beliefs of others?

Bart Hollebrandse, Angeliek van Hout and Petra Hendriks

University of Groningen

1 Introduction[1]

We understand and act upon the beliefs of other people, even if these conflict with our own beliefs. Children's development of this ability, known as Theory of Mind, has been extensively studied over the past twenty-five years, starting with the seminal study of Wimmer and Perner (1983). Theory of Mind (ToM) development involves various aspects of reasoning about others, including social awareness, joint attention, and anticipation of other people's behavior. Reasoning about false beliefs—the ability to handle the contrast between true and false beliefs—seems to develop rather late. It is typically not until the age of four that children understand that, for instance, *John thinks that it is raining outside* contains a belief about the weather attributed to John, and know that John may be incorrect in his belief, thus attributing a false belief to another person (Astington, 1993; Wellman & Bartsch, 1988). Our study involves more complex false-belief reasoning adding another belief layer, for instance, *Tom believes that John thinks that it is raining outside*. We investigated when children succeed at complex false-belief reasoning as tested with a verbal story task at the age

[1] The tests used in this study have been developed by the first author in collaboration with Tom Roeper, Jill and Peter de Villiers and Kate Hobbs. We thank the children and teachers of the Annie M.G. Schmidt school and the Joseph Haydnschool in Groningen for their hospitality. This paper has benefitted from the comments of three anonymous reviewers. We gratefully acknowledge ESF (grant no. 028395, PI M. Krifka and U. Sauerland) and NWO (grant no. 277-70-005, PI P. Hendriks).

of 7, whereas they pass a similarly complex task with non-verbal movie clips not until the age of 8 or 9.

Why does false-belief reasoning develop so late? De Villiers (2005) argues that the acquisition of the syntax of linguistic embedding, with verbs like *say (Mom said that it was raining)*, is a prerequisite for developing the cognitive representations required for false beliefs. The child has to be able to embed sentential complements before she can represent a false belief in a cognitive embedding and reason about it accurately. Recent work, however, found that toddlers as young as 15 months old are able to pass a false-belief task. In a non-verbal version of the so-called Sally-Ann task, Onishi and Baillargeon (2005) showed toddlers movies in which a toy is hidden in one location while the actor is watching; when the actor's view is blocked by a screen, the toy is hidden in another location. When the screen opens again and the actor is about to reclaim the toy, the toddlers looked longer at the location where the toy was initially hidden (i.e., where the actor thinks that the toy is), than at the location where the toy was hidden now. The children's looks reveal their expectation of the actor's behavior on the basis of that person's belief about the hiding place of a toy, which is different from their own beliefs about it (cf. Southgate, Senju & Csibra, 2007, who replicated this finding with 25-month-old children). The children thus seem to track the actor's false belief about the location of a toy vis-à-vis their own, true beliefs.

Although one may doubt whether these tasks, which measure expectation, test the actual reasoning involved in considering false beliefs (see Apperly & Butterfill, 2009; De Bruin, 2011), the Onishi and Baillargeon findings undoubtedly show that 15-month-olds effectively represent false beliefs. Toddlers develop implicit knowledge of false-belief attribution well before they can verbalize that knowledge explicitly,

which suggests that they have some form of cognitive representation of false beliefs that does not rely on language. These young learners pass this ToM test before they pass any of the verbal false-belief tasks and before they have acquired any complex syntax, thereby refuting the basis of De Villiers's (2005) hypothesis.

In this study we present a more complex case of false-belief reasoning with older children that finds just the opposite: 7-year-olds pass a verbal false-belief reasoning task, but fail on an equally complex non-verbal task. It is not until the age of 8 or 9 that children pass this non-verbal task as well. The case under investigation involves two layers of belief representations: the ability to understand one person's belief (the first layer) about a belief attributed to another person (the second layer), as in *Tom believes that John thinks that it is raining outside*, where Tom entertains the knowledge that John holds a certain belief about the weather. Perner and Wimmer (1985) claim that this latter type of ToM development—second-order reasoning—is not mastered until the age of 7 or 8 (see also Sullivan, Zaitchik & Tager-Flusberg, 1994). We probed second-order ToM reasoning in 6-9 year-olds with a verbal and a non-verbal task. We argue that for such complex false-belief tasks, language may support the development of the cognitive representations of reasoning required to perform these tasks, reviving De Villiers' hypothesis about the role of language in false-belief reasoning.

2 Participants

43 Dutch children were tested, divided over two age groups: twenty-one 6 and 7 year-olds (mean age = 6;9, range = 6;2 – 7;3) and twenty-two 8 and 9 year-olds (mean age = 8;10, range = 8;2 – 9;11). We also tested a control group of seventeen adults.

3 Method

We used two tasks to test false-belief (FB) reasoning at first-order and second-order levels, the designs and materials of which were taken from the study of Hollebrandse, Hobbs, De Villiers and Roeper (2008) with English children. The essence of both tasks is that the protagonists in the stories and video clips have beliefs about situations that are different from the participants' beliefs (first order), as well as from the beliefs of others (second order). The two tasks differ as to how the clues for the beliefs were presented. In the verbal task, participants were told a story about four characters which provided the necessary clues for FB reasoning (see Appendix 1 for an illustration). In the non-verbal task, participants watched silent movies with one or two actors. The experimenter occasionally pointed out some features in the movies, but, crucially, no language clues about beliefs were given. Instead, the clues for the beliefs of the different actors had to be deduced from the visual context (see Appendix 2 for an illustration).

All subjects participated in both tasks. The data was collected in two sessions. The order in which the tasks were conducted was balanced across participants.

3.1 Verbal false-belief task

In the verbal task an elaborate story was told in which the beliefs of various people in the story were manipulated. The stories were accompanied by pictures, which were presented one by one and served as a memory aid. The stories were modeled after Wimmer and Perner's (1985) "ice cream truck story", but in contrast to their stories, we made sure that there were no overlapping beliefs, not only at the second-order

level, but also at the first-order level: each protagonist had his or her own distinct belief which was different from those of the other protagonists.

All the stories have the same set up. Protagonist 1 and 2 initially share the same belief. In the sample story in Appendix 1, both main characters (Sam and Maria) initially think that there are chocolate-chip cookies at the bake sale of the church. Then character 1's belief changes without character 2 knowing about it (Sam's mom tells Sam that they are selling pumpkin pie). Next, character 2 learns that the reality is different, without character 1 knowing about this (Maria finds out that there are only brownies left). At this point character 1 has a first-order belief which differs from his initial belief and also from the reality (Sam now thinks they're selling pumpkin pie, not chocolate-chip cookies; he doesn't know that in reality they're selling brownies). Character 2 knows the reality which is different from her second-order belief about character 1 (Maria knows they're selling brownies, but thinks that Sam still thinks that they sell chocolate-chip cookies).

We did not use any second-order embedding constructions of the type *Maria thinks that Sam thinks they are selling cookies at the bake sale* in the story. Instead we elicited a second-order answer by asking a "double" first-order question. The mailman asks Maria a first-order question *What does Sam think they are selling at the bake sale?* The experimenter then asks the participant what Maria answered to the mailman (see also Sullivan, Zaitchik & Tager-Flusberg, 1994). The child thus did not need to process second-order embedding structures in language, but was still required to do a second-order reasoning task.

There were eight stories with this format, each containing one second-order question and two first-order ones. One first-order question was asked in the middle of the

story and the second one was asked at the end of the story. The purpose of asking the same first-order question once more at the end of the story was to check whether children had difficulties with the length and complexity of the story. The second first-order question thus effectively served as a control of how well participants were able to keep track of the different beliefs despite the length and complexity of the story.

3.2 Non-verbal false-belief task

For the non-verbal task, participants also had to keep track of the different beliefs of different protagonists in the same situation. Whereas the former task was a fully verbal one, this one limited the use of language as much as possible. The experimenter only drew attention to the contents of a box and pointed out whether or not the observers in the movies were watching the changes of the content. This was done without using any propositional-attitude verbs (such as *think* or *believe*), and without referring to thoughts or beliefs in any other way.

Participants watched short movies. In half of the movies the contents of a box were changed once or twice (the Unknown-Change-of-Content set-up), and in the other half an object was moved between two or three different locations (the Unknown-Change-of-Location set-up). The task was presented as a game in two parts. In the first part the participant himself was the player of the game and had to keep track of what the observer in the movie knew about the changes of the contents of a box (or the different locations in the change-of-location variant of this task). These were the first-order trials (see Appendix 2a).

In the second part participants were told that it was the same game, but now there was a different player: an additional observer in the movie (i.e., the man in the win-

dow on the right in Appendix 2b). The task of the participant was to keep track of what this observer knew about the contents of the box and what he knew about the other observer (the woman)'s beliefs, which involves second-order reasoning. The set of events in the second-order movies was essentially the same as in the first-order movies, except that here we introduced a second-order false belief for the man about the woman. For example, the man would incorrectly believe that the woman thinks there is an apple in the box, whereas she actually believes that there is small basket in the box; in reality, however, there is a turtle in the box.

Four movies tested first-order FB reasoning and four others tested second-order FB reasoning. For the younger children, a first-order question was added to the end of the second-order trials to check whether they were able to follow the complex series of events.[2]

In both tasks then, participants had to reason about first-order and second-order false beliefs in a setting in which none of the protagonists' beliefs overlapped. In both tasks participants had to keep track of two first-order beliefs and one second-order belief. In the story task they had to keep track of Sam's and Mary's first-order beliefs as well as Mary's second-order belief about Sam's first-order belief. In the movie task they had to keep track of the first-order beliefs of the protagonist in the right window and the protagonist in the left window, and the second-order belief of

[2] The second first-order question was not asked with the older children as they were tested before the younger children. At that time, we feared that adding another question would make the task too demanding. However, as the older children turned out to be quite successful with the second first-order question in the verbal task, our fears appeared to be unwarranted. Hence, we decided to add the second first-order question to the non-verbal task for the younger children.

the protagonist in the right window about the belief of the other protagonist in the left window.

The tasks differed in whether the clues about who believes what about whom were presented verbally, or whether they had to be deduced from the movies, hence our labels verbal versus non-verbal task. Note that even though we call the movie task non-verbal, it is not completely non-verbal, as the experimenter draws verbal attention to the changes of the contents of the box; moreover, the test questions were also verbal. Conversely, the verbal task was supported with pictures. The two tasks also differed in the number of protagonists; the non-verbal task has only two protagonists (the woman and the man), whereas the verbal task has four (Sam, Maria, the mom and the mailman). Having more protagonists adds to the complexity of the mental representations involved and potentially makes the verbal task more demanding. However, as we will see in the next section, this is not reflected in the results. Children are more accurate in the verbal task than in the non-verbal task.

4 Results

Figures 1 and 2 present the children's scores on the verbal and non-verbal tasks. For both, the results show a sharp difference between first-order and second-order questions. Moreover, for the second-order items, children performed better in the verbal task than in the non-verbal task. The adults performed nearly at ceiling at all test questions, with 96% correct responses on the second-order question in the verbal task and 91% in the non-verbal task.

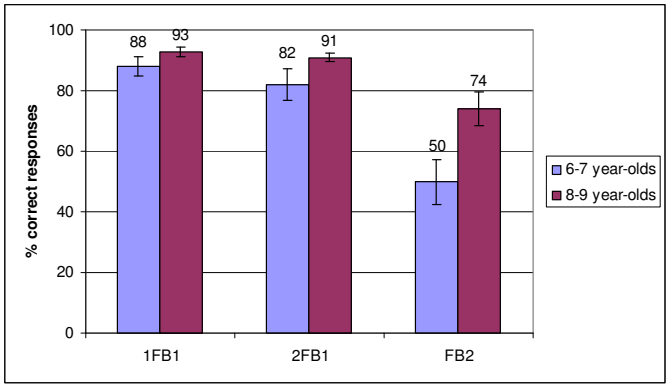


Figure 1: Verbal false-belief task: Percentage of correct responses for both age groups (error bars show standard errors) on 1FB1 (first first-order FB question), 2FB1 (second first-order FB question) and FB2 (second-order FB question).

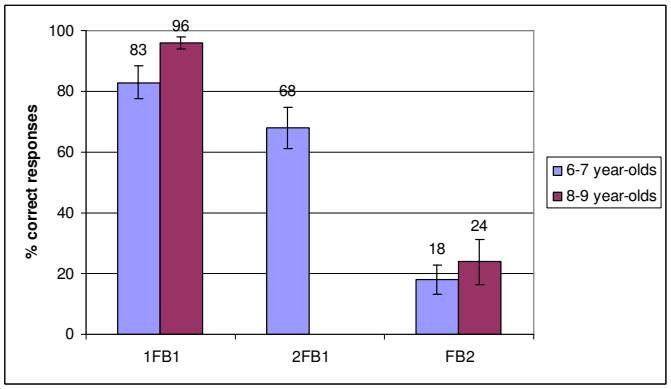


Figure 2: Non-verbal false-belief task: Percentage of correct responses for both age groups (error bars show standard errors) on 1FB1 (first first-order FB question), 2FB1 (first-order FB question in the second-order trial) and FB2 (second-order FB question). Note that the older children did not receive the 2FB1 question.

Repeated Measures ANOVAs were performed with Verbal Level (verbal – non-verbal) and False Belief Order (first order – second order) as within-participants factors, and Age (younger children – older children) as the between-participants factor. There were main effects for Verbal Level, $F(2,43) = 51.4$; $p < 0.001$) and False Belief Order, $F(2,43) = 160.7$; $p < 0.001$), and a significant interaction between the two ($F(2,43) = 37.4$; $p < 0.001$). The children performed better on the verbal than on the

non-verbal task and they also performed better on first-order questions than on second-order questions. There was a difference between performance on first-order and second-order questions in the non-verbal but not in the verbal task.

Taking a closer look at these effects, we performed paired-sample t-tests. These only revealed significant differences between the verbal and the non-verbal second-order responses, both for the younger group ($t(20) = 4.1$; $p = 0.001$) and the older group ($t(20) = 6.5$; $p < 0.001$). Importantly, there were no significant differences between any of the first-order false-belief responses, except for one: in the non-verbal task the younger group showed a significant difference between the first and second first-order question ($t(20) = 2.6$; $p = 0.015$). This general lack of effect at first-order level indicates that participants had no problem at this level of reasoning. The difficulties lie instead at the second-order level of reasoning.

Furthermore, Age was also significant ($F(2,43) = 7.9$; $p = 0.008$). The younger group performed worse than the older group. Paired sample t-tests reveal that this difference mostly lies in the difference between the two age groups on the verbal second-order false belief question ($t(20) = 2.1$; $p = 0.050$). Moreover, there was a trend for Age for the first first-order FB question in the non-verbal task ($t(20) = 2.0$; $p = 0.056$).

5 Discussion

Onishi and Baillargeon (2005) show that implicit non-verbal FB reasoning is accomplished at a very young age, which suggests that language is not required for the implicit representation of first-order false beliefs. Explicit verbal FB reasoning is acquired around the age of four, possibly with the help of language (de Villiers,

2005). In this study, we investigated second-order reasoning about false beliefs with a verbal and a non-verbal task. Children's success on the first-order false-belief items in the two tasks indicates that they were able to keep track of the different beliefs despite the complexity of the tasks with several protagonists, each with their own beliefs. Their performance on these first-order items contrasts with their much poorer performance on the second-order items. Importantly, children pass a verbal second-order task before they pass a non-verbal second-order task. We thus find a strong effect of language at the second-order level, as the verbal second-order FB task turned out to be easier than the corresponding non-verbal task. What does this suggest about the relation between language development and false-belief reasoning? And why is the verbal task easier for children?

A first possibility is that children's different performance on the two tasks is a task effect that does not relate to their capacity for FB reasoning. Retrieval of the relevant belief representations from memory in order to respond to the test question may be aided by the manner in which these representations have been processed and stored during the task. In the verbal task the story was presented verbally. Moreover there were probe questions at various moments throughout the story which were designed to prompt participants to verbalize their knowledge about false beliefs. In the non-verbal task, on the other hand, the story was presented purely visually and no probe questions were asked. It is conceivable that the explicit nature of the verbal task makes this task easier for children. However, this is at odds with the observed difference between toddler's early good performance on implicit false-belief tasks and young children's difficulty with explicit false-belief tasks (see Section 1).

An alternative explanation in terms of task effects is that the mismatch between mode of presentation and mode of response may make the non-verbal task more difficult for children. In both the verbal and the non-verbal task a verbal response was required in response to the test question. If this explanation is correct, we predict that children will perform better on a non-verbal task if performance is measured by looking behavior, as one can do in an eye-tracking study. Also, if this explanation is correct, it is expected that in implicit higher-order reasoning tasks (such as Meijering, Van Maanen, Van Rijn & Verbrugge's (2010) marble task) that do not rely on language either in their presentation or for the response, children may not experience the same difficulty as in our non-verbal task. However, even for adults implicit second-order reasoning does not come readily, as Hedden and Zhang (2002) have shown.

An third possibility is that children perform better on the verbal task because grammatical representations help them in their FB reasoning. Second-order FB reasoning requires embedding of beliefs in a way similar to how language structures syntactic embedding, The recursive linguistic representations involved in syntactic embedding may therefore provide the scaffolding to perform the recursive step of a second-order FB reasoning task (Hollebrandse & Roeper, submitted). This explanation extends the ideas of De Villiers (2005) about the role of language in acquiring explicit FB representations to second-order FB reasoning (see also Hollebrandse, 2000).

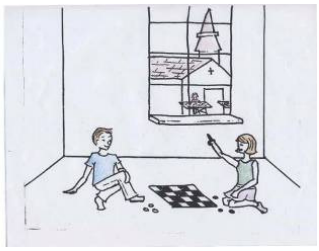
We conclude that a verbal second-order FB task is easier for children than a corresponding non-verbal FB task. This suggests that language supports explicit reasoning about beliefs, perhaps by facilitating the cognitive system that keeps track of beliefs attributed by people to other people.

References

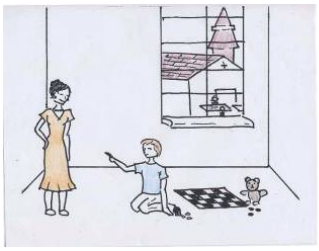
1. Apperly, I.A., Butterfill, S.A.: Do humans have two systems to track beliefs and belief-like states? American Psychological Association 116, 953-970 (2009)
2. Astington, J. W.: The child's discovery of the mind. Harvard University Press, Cambridge (1993)
3. Bruin, L. de: An association based account of false belief understanding. Ms. University of Bochum (2011)
4. De Villiers, J.G.: Can language acquisition give children a point of view? In: Astington, J., Baird, J. (eds.) Why Language Matters for Theory of Mind. pp. 186-219. Oxford Press, Oxford (2005)
5. Hedden, T., Zhang, J.: What do you think I think you think?: Strategic reasoning in matrix games. Cognition 85, 1-36 (2002)
6. Hollebrandse, B, Roeper, T.: Recursion and propositional exclusivity. submitted.
7. Hollebrandse, B., Hobbs, K., De Villiers, J.G., Roeper, T.: Second order embedding and second order false belief. In: Gavarro, A., Freitas, M.J. (eds.) Language Acquisition and Development, Proceedings of GALA 2007. pp. 270-280. Cambridge Scholar Press, Cambridge (2008)
8. Hollebrandse, B.: The Acquisition of Sequence of Tense. University of Massachusetts dissertation (2000)
9. Meijering, B., Maanen, L. van, Rijn, H. van, Verbrugge, R.: The facilitative effect of context on second-order social reasoning. In: Catrambone, R., Ohlsson, R. (eds.), Proceeding of the 32nd Annual conference of the Cognitive Science Society, pp. 1423-1428. Cognitive Science Society, Austin (2010)
10. Onishi, K.H., Baillargeon, R.: Do 15-month-old infants understand false beliefs? Science 308, 255-258 (2005)
11. Perner, J., Wimmer, H.: "John thinks that Mary thinks that …." Attribution of second order beliefs by 5- to 10-year-old children. Journal of Experimental Child Psychology 39, 437-471(1985)
12. Southgate, V., Senju, A., Csibra, G.: Action anticipation through attribution of false belief by 2-year-olds. Psychological Science 17:7, 587-592 (2007)
13. Sullivan, K., Zaitchik, D. Tager-Flusberg, H.: Preschoolers can attribute second order beliefs. Developmental Psychology 30, 395-402 (1994)
14. Wellman, H. M., & Bartsch, K.: Young children's reasoning about beliefs. Cognition 30, 239–277 (1988)
15. Wimmer, H., Perner, J.: (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13, 103–128 (1983)

Appendix 1: Illustration of verbal false belief task: the *Bake Sale* Story

Abbreviations: Q1FB1 = first first-order false-belief question; Q2FB1 = second first-order false -belief question; QFB2 = second-order false-belief question.

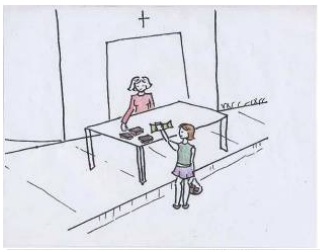


Sam and Maria are playing together. They look outside and see that the church is having a bake sale. Maria tells Sam: "I am going to buy chocolate chip cookies for us there," and she walks away.



Mom comes home and she tells Sam that she just drove past the bake sale. "Are they selling chocolate chip cookies?" Sam asks. "No," mum says, "they are only selling pumpkin pie." "Maria will now probably get pumpkin pie at the bake sale," Sam says.

Probe 1: Does Maria know they are selling pumpkin pie at the bake sale?



Maria has arrived at the bake sale. "I would like to buy chocolate chip cookies," she says. "All we have left are brownies," says the lady behind the stall. Since Maria also likes brownies, she decides to get some brownies.

Probe 2: Does Sam know that Maria bought some brownies?

Q1FB1: What does Sam think they are selling at the bake sale? Why does he think that?

On her way back, Maria meets the mailman. She tells the mailman: "I have just bought some brownies. I am going to share them with my brother Sam. It is a surprise". "That is nice of you," says the mailman. Then he asks Maria: "Does Sam know what you bought him?"

Ignorance: What does Maria tell the mailman?

Then the mailman asks: "What does Sam think they are selling at the bake sale?"

QFB2: What does Maria tell the mailman? Why does she say that?
Q2FB1: What does Sam think they are selling at the bake sale? Why does he think that?

Appendix 2: Illustration of non-verbal false-belief task (Unknown-Change-of-Content set-up)

Abbreviations: Q1FB1 = first first-order false-belief question; Q2FB1 = second first-order false-belief question; QFB2 = second-order false-belief question.

a. Non-verbal first-order task

First order trial

Q1FB1: Remember you are the player. Now, what does she think is in the box? (experimenter points to the woman)

b. Non-verbal second-order task

Second order trial

QFB2: Remember, first you were the player, but now he (experimenter points to the man in the right window) is the player and we are going to ask him the same question as we asked you: "What does she think is in the box?" What will he answer?

Q2FB1: What does she herself think is in the box?

Examples of non-verbal movies can be found at:
http://www.let.rug.nl/hollebr/FB-movies/1stOrderNonVerbal.wmv
http://www.let.rug.nl/hollebr/FB-movies/2ndOrderNonVerbal.wmv

A Dynamic Analysis of Interactive Rationality

Eric Pacuit[1] Olivier Roy[2]

[1] Tilburg Institute for Logic and Philosophy of Science, e.j.pacuit@uvt.nl
[2] Center for Mathematical Philosophy, LMU, Olivier.Roy@lrz.uni-muenchen.de

Abstract. Epistemic game theory has shown the importance of informational contexts in understanding strategic interaction. We propose a general framework to analyze how such contexts may arise. The idea is to view informational contexts as the fixed-points of iterated, "rational responses" to incoming information about the agents' possible choices. We show general conditions for the stabilization of such sequences of rational responses, in terms of structural properties of both the decision rule and the information update policy.

1 Background and Motivation

An increasingly popular[3] view is that "*the* fundamental insight of game theory [is] that a rational player must take into account that the players reason about each other in deciding how to play" [6, pg. 81]. Exactly *how* the players (should) incorporate the fact that they are interacting with other (actively reasoning) agents into their own decision making process is the subject of much debate. A variety of frameworks explicitly model the *reasoning* of rational agents in a strategic situation. Key examples include Brian Skyrms' models of "dynamic deliberation" [32], Ken Binmore's analysis of "eductive reasoning" [11], and Robin Cubitt and Robert Sugden's "common modes of reasoning" [17]. Although the details of these frameworks are quite different they share a common line of thought: In contrast to classical game theory, *solution concepts* are no longer the basic object of study. Instead, the "rational solutions" of a game are the result of individual (rational) decisions in specific informational "contexts".

This perspective on the foundations of game theory is best exemplified by the so-called epistemic program in game theory (cf. [15]). The central thesis here is that the basic mathematical model of a game should include an explicit parameter describing the players' *informational attitudes*. However, this broadly decision-theoretic stance does not simply *reduce* the question of decision-making in interaction to that of rational decision making in the face of uncertainty or ignorance. Crucially, *higher-order* information (belief about beliefs, etc.) are key components of the informational context of a game[4]. Of course, different contexts

[3] But, of course, not uncontroversial. See, for example, [22, pg. 239].

[4] That is, strategic behavior *depends*, in part, on the players' higher-order beliefs. However, the question of what precisely is being claimed should be treated with some care. The well-known *email game* of Ariel Rubinstein [30] demonstrates that

of a game can lead to drastically different outcomes, but this means that the informational contexts themselves are open to rational criticism:

> "It is important to understand that we have two forms of irrationality [...]. For us, a player is rational if he optimizes and also rules nothing out. So irrationality might mean not optimizing. But it can also mean optimizing while not considering everything possible." [16, pg. 314]

Thus, a player can be rationally criticized for not choosing what is *best given their information*, but also for not reasoning *to* a "proper" context. Of course, what counts as a "proper" context is debatable. There might be rational pressure for or against making certain *substantive assumptions*[5] about the beliefs of one's opponents, for instance, always entertaining the possibility that one of the players might not choose optimally.

Recently, researchers using methods from dynamic-epistemic logic have taken steps to understanding this idea of reasoning *to* a "proper" or "rational" context [10, 9, 8, 36]. Building on this literature[6], we provide a general characterization of when players can or cannot rationally reason to an informational context.

2 Belief Dynamics for Strategic Games

Our goal is to understand well-known solutions concepts, not in terms of fixed informational contexts—for instance, models (e.g., type spaces or epistemic models) satisfying rationality and common belief of rationality—but rather as a result of a dynamic, interactive process of "information exchanges". It is important to note that we do *not* see this work as an attempt to represent some type of "preplay communication" or form of "cheap talk". Instead, the idea is to represent the process of *rational deliberation* that takes the players from the *ex ante* stage to the *ex interim* stage of decision making. Thus, the "informational exchanges" are the result of the players' *practical reasoning* about what they should do, given their current beliefs. This is in line with the current research program using dynamic epistemic and doxastic logics to analyze well-known solution concepts (cf. [2, 9, 10] where the "rationality announcements" do not capture any type of communication between the players, but rather internal observations about which outcomes of the game are "rational").

misspecification of arbitrarily high-orders of beliefs can have a great impact on (predicted) strategic behavior. So there are simple examples where (predicted) strategic behavior is *too sensitive* to the players' higher-order beliefs. We are not claiming that a rational agent is *required* to consider *all* higher-order beliefs, but only that a rational player recognizes that her opponents are actively reasoning, rational agents, which means that a rational player does take into account *some* of her higher-order beliefs (e.g., what she believes her opponents believe she will do) as she deliberates. Precisely "how much" higher-order information should be taken into account is a very interesting, open question which we set aside in this paper.

[5] The notion of substantive assumption is explored in more detail in [29].

[6] The reader not familiar with this area can consult the recent textbook [35] for details.

2.1 Describing an Informational Context

Let $G = \langle N, \{S_i\}_{i \in N}, u_i \rangle$ be a strategic game (where N is the set of players and for each $i \in N$, S_i is the set of actions for player i and $u_i : \Pi_i S_i \to \mathbb{R}$ is a utility function).[7] The informational context of a game describes the players' *hard* and *soft* information about the possible outcomes of the game. Many different formal models have been used to represent an informational context of a game (for a sample of the extensive literature, see [13, 10] and references therein). In this paper we employ one such model: a *plausibility structure* consisting of a set of states and a single plausibility ordering (which is reflexive, transitive and connected) $w \preceq v$ that says "v is at least as plausible as w." Originally used as a semantics for conditionals (cf. [24]), these *plausibility models* have been extensively used by logicians [34, 35, 8], game theorists [12] and computer scientists [14, 23] to represent rational agents' (all-out) beliefs. We thus take for granted that they provide a natural model of beliefs in games:

Definition 1. *Let $G = \langle N, \{S_i\}_{i \in N}, u_i \rangle$ be a strategic form game. An* **informational context** *of G is a plausibility model $\mathcal{M}_G = \langle W, \preceq, \sigma \rangle$ where $\preceq$ is a connected, reflexive, transitive and well-founded*[8] *relation on W and σ is a* **strategy function***: a function $\sigma : W \to \Pi_i S_i$ assigning strategy profiles to each state. To simplify notation, we write $\sigma_i(w)$ for $(\sigma(w))_i$ (similarly, write $\sigma_{-i}(w)$ for the sequence of strategies of all players except i).*

A few comments about this definition are in order. First of all, note that there is only one plausibility ordering in the above models, yet we are interested in games with more than one player. There are different ways to interpret the fact that there is only one plausibility ordering. One is that the models represent the beliefs of a single player before she has made up her mind about which option to choose in the game. A second interpretation is to think of a model as representing the modeler's or game theorist's point of view about which outcomes are more or less plausible given the reasoning of the players. Thus, a model describes a stage of the rational deliberation of *all* the players starting from an initial model where the players have the same beliefs (i.e., the *common prior*). The private information about which outcomes the *players* consider possible given their actual choice can then be defined from the *conditional beliefs*.[9] Our second comment on the above definition is that since we are representing the rational

[7] We assume the reader is familiar with the basic concepts of game theory. For example, strategic games and various solution concepts, such as iterated removal of strictly (weakly) dominated strategies.

[8] Well-foundedness is only needed to ensure that, for any set X, the set of minimal elements in X is nonempty. This is important only when W is infinite – and there are ways around this in current logics. Moreover, the condition of connectedness can also be lifted, but we use it here for convenience.

[9] The suggestion here is that one can define a partition model á la Aumann [5] from a plausibility model. Working out the details is left for future work, but we note that such a construction blurs the distinction between so-called *belief*-based and *knowledge*-based analyses of solution concepts (cf. the discussion in [15]).

deliberation process, we do not assume that the players have made up their minds about which actions they will choose. Finally, note that the strategy functions need not be onto. Thus, the model represents the player's(s') opinions about which outcomes of the game are more or less plausible *among the ones that have not been ruled out*.

Of course, this model can be (and has been: see [8, 35]) extended to include beliefs for each of the players, an explicit relation representing the player(s) hard information or by making the plausibility orders state-dependent. In order to keep things simple we focus on models with a single plausibility ordering.

We conclude this brief introduction to plausibility models by giving the well-known definitions of a conditional belief. For $X \subseteq W$, let $Min_{\preceq}(X) = \{v \in X \mid v \preceq w \text{ for all } w \in X\}$ be the set of minimal elements of X according to $\preceq$.

Definition 2 (Belief and Conditional Belief). *Let $\mathcal{M}_G = \langle W, \preceq, \sigma \rangle$ be a model of a game G. Let E and F be subsets of W, we say:*

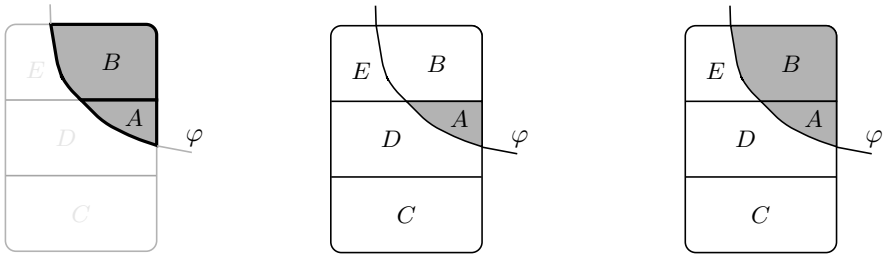
- *E is* **believed conditional on** *F in $\mathcal{M}_G$ provided $Min_{\preceq}(F) \subseteq E$.*

Also, we say E is **believed** *in $\mathcal{M}_G$ if E is believed conditional on W. Thus, E is believed provided $Min_{\preceq}(W) \subseteq E$*

2.2 A Primer on Belief Dynamics

We are not interested in informational contexts *per se*, but rather how the informational context changes during the process of rational deliberation. The type of change we are interested in is how a model $\mathcal{M}_G$ of a game G incorporates new information about what the players *should* do (according to a particular choice rule). As is well known from the belief revision literature, there are many ways to transform a plausibility model given some new information [28]. We do not have the space to survey the entire body of relevant literature here (cf., [35, 7]). Instead we sketch some key ideas, assuming the reader is already familiar with this approach to belief revision.

The general approach is to define a way of *transforming* a plausibility model $\mathcal{M}_G$ given a proposition φ. A transformation τ maps plausibility models and propositions to plausibility models (we write $\mathcal{M}_G^{\tau(\varphi)}$ for $\tau(\mathcal{M}_G, \varphi)$). Different definitions of τ represent the different attitudes an agent can take to the incoming information. The picture below provides three typical examples:



$!(\varphi) : A \prec B$ $\quad\quad$ $\uparrow(\varphi) : A \prec C \prec D \prec B \cup E$ $\quad\quad$ $\Uparrow(\varphi) : A \prec B \prec C \prec D \prec E$

The operation on the left is the well-known *public announcement* operation [25, 19], which assumes that the source of φ is *infallible*, ruling out any possibilities that are inconsistent with φ. For the other transformations, while the players do *trust* the source of φ, they do not treat the source as infallible. Perhaps the most ubiquitous policy is *conservative upgrade* ($\uparrow\varphi$), which allows the player(s) only tentatively to accept the incoming information φ by making the best φ-worlds the new minimal set while keeping the old plausibility ordering the same on all other worlds. The operation on the right, *radical upgrade* ($\Uparrow\varphi$), is stronger, moving *all* φ worlds before all the $\neg\varphi$ worlds and otherwise keeping the plausibility ordering the same. These dynamic operations satisfy a number of interesting logical principles [35, 7], which we do not discuss further here.

We are interested in the operations that transform the informational context as the players deliberate about what they should do in a game situation. In each informational context (viewed as describing one stage of the deliberation process), the players determine which options are "*rationally permissible*" and which options the players ought to avoid (which is guided by some fixed choice rule). This leads to a transformation of the informational context as the players adopt the relevant beliefs about the outcome of their *practical reasoning*. The different types of transformation mentioned above then represent how confident the player(s) (or modeler) is (are) in the assessment of which outcomes are rational. In this new informational context, the players again think about what they should do, leading to another transformation. The main question is does this process *stabilize*?

The answer to this question will depend on a number of factors. The general picture is

$$\mathcal{M}_0 \overset{\tau(D_0)}{\Longrightarrow} \mathcal{M}_1 \overset{\tau(D_1)}{\Longrightarrow} \mathcal{M}_2 \overset{\tau(D_2)}{\Longrightarrow} \cdots \overset{\tau(D_n)}{\Longrightarrow} \mathcal{M}_{n+1} \Longrightarrow \cdots$$

where each D_i is some proposition and τ is a model transformer. Two questions are important for the analysis of this process. First, what type of transformations are the players using? For example, if τ is a public announcement, then it is not hard to see that, for purely logical reasons, this process must eventually stop at a limit model (see [8] for a discussion and proof). The second question is where do the propositions D_i come from? To see why this matters, consider the situation where you iteratively perform a radical upgrade with p and $\neg p$ (i.e., $\Uparrow(p), \Uparrow(\neg p), \ldots$). Of course, this sequence of upgrades never stabilizes. However, in the context of reasoning about what to do in a game situation, this situation may not arise thanks to special properties of the choice rule that is being used to describe (or guide) the players' decisions.

2.3 Deliberating about What to Do

It is not our intention to have the dynamic operations of belief change discussed in the previous section directly represent the players' (practical) *reasoning*. Instead, we treat practical reasoning as a "black box" and focus on general *choice rules* that are intended to describe rational decision making (under ignorance). To make this precise, we need some notation:

Definition 3 (Strategies in Play). *Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a strategic game and $\mathcal{M}_G = \langle W, \preceq, \sigma\rangle$ an informational context of G. For each $i \in N$, the strategies* **in play for** *i is the set*

$$S_{-i}(\mathcal{M}_G) = \{s_{-i} \in \Pi_{j\neq i} S_j \mid \text{ there is } w \in Min_{\preceq}(W) \text{ with } \sigma_{-i}(w) = s_{-i}\}$$

This set $S_{-i}(\mathcal{M}_G)$ is the set of strategies that are believed to be available for player i at some stage of the deliberation process represented by the model $\mathcal{M}_G$. Given $S_{-i}(\mathcal{M}_G)$, different choice rules offer recommendations about which options to choose. There are many choice rules that could be analyzed here (e.g., strict dominance, weak dominance or admissibility, minimax, minmax regret, etc.). For the present purposes we focus primarily on weak dominance (or admissibility), although our main theorem in Seciton 3 applies to all choice rules.

Weak Dominance (pure strategies[10]) Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a strategic game and $\mathcal{M}_G$ an model of G. For each i and $a \in S_i$, put $a \in S_i^{wd}(\mathcal{M}_G)$ provided there is $b \in S_i$ such that for all $s_{-i} \in S_{-i}(\mathcal{M}_G)$, $u_i(s_{-i}, b) \geq u_i(s_{-i}, a)$ and there is some $s_{-i} \in S_{-i}(\mathcal{M}_G)$ such that $u_i(s_{-i}, b) > u_i(s_{-i}, a)$.

So an action a is weakly dominated for player i if it is weakly dominated with respect to all of i's available actions and the (joint) strategies believed to be still in play for i's opponents.

More generally, we assume that given the beliefs about which strategies are in play the players *categorize* their available options (i.e., the set S_i) into "good" (or "rationally permissible") strategies and those strategies that are "bad" (or "irrational"). Formally, a **categorization** for player i is a pair $\mathbf{S}_i(\mathcal{M}_G) = (S_i^+, S_i^-)$ where $S_i^+ \cup S_i^- \subseteq S_i$. (We write $\mathbf{S}_i(\mathcal{M}_G)$ to signal that the categorization depends on current beliefs about which strategies are in play.) Note that, in general, a categorization need not be a partition (i.e., $S_i^+ \cup S_i^- \neq S_i$) . See [18] for an example of such a categorization algorithm. However, in the remainder of this paper we focus on familiar choice rules where the categorization does form a partition. For example, for weak dominance we let $S_i^- = S_i^{wd}(\mathcal{M}_G)$ and $S_i^+ = S_i - S_i^-$.

Given a model of a game $\mathcal{M}_G$ and for each player i a categorization is $\mathbf{S}_i(\mathcal{M}_G)$; the next step is to incorporate this information into $\mathcal{M}_G$ using some model transformation. We start by introducing a simple propositional language to describe a categorization.

Definition 4 (Language for a Game). *Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a strategic game. Without loss of generality, assume that each of the S_i is disjoint and let $\mathsf{At}_G = \{P_a^i \mid a \in S_i\}$ be a set of atomic formulas (one for each $a \in S_i$). The propositional language for G, denoted $\mathcal{L}_G$, is the smallest set of formulas containing At_G and closed under the Boolean connectives $\neg$ and $\wedge$.*

Formulas of $\mathcal{L}_G$ are intended to describe possible outcomes of the game. Given an informational context of a game $\mathcal{M}_G$, the formulas $\varphi \in \mathcal{L}_G$ is can be associated with subsets of the set of states in the usual way:

[10] This definition can be modified to allow for dominance by mixed strategies, but we leave issues about how to incorporate probabilities to another occasion.

Definition 5. *Let G be a strategic game, $\mathcal{M}_G = \langle W, \preceq, \sigma\rangle$ an informational context of G and $\mathcal{L}_G$ a propositional language for G. We define a map $[\![\cdot]\!]_{\mathcal{M}_G} : \mathcal{L}_G \to \wp(W)$ by induction as follows: $[\![P_a^i]\!]_{\mathcal{M}_G} = \{w \mid \sigma(w)_i = a\}$, $[\![\neg\varphi]\!]_{\mathcal{M}_G} = W - [\![\varphi]\!]_{\mathcal{M}_G}$ and $[\![\varphi \wedge \psi]\!]_{\mathcal{M}_G} = [\![\varphi]\!]_{\mathcal{M}_G} \cap [\![\psi]\!]_{\mathcal{M}_G}$.*

Using the above language, for each informational context of a game $\mathcal{M}_G$, we can define $Do(\mathcal{M}_G)$, which describes what the players are going to do according to a fixed categorization procedure. To make this precise, suppose that $\mathbf{S}_i(\mathcal{M}_G) = (S_i^+, S_i^-)$ is a categorization for each i and define:

$$Do_i(\mathcal{M}_G) \;:=\; \bigvee_{a \in S_i^+} P_i^a \wedge \bigwedge_{b \in S_i^-} \neg P_i^b$$

Then, let $Do(\mathcal{M}_G) = \bigwedge_i Do_i(\mathcal{M}_G)$.[11]

The general project is to understand the interaction between types of categorizations (eg., choice rules) and types of model transformations (representing the rational deliberation process). One key question is: Does a deliberation process *stabilize*(and if so, under what conditions)? (See [8] for general results here.) In this paper there are two main reasons why an upgrade stream would stabilize. The first is from properties of the transformation. The second is because the choice rule satisfies a monotonicity property so that, eventually, the categorizations stabilize and no new transformations can change the plausibility ordering. We are now ready to give a formal definition of a "deliberation sequence":

Definition 6 (Deliberation Sequence). *Given a game G and an informational context $\mathcal{M}_G$, a deliberation sequence of type τ (which we also call an upgrade sequence), induced by $\mathcal{M}_G$ is an infinite sequence of plausibility models $(\mathcal{M}_m)_{m\in\mathbb{N}}$ defined as follows:*

$$\mathcal{M}_0 = \mathcal{M}_G \qquad \mathcal{M}_{m+1} = \tau(\mathcal{M}_m, Do(\mathcal{M}_m))$$

An upgrade sequence **stabilizes** *if there is an $n \geq 0$ such that $\mathcal{M}_n = \mathcal{M}_{n+1}$.*

3 Case Study: Iterated Admissibility

A key issue in the epistemic foundations of game theory is the epistemic analysis of iterated removal of *weakly* dominated strategies. Many authors have pointed out puzzles surrounding such an analysis [4, 31, 16]. For example, Samuelson [31] showed (among other things) that "common knowledge of admissibility" may be an inconsistent concept (in the sense that there is a game which does not have a model with a state satisfying 'common knowledge of rationality' [31, Example 8, pg. 305]).[12] This is illustrated by the following game:

[11] There are other ways to describe a categorization, but we leave this for further research.

[12] Compare with strict dominance: it is well known that common knowledge that players do not play weakly dominated strategies *implies* that the players choose a strategy profile that survives iterated removal of strictly dominated strategies.

		Bob	
		L	R
Ann	u	$1,1$	$1,0$
	d	$1,0$	$0,1$

The key issue is that the assumption that players only play *admissible* strategies conflicts with the logic of iteratively removing strategies deemed "irrational". The general framework introduced above offers a new, dynamic perspective on this issue, and on reasoning with admissibility more generally.[13] Dynamically, Samuelson's non-existence result corresponds to the fact that the players' rational upgrade streams do not stabilize. That is, the players are not able to deliberate their way to a stable, common belief in admissibility. In order to show this we need the "right" notion of model transformation.

Our first observation is that the model transformations we discussed in Section 2.2 do not explain Samuelson's result.

Observation 1 Suppose that the categorization method is weak dominance and that $Do(\mathcal{M})$ is defined as above. For each of the model transformations discussed in Section 2.2 (i.e., public announcement, radical upgrade and conservative upgrade), any deliberation sequence for the above game stabilizes.

The proof of this Observation is straightforward since the language used to describe the categorization does not contain belief modalities[14]. This observation is nice, but it does not explain the phenomena noticed by Samuelson [31]. The problem lies in the way we incorporate information when there is more than one element of $S_i^+(\mathcal{M})$ for some agent i.

It is well known that, in general, there are no rational principles of decision making (under ignorance or uncertainty) which *always* recommend a *unique* choice. In particular, it is not hard to find a game and an informational context where there is at least one player without a *unique* "rational choice". How should a rational player incorporate the information that more than one action is classified as "choice-worthy" or "rationally permissible" (according to some choice rule) for her opponent(s)? Making use of a well-known distinction due to Edna Ullmann-Margalit and Sidney Morgenbesser [33], the assumption that all players are rational can help determine which options the player will *choose*, but rationality alone does not help determine which of the rationally permissible options will be "picked"[15]. What interests us is how to transform a plausibility

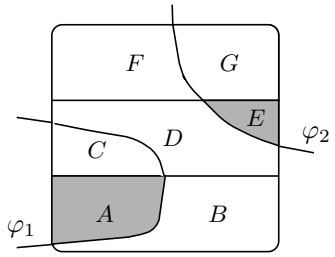
[13] We do not provide an alternative epistemic characterization of this solution concept. Both [16] and [20] have convincing results here. Our goal is to use this solution concept as an illustration of our general approach.

[14] An interesting extension would be to start with a multiagent belief model and allow players not only to incorporate information about which options are "choice-worthy", but also what beliefs their opponents may have. We leave this extension for future work, focusing here on setting up the basic framework.

[15] This line of thought led Cubitt and Sugden to impose a "privacy of tie breaking" property which says that players cannot *know* that her opponent will not pick an

model to incorporate the fact that there is a *set* of choice-worthy options for (some of) the players.

We suggest that a generalization of *conservative upgrade* is the notion we are looking for (see [21] for more on this operation). The idea is to do an upgrade with a *set* of propositions $\{\varphi_1, \ldots, \varphi_n\}$ by letting the most plausible worlds be the union of each of the most plausible φ_i worlds:



$$\uparrow\{\varphi_1, \varphi_2\} : A \cup E \prec B \prec C \cup D \prec F \cup G$$

We do not give the formal definition here, but it should be clear from the example given above. It is not hard to see that this is not the same as $\uparrow\varphi_1 \vee \cdots \vee \varphi_n$, since, in general, $Min_{\preceq}([\![\varphi_1]\!] \cup \cdots \cup [\![\varphi_n]\!]) \neq \bigcup_i Min_{\preceq}([\![\varphi_i]\!])$. We must modify our definition of $Do(\mathcal{M})$: for each $i \in N$ let:

$$Do_i(\mathbf{S}_i(\mathcal{M}_G)) = \{P_a^i \mid a \in \mathbf{S}_i^+(\mathcal{M}_G)\} \cup \{\neg P_b^i \mid b \in \mathbf{S}_i^-(\mathcal{M}_G)\}$$

Then define $Do(\mathbf{S}(\mathcal{M}_G)) = Do_i(\mathbf{S}_i(\mathcal{M}_G)) \bigwedge Do_2(\mathbf{S}_2(\mathcal{M}_G)) \cdots \bigwedge Do_n(\mathbf{S}_n(\mathcal{M}_G))$, where if X and Y are two sets of propositions, then let $X \wedge Y := \{\varphi \wedge \psi \mid \varphi \in X, \psi \in Y\}$.

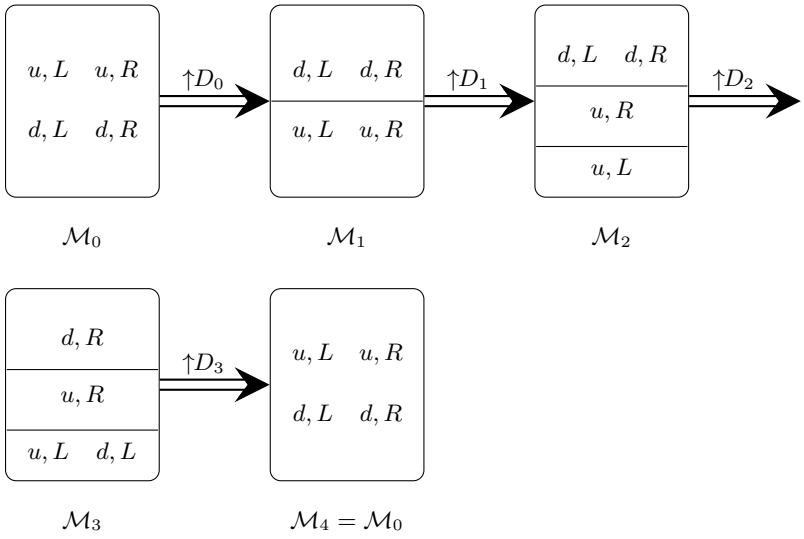
Observation 2 Suppose that the categorization method is weak dominance as explained in Section 2.3 and that $Do(\mathcal{M})$ is defined as above. Then, starting with the initial full model of the above game,[16] a generalized conservative upgrade stream does not stabilize.

The following upgrade stream illustrates this observation:

---

option that is classified as "choice-worthy" [17, pg. 8] (cf. also [4]'s "no extraneous restrictions on beliefs" property). Wlodeck Rabinovich takes this even further and argues that from the principle of indifference, players must assign equal probability to all choice-worthy options [27].

[16] A full model is one where it is common knowledge that each outcome of the game is equally plausible.



Intuitively, from $\mathcal{M}_0$ to $\mathcal{M}_2$ the agents have reasons to exclude d and R, leading them to the common belief that u, L is played. At that stage, however, d is admissible for Ann, canceling the reason the agents had to rule out this strategy. The rational response here is thus to suspend judgment on d, leading to $\mathcal{M}_3$. In this new model the agents are similarly led to suspend judgment on not playing R, bringing them back to M_0. This process loops forever: the agents' reasoning does not stabilize.

A corollary of this observation is that common belief in admissibility is not sufficient for the stabilization of upgrade streams. Stabilization also requires that all *and only* those profiles that are most plausible are admissible.

4 Stabilization Theorem

In this section we informally state and discuss a number of abstract principles which guarantee that a rational deliberation sequence will *stabilize*. The principles ensure that the categorizations are "sensitive" to the players' beliefs and that the players respond to the categorizations in the appropriate way.

We start by fixing some notation. Let U be a fixed set of states and G a fixed strategic game. We confine our attention to transformations between models of G whose states come from the universe of states U. Let $\mathbb{M}_G$ be the set of all such plausibility models. A model transformation is then a function that maps a model of G and a finite set of formulas of $\mathcal{L}_G$ to a model in $\mathbb{M}_G$:

$$\tau : \mathbb{M}_G \times \wp_{<\omega}(\mathcal{L}_G) \to \mathbb{M}_G$$

where $\wp_{<\omega}(\mathcal{L}_G)$ is the set of finite subsets of $\mathcal{L}_G$. Of course, not all transformations τ make sense in this context.

The first set of principles that τ must satisfy ensure that the categorizations and belief transformation τ are connected in the "right way". One natural property is that the belief transformations treat *equivalent* formulas the same way. A second property we impose is that receiving exactly the same (ground) information twice does not have any effect on the players' beliefs. These are general properties of the belief transformation. Certainly, there are other natural properties that one may want to impose (for example, variants of the AGM postulates [1]), but for now we are interested in the minimal principles needed to prove a stabilization result.

The next set of properties ensure that the transformations respond "properly" to a categorization. First, we need a property to guarantee that the categorizations depend only on the players' beliefs. Second, we need to ensure that all upgrade sequences respond to the categorizations in the right way:

C2$^-$ For any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$ in τ, if $a \in S_i^-(\mathcal{M}_n)$ then $\neg P_i^a$ is believed in $\mathcal{M}_{n+1}$.

C2$^+$ For any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$ in τ, if $a \in S_i^+(\mathcal{M}_n)$ then $\neg P_i^a$ is not believed in $\mathcal{M}_{n+1}$

Finally, we need to assume that the categorizations are monotonic:

Mon$^-$ For any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$, for all $n \geq 0$, for all players $i \in N$, $\mathbf{S}_i^-(\mathcal{M}_n) \subseteq \mathbf{S}_i^-(\mathcal{M}_{n+1})$

Mon$^+$ Either for all models $\mathcal{M}_G$, $\mathbf{S}_i^+(\mathcal{M}_G) = S_i - \mathbf{S}_i^-(\mathcal{M}_G)$ or for any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$, for all $n \geq 0$, for all players $i \in N$, $\mathbf{S}_i^+(\mathcal{M}_n) \subseteq \mathbf{S}_i^+(\mathcal{M}_{n+1})$

In particular, **Mon**$^-$ means that once an option for a player is classified as "not rationally permissible", it cannot drop this classification at a later stage of the deliberation process.

Theorem 3. *Suppose that G is a finite game and all of the above properties are satisfied. Then every upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$ stabilizes.*

The proof can be found in the full version of the paper. The role of monotonicity of the choice has been noticed by a number of researchers (see [3] for a discussion). This theorem generalizes van Benthem's analysis of rational dynamics [10] to soft information, both in terms of attitudes and announcements. It is also closely related to the result in [3] (a complete discussion can be found in the full paper).

5 Concluding remarks

In this paper we have proposed a general framework to analyze how "proper" informational contexts my arise. We have provided general conditions for the stabilization of deliberation sequences in terms of structural properties of both the

decision rule and the information update policy. We have also applied the framework to admissibility, giving a dynamic analysis of Samuelson's non-existence result.

Throughout the paper we have worked with (logical) models of *all out* attitudes, leaving aside probabilistic andgraded beliefs, even though the latter are arguably most widely used in the current literature on epistemic foundations of game theory. It is an important but non-trivial task to transpose the dynamic perspective on informational contexts that we advocate here to such probabilistic models. This we leave for future work.

Finally, we stress that the dynamic perspective on informational contexts is a natural complement and not an alternative to existing epistemic characterizations of solution concepts [37], which offer rich insights into the consequences of taking seriously the informational contexts of strategic interaction. What we have proposed here is a first step towards understanding how or why such contexts might arise.

References

1. Alchourrón, C. E., Gärdenfors, P., and Makinson, D. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic 50* (1985), 510 – 530.
2. Apt, K., and Zvesper, J. Public announcements in strategic games with arbitrary strategy sets. In *Proceedings of LOFT 2010* (2010).
3. Apt, K., and Zvesper, J. The role of monotonicity in the epistemic analysis of strategic games. *Games 1*, 4 (2010), 381 – 394.
4. Asheim, G., and Dufwenberg, M. Admissibility and common belief. *Game and Economic Behavior 42* (2003), 208 – 234.
5. Aumann, R. Interactive epistemology I: Knowledge. *International Journal of Game Theory 28* (1999), 263–300.
6. Aumann, R., and Dreze, J. Rational expectations in games. *American Economic Review 98* (2008), 72 – 86.
7. Baltag, A., and Smets, S. ESSLLI 2009 course: Dynamic logics for interactive belief revision. Slides available online at `http://alexandru.tiddlyspot.com/#%5B%5BESSLLI09%20COURSE%5D%5D`, 2009.
8. Baltag, A., and Smets, S. Group belief dynamics under iterated revision: Fixed points and cycles of joint upgrades. In *Proceedings of Theoretical Aspects of Rationality and Knowledge* (2009).
9. Baltag, A., Smets, S., and Zvesper, J. Keep 'hoping' for rationality: a solution to the backwards induction paradox. *Synthese 169* (2009), 301–333.
10. Benthem, J. v. Rational dynamics and epistemic logic in games. *International Game Theory Review 9*, 1 (2007), 13–45.
11. Binmore, K. Modeling rational players: Part I. *Economics and Philosophy 3* (1987), 179 – 214.
12. Board, O. Dynamic interactive epistemology. *Games and Economic Behavior 49* (2004), 49 – 80.
13. Bonanno, G., and Battigalli, P. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics 53*, 2 (June 1999), 149–225.

14. Boutilier, C. *Conditional Logics for Default Reasoning and Belief Revision*. PhD thesis, University of Toronto, 1992.
15. Brandenburger, A. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory 35* (2007), 465–492.
16. Brandenburger, A., Friedenberg, A., and Keisler, H. J. Admissibility in games. *Econometrica 76* (2008), 307–352.
17. Cubitt, R., and Sugden, R. Common reasoning in games: A Lewisian analysis of common knowledge of rationality. CeDEx Discussion Paper, 2011.
18. Cubitt, R., and Sugden, R. The reasoning-based expected utility procedure. *Games and Economic Behavior* (2011), In Press.
19. Gerbrandy, J. *Bisimulations on Planet Kripke*. PhD thesis, University of Amsterdam, 1999.
20. Halpern, J., and Pass, R. A logical characterization of iterated admissibility. In *Proceedings of the Twelfth Conference on Theoretical Aspects of Rationality and Knoweldge* (2009), A. Heifetz, Ed., pp. 146 – 155.
21. Holliday, W. Trust and the dynamics of testimony. In *Logic and Interaction Rationality: Seminar's Yearbook 2009* (2009), ILLC Technical Reports, pp. 147 – 178.
22. Kadane, J. B., and Larkey, P. D. Subjective probability and the theory of games. *Management Science 28*, 2 (1982), 113–120.
23. Lamarre, P., and Shoham, Y. Knowledge, certainty, belief and conditionalisation. In *Proceedings of the International Conference on Knoweldge Representation and Reasoning* (1994), pp. 415 – 424.
24. Lewis, D. *Counterfactuals*. Blackwell Publishers, Oxford, 1973.
25. Plaza, J. Logics of public communications. In *Proceedings, 4th International Symposium on Methodologies for Intelligent Systems* (1989), M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, and Z. Ras, Eds., pp. 201–216 (republished as [26]).
26. Plaza, J. Logics of public communications. *Synthese: Knowledge, Rationality, and Action 158*, 2 (2007), 165 – 179.
27. Rabinowicz, W. Tortous labyrinth: Noncooperative normal-form games between hyperrational players. In *Knowledge, Belief and Strategic Interaction* (1992), C. Bicchieri and M. L. D. Chiara, Eds., pp. 107 – 125.
28. Rott, H. Shifting priorities: Simple representations for 27 itereated theory change operators. In *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg* (2006), H. Lagerlund, S. Lindström, and R. Sliwinski, Eds., vol. 53 of *Uppsala Philosophical Studies*, pp. 359 – 384.
29. Roy, O., and Pacuit, E. Substantive assumptions and the existence of universal knowledge structures: A logical perspective. Under submission, 2010.
30. Rubinstein, A. The electronic mail game: A game with almost common knowledge. *American Economic Review 79* (1989), 385 – 391.
31. Samuelson, L. Dominated strategies and common knowledge. *Game and Economic Behavior 4* (1992), 284 – 313.
32. Skyrms, B. *The Dynamics of Raitonal Deliberation*. Harvard University Press, 1990.
33. Ullmann-Margalit, E., and Morgenbesser, S. Picking and choosing. *Social Research 44* (1977), 757 – 785.
34. van Benthem, J. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics 14*, 2 (2004), 129 – 155.
35. van Benthem, J. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2010.

36. van Benthem, J., and Gheerbrant, A. Game solution, epistemic dynamics and fixed-point logics. *Fund. Inform. 100* (2010), 1–23.
37. van Benthem, J., Pacuit, E., and Roy, O. Towards a theory of play: A logical perspective on games and interaction. *Games 2*, 1 (2011), 52–86.

The Power of Knowledge in Games

Rohit Parikh[1,2], Çağıl Taşdemir[2], and Andreas Witzel[3]

[1] Brooklyn College of the City University of New York, rparikh@gc.cuny.edu
[2] Graduate Center of the City University of New York
[3] Courant Institute, New York University

10AM
July 5, 2011

Abstract: We propose a theory of the interaction between knowledge and games. Epistemic game theory is of course a well developed subject [4,5]. But there is also a need for a theory of how some agents can *affect* the outcome of a game by affecting the knowledge which other agents have and thereby affecting their actions.

We concentrate on games of incomplete or imperfect information, and study how conservative, moderate, or aggressive players might play such games. We provide models for the behavior of a knowledge manipulator who seeks to manipulate the knowledge states of active players in order to affect their moves and to maximize her own payoff even while she herself remains inactive.

1 Introduction

It is a commonplace that what we do depends on what we know[4]. And the *theory of mind* (Premack and Woodruff [15]) predicts that we also know that what others do will depend on what *they* know.

Bill can copy the answers from a fellow student's exambook if he knows that the teacher is not looking. And Betty can shoplift if she knows that the store does not have cameras, and that the guard is distracted.[5]

[4] or believe. We will use the word 'knowledge' neutrally, being well aware that actions often proceed from false beliefs.

[5] Deception also occurs among non-human primates. As [11] note, "... chimpanzees, one of humans' two closest primate relatives, sometimes attempt to actively conceal things from others. Specifically, when competing with a human in three novel tests, eight chimpanzees, from their first trials, chose to approach a contested food item via a route hidden from the human's view (sometimes using a circuitous path to do so)." Note that we are not claiming that chimps actually have what is called a *theory of mind*. Merely that some of their behavior seems deceptive.

But we can also proceed one level up. If Bill wants to copy from Jack's answerbook, he might ask Betty to *distract* the teacher, perhaps by asking his permission to go to the bathroom. If the store does not want Betty to shoplift then it might install TV cameras which record what happens in the store.

This second level of *arranging* for some level of knowledge or ignorance in others, in order to influence their actions, seems not to be sufficiently investigated formally.

Manipulation of knowledge can happen in two different contexts. One might manipulate knowledge for a particular purpose. For instance if Jack and Ann are going on a picnic and do not want Betty to come along, they may simply not reveal the existence of the picnic to her. That way they can avoid the situation where she says, "Oh, can I come too?" and have to either put up with her presence or else offend her by saying no.

But manipulation can also happen in a more general context. For instance a university may reveal the email address of a student to a professor who is teaching a course which the student is taking; and yet, not reveal the email address to another professor even though there is no *specific reason* why this knowledge would be harmful in some way.

1.1 Our Model

In our model we have a number of active players as well as a knowledge manipulator (KM). The knowledge manipulator arranges for the players to have certain restricted amounts of knowledge, both about the situation and about the knowledge of the other players. But she makes no moves herself. When the game ends, all the players including KM receive payoffs.

As we show later, our games can be reduced to more familiar forms treating KM as yet another active player. We choose not to do that since the role of the manipulator in real life is different, whether we are speaking about Julian Assange revealing certain secret messages or the government of some country restricting access to the internet. Iago in Shakespeare's play Othello is also a knowledge manipulator, although what he supplies to Othello is false beliefs rather than knowledge. It is important that Othello trusts Iago rather than questioning his motives. So in this paper, we will assume that the active players do not concern themselves with the motives of KM.

1.2 Defining rationality

Suppose an agent is in a situation of uncertainty where it has to choose between two moves L and R but does not know for sure what the outcome will be with either choice. How will the agent choose?

One option is the maxmin route. The agent can choose L if the *worst* possible outcome with L is better than the *worst* outcome with R. We will describe such an agent as *conservative*. However, an ambitious agent may choose R if the *best* outcome under R is better than the *best* outcome with L. We will describe such an agent as *aggressive*.

It is clear then that in the same situation, an aggressive agent with the same preferences as a conservative one may still make a different choice. Some people never buy lottery tickets on the ground that the worst outcome under buying, namely losing one's money, is worse than the certain outcome under not buying. But those who do buy such tickets are clearly judging by the best outcome.

In most of this paper we assume that utilities are ordinal. In other words, between any two choices a, b, the agent may be neutral, prefer a or prefer b. Numbers can be assigned to a and b so that $u(a) < u(b)$ iff b is preferred to a. However, ordinal utilities are preserved by all order preserving transformations. If c is preferred to b and b to a (which we may write $c > b > a$) then there is no difference between utility assignments to a, b, c of 1, 2, 3 or 1, 2, 4 or 1, 3, 4. It is also generally assumed that comparing utilities between different players makes no sense.

If utilities are cardinal and a subjective probability is available, we could also use expected value as a measure. However, in this work our utilities will be ordinal, and the notion of expected utility will not be available to us.

In addition to conservative and aggressive players, we can also consider moderate players who try to find the middle way. The general issue is that a player in uncertainty is choosing between two sets (or sequences) of payoffs. The payoff with L is say, $a_1 > a_2 > ... > a_k$ and with R it is $b_1 > b_2 > ... > b_m$. A conservative player chooses L over R iff a_k is preferred to b_m. An aggressive player chooses R over L iff b_1 is better than a_1. More generally, let a player use a function f to represent a sequence of outcomes by a single element. A conservative player uses the minimum, an aggressive player uses the maximum, and a moderate player uses (say) the median.

Such points of view are often taken into account by stockbrokers advising people on investments. A younger investor may prefer a stock with a high potential growth but significant risk. An investor close to retirement age may, on the contrary prefer a stock with less growth but also less risk. A middle aged investor may accept a moderate amount of risk.

The function f should satisfy some rationality conditions.

Definition 1. *A choice function f is* suitable *if it satisfies the following two conditions:*

1. *If X is a final segment of Y, then $f(X) \geq f(Y)$ and if X is an initial segment of Y then $f(Y) \geq f(X)$.*
2. *(Dubey) If sequences X and Y overlap, but all elements in $X - Y$ are higher than all elements of $Y - X$, then $f(X)$ is higher than $f(Y)$.*
3. *If sequences X and Y are in an order preserving one-one correspondence g then $g(f(X)) = f(Y)$.*

Lemma 11 *The minimum, the median and the maximum are all suitable functions in the sense above (and the corresponding notions of f-rationality are equivalent to being conservative, moderate, and aggressive respectively).*

Note that an SCF need not satisfy Nash's IIA condition that if $a = f(X)$, $Y \subseteq X$, and $a \in Y$ then $a = f(Y)$. It so happens that both the maximum and the minimum do satisfy this condition, but not the median. Of course there is no particular reason why IIA *should* be obeyed in such a case. The role of $f(X)$ is to play the role of an element which in some sense *represents* X rather than that of a *most preferred* element of X. Thus the median is probably the closest to the expected value which we tend to use when we have cardinal utilities and a subjective probability.

Definition 2. *Given an SCF f, An f-rational agent is an agent who, when uncertain between sets X and Y of alternatives, always picks X if $f(X) > f(Y)$.*

It is easily seen that if all payoffs in X are higher than those in Y then an f-rational player will choose X over Y. Thus all three kinds of players, conservative, moderate and aggressive will never pick a strictly dominated strategy.[6]

[6] By a *strictly dominated* strategy we will mean a strategy which is dominated by another pure strategy. See appendix for details.

1.3 An Example

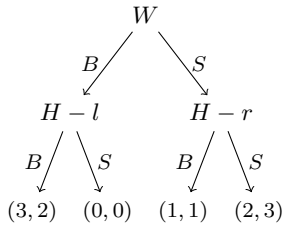


Fig. 1.

In figure 1, we assume that the wife moves first and the husband after. We consider various scenarios involving the husband's knowledge and temperament. We assume that the wife knows the husband's payoffs and temperament and he does not know hers.

Case 1) Husband does not know wife's move (and she knows this).

a) He is aggressive. Then being aggressive, he will choose S (Stravinsky) for his move since the highest possible payoff is 3. Anticipating his move, she will also choose S, and they will end up with payoffs of (2,3).

b) The husband is conservative. Then not knowing what his wife chose, he will choose B since the minimum payoff of 1 is better than the minimum payoff of 0. Anticipating this, the wife will also choose B and they will end up with (3,2).

2) Finally if the husband *will* know what node he is at (and the wife knows this), then the wife will choose B, the husband will also choose B and they will end up at (3,2).

1.4 Example 2

Artemov [2] is concerned with rationality in the presence of uncertainty.[7] A rational player for him is one who makes a decision based on the highest *guaranteed* payoff, subject to the player's knowledge. In other words he describes as rational the kind of player we have chosen to call conservative.

[7] His utilities are also ordinal.

Artemov shows (his theorem 1) that a rational player *in his sense* will follow the backward induction solution *even in the absence of common knowledge of rationality*. Thus Artemov generalizes Aumann's result, replacing common knowledge of rationality by plain rationality.[8]

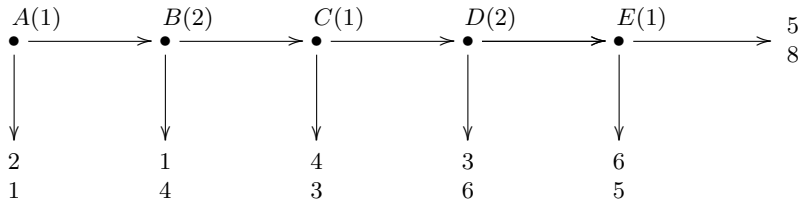


Fig. 2. Centipede game

Now consider a *moderate* player playing this game. If he had been conservative and used backward induction, he would go down at once and get a certain payoff. But since he is a moderate, he will see that the median from going across is much higher. If both players are moderate players, and agnostic about the rest of the game, then the game will continue for quite a while, with both players going across and earning much larger payoffs. Thus our notion of a moderate player shows the rationality of the common pattern seen in ordinary behaviour where players play across for quite a while.[9]

1.5 Comparison with previous work

Two relevant sources are the book by Chwe [7] and the recent paper by Artemov [2]. Chwe's book is largely concerned with the manipulation of beliefs through some form of advertising. An advertiser may seek to create the common belief that everyone is drinking beer X and so the viewer of the TV show should also drink beer X. However, Chwe's treatment is largely non-technical and does not bring in game theoretic techniques for the most part.

Artemov does mention a case (section 5.2 of [2]) where revealing true information changes the behavior of the players.

[8] Artemov's argument applies only to the tree. For other games Artemov's solution could diverge from the backward induction solution.

[9] We are assuming here that the players will be agnostic about the actions of the other player rather than carry out the elaborate backward induction argument.

What is novel in our present work is that we make knowledge manipulation the central aspect of our considerations and we do bring in some technical considerations.

Thus while we acknowledge a debt to Chwe and Artemov, we are carrying the ideas considerably further.

Other work like that of Brandenburger et al [6] is also relevant but unlike us they rely on cardinal utilities. They also do not speak about actual manipulation of behavior by limiting knowledge.

Finally, Agotnes et al have written a very interesting paper about the power which agents have over other agents who want some knowledge. Suppose A knows $P \to R$, B knows $Q \to R$ and C knows $P \wedge Q$. Then, if the interest is in knowing that R is true, C has the most power since *either* of the pair A and C or the pair B and C could derive R; but A and B together could not.

2 Game Theory

Let us consider a game tree for two *(The number two has no special significance and is only used to simplify notation.)* players with a set X of nodes, divided into X_1, the nodes where player 1 moves, X_2 where player 2 moves, and T the set of terminal nodes so that X is the disjoint union of X_1, X_2, T. Moreover payoff functions p_1 and p_2 are defined on T. To simplify matters we will usually assume that both p_1 and p_2 are 1-1. *(I.e., the payoffs at distinct leaves are distinct, i.e., the tree is generic.)*

In that case we know that if we have a *perfect information* game, then backward induction yields a unique way in which the game is played and according to Aumann, that will indeed be the way the game will be played if there is common knowledge of rationality, see [3,2].

But of course a perfect information game might be played differently from an imperfect information game with the same structure, same moves, and the same payoffs. As we saw with the example in figure 1, this matters, because someone who can manipulate the knowledge of others can also affect the way they play some particular game. If the game has payoffs not only for the active players, but also for the KM, then KM will seek to manipulate the active players' knowledge in such a way as to maximize her own payoff.

2.1 States of Knowledge

We now describe a model for representing a game with possibly complex knowledge situations. We will use Kripke models for that. X is the set of nodes of the game tree.

Let us stipulate that for each element $A \in X$, A is also an atomic formula which is true precisely when the play is at node A. We create a formal language L by closing under truth functions, operators K_1, K_2 and the operator C. (Here K_1 means that 1 knows, K_2 means that 2 knows, and C stands for common knowledge).

Then a perfect information game is simply a game where formulas of the form $A \to C(A)$ are true for all nodes A.

But now consider a game with two players 1 and 2 and where the formula $A \to (K_1(A) \wedge K_2(A))$ holds at all nodes A, but for instance $K_1K_2(A)$ does not hold at node A. At each node, both players know what node it is but they do not know that the other knows.

With $(K_1(A) \wedge K_2(A))$, both players know which node they are at. But if 1 makes a choice between L and R, 2 knows which choice 1 made, but 1 does not know that 2 *will* know, then 1 might well play differently. So it is not a perfect information game, strictly speaking. Yet we cannot indicate the 'imperfection' by indicating an information set.

To represent such situations, we modify the knowledge requirement. We stipulate that with each node A is associated a Kripke structure M_A with two knowers 1 and 2. *Such a Kripke structure would represent a state of partial knowledge on the part of the players.*

We assume that the map $A \rightsquigarrow M_A$ is common knowledge.[10] To fix thoughts, we also assume that common knowledge of temperaments (conservative, moderate or aggressive) exists. Each player plays according to his own temperament subject to what he believes about the choice situation he will be in.[11] Thus the class of knowledge situations we can consider is more general than perfect information games or games whose imperfection can be indicated simply by information sets.

[10] We of course mean the unpointed Kripke structure M_A, since an agent who knows also what the real world is would know everything.

[11] Thus it is even open in our model to consider players who have not carried out certain deductions which they were entitled to carry out. They choose according to their belief.

We define an *extended knowledge-based game* (or KB-game) as an extended game supplemented by such a function M_A. As we noted, a perfect information game is a special case of such a KB-game. For in that case, for each A, the structure M_A has a single state satisfying A and no other states are accessible to any player.

2.2 Creating knowledge States

How would the KM create the structure M_A? One way that KM can create such structures is, at each node she sends signals to the players – the signal function s is common knowledge, and based on the signal received by the player he can infer something about the node he is at.

Definition 3. *A* game tree with knowledge function *is a standard extensive-form game tree with nodes A along with a set of signals Σ and a function $s : A \to P(\Sigma^n)$ where n is the number of players and P stands for the power set. We extend s to sequences $\in A^*$ in the obvious way. The associated* protocol *(see [14]) $H(A)$ consists of all sequences $(a_1, \sigma_1), m_1, (a_2, \sigma_2) \ldots, m_{k-1}, (a_k, \sigma_k) \in (A \times \Sigma)^*$ such that $a_1, \ldots, a_k$ is a path in the game tree starting at the root, for all $i < k$, a_{i+1} is a child of a_i resulting from the move m_i, and $\sigma_i \in s(a_i)$ for each $1 \leq i \leq k$. We define a* valuation function $V : H(A) \to 2^A$ *by setting $V((a_1, \sigma_1), \ldots, (a_k, \sigma_k)) := \{a_k\}$ (re-using the nodes as propositions). Further, we assume an* observability function *on Σ for each player which gives rise to synchronous epistemic accessibility relations in the usual way. Thus for each $\sigma_i = (s_1, ..., s_n)$, player j observes s_j and moreover the player observes all the m_i which were his own moves.*

Pradeep Dubey [8] has pointed out that by including KM as an additional active player and interpreting her signals as moves, a knowledge based game can be understood as a conventional game of partial information with information sets. For details see the appendix.

2.3 Example 1 revisited

We consider now the question of how KM can create these various knowledge scenarios of example 1.

KM is capable of creating all these three situations by means of signals, as well as the one we did not mention where the husband does not know but the wife does not know that he will not.

For case 1a), $s(H-l) = (l, a)$ and $s(H-r) = (r, a)$. The wife knows (if she did not already) which node they are at, but the husband will not.

For case 2, $s(H-l) = (l, l)$ and $s(H-r) = (r, r)$. Both will know which node they are at.

Finally if KM wants the wife to be in doubt whether the husband knows, he could make $s(H-l) = \{(l, l), (l, a)\}$ and $s(H-r) = \{(r, r), (r, a)\}$. Then if the wife chose left and receives an l, she will not know if the husband got an l or the neutral a. If KM does send (l, l) then the husband will know, but will also know that his wife did not know whether he *would* know.

We have not indicated KM's utilities above. They could appear as a third component of the payoff function. When the game finishes, all three players including KM receive their payoffs and so KM has an interest in seeing to it that the game is played in a certain way. She can do this, to a limited extent, by influencing the structures M_A.

The Kripke structures which arise this way will be special in three ways. In the first place it will be common knowledge that wherever the players are, they are all at *some* node of the game tree (but they may not know the actual node). Secondly, (assuming perfect recall) if a player was uncertain among nodes A and B, and only these, then she will know in the future that she must be at some node below one of A and B. Finally, if she herself performed an action α when she was so uncertain, then whatever node she is at now will have be below either the α successor of A or the α successor of B.

2.4 Predicting the play

Can the KM *always* predict how a game will be played in a less than perfect information state which he has brought about? This is indeed true in a decision theoretic situation if the temperament of the player is common knowledge[12] For instance a conservative agent faced with uncertainty will choose the least risky alternative. And since we assume that no two outcomes have the same value, the least risky alternative will always be well defined and *known to the KM.*

With two person games, there may not be a unique way that the players will play in a game with imperfect information and so the KM may not

[12] By decision theoretic we mean that there is only one agent *apart* from KM, who has a decision theoretic problem to solve.

be able to predict *how* they will indeed play. In particular the reasoning process of the players can be order-dependent, for consider figure 3 below.

With perfect information and CK of (conventional) rationality, the backward induction solution applies. In figure 3, 2 would choose right at B and left at A. The resulting payoffs for 1 are 4 with left, and 3 with right. He chooses L and so does 2 so they get (4,4).

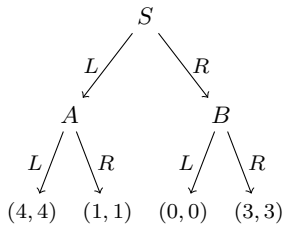


Fig. 3.

But now suppose that when it is his turn to play, 2 (who is conservative) does not know whether he is at node A or B. Then he will choose Right which gives him one of {1,3}, safer than {4,0}, which he would get with Left. 1 will anticipate this and choose Right. So they end up at (3,3).

However, 2 might start his reasoning by trying to figure out 1's move. 1 will get one of {4,1} if she plays Left, and one of {0,3} if she plays Right. So she will play Left. 2 will anticipate this and will play Left. So they end up at (4,4).

Clearly the KM (whose payoffs we have not included) cannot count on any particular play.

Theorem 21 *If player 2 does not know player 1's payoffs but player 1 does know player 2's payoffs, then (given their temperaments) there is a unique solution to the game.*

More generally, with 2 or more players, if the players are linearly ordered so that no player knows the payoffs of any player above him then there is a unique solution.

Future work: In the setup we investigated, there is only one knowledge manipulator who, moreover, is trusted by the other players. But we can consider variants.

One possibility is where the manipulator is, well, manipulative. Her payoff function is known to other players, and they are aware that they cannot fully trust her. This is the direction of cheap talk [10].

Another possibility to consider is that while the KM is presumed honest, every player is both an actor and an informer. This case would be investigated by enriching the purely informational structure of [14] and augmenting it with actions.

Acknowledgement: We thank Sergei Artemov, Pradeep Dubey, and Johan van Benthem for comments.

References

1. Thomas Agotnes, Wiebe van der Hoek and Michael Wooldridge, Scientia Potentia Est, *Proc. 10th Int. Conf. on Autonomous Agents and Multiagent Systems*, Tumer, Yolum, Sonenberg and Stone (eds.) May 2-6, 2011, Taipei, Taiwan.
2. Sergei Artemov, Rational Decisions in non-probabilistic settings, technical report TR-2009012, CUNY Ph.D. program in computer science, 2009.
3. R. Aumann. Backward Induction and Common Knowledge of Rationality, *Games and Economic Behavior*, 1995.
4. Battigalli, Pierpaolo and Bonanno, Giacomo, 1999. "Recent results on belief, knowledge and the epistemic foundations of game theory, " Research in Economics, Elsevier, vol. 53(2), pages 149-225, June
5. Adam Brandenburger, "Epistemic Game Theory: An Overview" (pre-publication version) and "Epistemic Game Theory: Complete Information" (pre-publication version) in The New Palgrave Dictionary of Economics, 2nd edition, edited by Steven Durlauf and Lawrence Blume, Palgrave Macmillan 2008
6. Adam Brandenburger, Amanda Friedenberg and Jerome Keisler, Admissibility in Games, *Econometrica* **76** (2) 307-352 (2008)
7. Michael Chwe, *Rational Ritual: Culture, Coordination and Common Knowledge*, Princeton University Press, 2001.
8. Pradeep Dubey, personal communication, March 18, 2011.
9. Jan van Eijck, and Rineke Verbrugge, *Discourses on Social Software*, Amsterdam University Press, (Paperback - Apr 1, 2010)
10. J. Farrell and M. Rabin, Cheap talk, *J. Economic Perspectives*, **10** (3) 1996, 113-118.
11. Brian Hare, Josep Call, and Michael Tomasello, Chimpanzees deceive a human competitor by hiding, *Cognition*, Volume 101, Issue 3, October 2006, Pages 495-514
12. Martin Osborne and Ariel Rubinstein, *A Course in Game Theory*, MIT Press (1994)
13. Rohit Parikh, Social Software, *Synthese*, **132**, Sep 2002, 187-211.
14. Parikh and Ramanujam, A Knowledge based Semantics of Messages, *J. Logic, Language and Information* **12** (2003), 453-467
15. Premack, D. G. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, **1**, 515-526 (1978.)
16. Tommy Tan and Sergio Werlang, The Bayesian foundations of solution concepts of games, *J. Economic Theory*, **45** (1988) 370-391.

3 Appendix

Proof of lemma 11 It is obvious that the median, the maximum and the minimum are preserved by isomorphism. we check the Dubey property just for the median. Suppose that X and Y overlap so that X is $a_1 > a_2 > ... > a_k > b_1 > ....b_m$ and Y is $b_1 > b_2 > ...b_m > c_1 > ... > c_p$. $X - Y$ is above $Y - X$. Clearly if the median of X is an a_i or the median of Y is a c_i then we are done. If both medians are b_i and b_j respectively. Then $i + k = m - i + 1$ and $j = p + m - j + 1$. Thus we get $2i = m + 1 - k$ and $2j = p + m + 1$. Thus $i < j$ and $b_i > b_j$. □

Sketch of proof of theorem 21: We do not assume that player 2 always plays after 1. For instance the game may be over more than two stages.

At any particular node, player 2 has a set of nodes X which he *might* be at. He considers all possible strategies s of player 1 which are compatible with their presently being in X. For each such s he considers various strategies s' which he himself could play and the payoff $p(s, s')$ to himself of s, s'. Then he chooses that s' for which $\min\{p(s, s')|s \in X\}$ is highest.

This defines the strategy s' of 2 as a function of the node. Player 1 can simulate player 2's reasoning and plays so as to maximize her own payoff.

This yields a unique outcome. □

Note that since player 2 does not know player 1's payoffs, he is not able now to think of a proper response to player 1's choice - he has no idea what it is. So there is no 'cycle of reasoning'.

We now provide definitions for the way in which a KM can create appropriate Kripke structures.

Intuitively, at each node a, the KM chooses and sends an n-tuple of signals $(s_1, ..., s_n) \in f(a)$. Player j observes only s_j but can infer something about the signals received by the other players. Moreover, he observes his own moves. Based on what he has seen, he can infer a set of possible sequences compatible with what he has seen, and what he knows is what is true in all these possible sequences [14].

As Dubey has pointed out, the kinds of structures we defined above can be replaced by traditional imperfect information games with information sets, *provided that* the knowledge player's signals are treated as actual moves arising *within the game tree* rather than outside it.

Consider the case from figure 1 where KM wants the wife to be in doubt whether the husband knows, she could make $s(H - l) = \{(l, l), (l, a)\}$ and

$s(H-r) = \{(r,r),(r,a)\}$. Then if the wife chose left and receives an l, she will not know if the husband got an l or the neutral a. If KM sends (l,l) then the husband will know, but will also know that his wife did not know whether he *would* know.

Thus KM could have two moves for each of the wife's moves. After her move L, she could have an L move corresponding to the signal pair (l,l) and an R move corresponding to the signal pair (l,a). Similarly after her R move, he could have an L move corresponding to the signal pair (r,r) and an R move corresponding to the signal pair (r,a). This gives us four nodes corresponding to the moves by the wife and KM, and let us denote them in the natural way as LL, LR, RL and RR. (See figure 4).

The nodes LL, LR are indistinguishable for the wife and similarly RL and RR. She knows what she moved, but does not know what the husband got. The husband cannot distinguish between LR and RR, because in the signal description he got an a in either case. But the other two, LL and RL are singletons for him. If he gets an l or an r he knows how the wife moved.

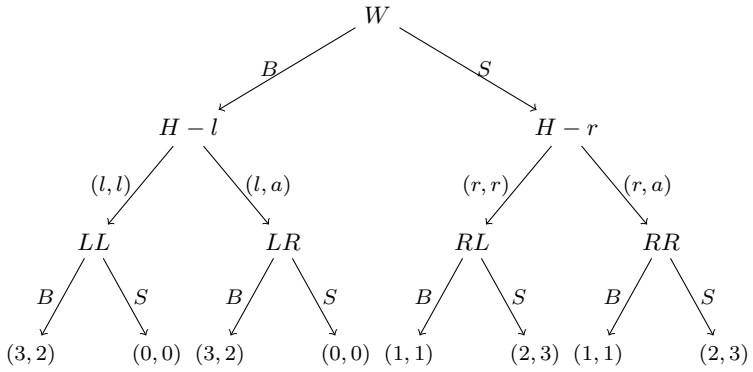


Fig. 4.

Definition 4. *A* D-tree *is a standard extensive-form game tree as above, but with the choices given by s interleaved after each move, and information sets added as follows: For each player i, at each depth in the tree, any two nodes share an information set iff their parents share an i-information set and*

(i) they both result from the same action by i himself, or
(ii) they result from two s-actions which lead to the same observation for i, or
(iii) they result from some other player's actions.

Additionally, we define a valuation function *which assigns a unique proposition to all nodes generated by f-actions from the same parent. The* knowledge situation *after n moves is the horizontal slice of this tree at depth* $2n$.

Theorem 31 *The knowledge situations in a game tree with knowledge function are isomporphic to the ones in the corresponding D-tree (modulo renaming of propositions).*

Proof. intuition: both constructions boil down to taking the product of a "normal" move and the signals that can be sent along with it, and in both constructions the indistinguishabilities are wired according to the observability of the signal part.

Theorem 32 *Any knowledge situation can be created in a single signaling step.*

Proof. Intuition: Take the Kripke structure representing the knowledge situation and create an edge from a unique (new) root node to each possible world. Label each edge with tuples $(\sigma_1, \ldots, \sigma_n)$ of signals, one for each player i, such that any two edge labels coincide in σ_i iff the worlds they lead to are indistinguishable to i. Define the observability function for player i as the restriction of a given tuple to its ith component.

Postscript on dominated strategies: Suppose that an agent believes he is facing various scenarios $s_1, ..., s_n$ but does not know which one. For each of these he has payoffs l_i from playing L and r_i from playing R and in each case $l_i > r_i$. then it is an easy one step argument that the set $\{l_1, ..., l_n\}$ has higher maximum, minimum and median values than the set $\{r_1, ..., r_n\}$. Thus whether a player is conservative, aggressive, or moderate, he will not choose R. Only the case of a moderate player requires a very short argument which we leave to the reader.

Children's Strategy Use in Playing Strategic Games

Maartje E. J. Raijmakers[1], Sara van Es[1], Marian Counihan[1]

[1]Department of Psychology, University of Amsterdam, M.E.J.Raijmakers@uva.nl

Abstract. Strategic games require to reason about other peoples and one's own beliefs or intentions. Although they have clear commonalities with psychological tests of theory of mind, they are not clearly related to these tests for children between 9 and 10 years old [6]. We study children's (5 – 12 years of age) individual differences in playing a strategic game by analyzing the strategies that they apply in a zero, first, and second-order reasoning tasks. For the zero-order task, there were two subgroups with different accuracy. For the first-order task subgroups apply different suboptimal strategies or an optimal strategy. For the second-order task only different suboptimal strategies were present. Strategy use for all tasks was related to age. For the 5 and 6 years old children strategy-use was related to working memory, and not to theory of mind, after correction for age, verbal ability and general IQ.

Keywords: Strategic games, child development, reasoning, theory of mind, strategy analysis.

1 Introduction

Strategic games require to reason about other peoples and one's own beliefs or intentions. Hedden and Zang [8] designed a matrix game to distinguish the use of first-order and second-order theory of mind in adults. First-order reasoning involves a proposition of the form: "The other person (A) plays X" and second-order reasoning involves a proposition of the form: "A knows that I will play X, so A will play Y". That is, second-order reasoning involves 2 propositions that are embedded. Hedden and Zang suggested that for optimal play in a strategic game one needs a theory of mind.

Since theory of mind is still developing into childhood [20], a limited theory of mind is expected to be a factor in children's ability to play strategic games. The development of theory of mind is most extensively tested with false-beliefs tasks, appearance-reality tasks, and deception tasks [23,15,5,7]. These tasks all require first-order reasoning, that is, they involve a proposition "A beliefs X". In development, theory of mind, as measured with these tasks, is strongly related to executive functions, especially a combination of working memory, inhibitory control and planning, independent of age, verbal abilities, and intelligence (e.g., [1,2]).

Second-order false-belief tests require reasoning with reference to what another person beliefs about your own intentions, that is, it involves two propositions "A believes that I believe X" that are embedded, as in the strategic game that requires

second-order reasoning. Second-order reasoning is mostly studied with stories from which one has to infer a person's belief [14,15]. Second-order reasoning involves more information, more complex worded sentences, and puts more demand on working memory. Success of second-order reasoning emerges around 5 and 6 years of age, but differs substantially between tasks and studies. Second-order reasoning abilities are related to inhibition, planning, and working memory, but not in all studies independently from verbal abilities and general intelligence [13].
Flobbe et al. [6] adapted the task by Hedden en Zang [8] such that the strategic game is understandable and appealing to children. They showed that 55% of the 8 to 10 years old children perform first-order reasoning (at least 5 out of 6 items correct) and these children can show second-order reasoning above chance level. However, the game results were not related to two theory of mind tasks, a false belief task and a sentence-comprehension task. As they concluded, successful first and second-order theory of mind in 8 to 10 years old children depends crucially on the domain in which it must be applied.
In summary, we could state that, looking at the structure of the tasks, first-order and second-order reasoning in theory of mind requires the same type of reasoning as in the strategic games. However, the ability to play the strategic game appears not to be related to other theory of mind tasks in 8 to 10 years old children. In development, abilities measured by typical theory of mind tasks are related to executive functions, inhibition, working memory, and planning, but verbal abilities, general intelligence and age partially contribute to this relation. As yet, the relation between playing strategic games and executive functions is not known.

1.1 Individual differences

In Flobbe et al. [6] children make more mistakes on first-order and second-order reasoning tasks than adults, but they found also considerable variation within the group of children. The source of this inter-individual variation did not become clear. Individuals can differ in game playing in multiple ways. They can play different strategies and/or they can differ in the number of mistakes in applying one and the same strategy. By inspecting sum scores of test items, it is difficult to disentangle these different sources of variation in the data. Only the use of different strategies would indicate different insight into the games, for example first-order and second-order reasoning. The aim of the present research is to study different strategies in playing strategic games and the way these strategies are related to age, executive functions, and a standard false-belief task. For multiple cognitive domains, children appear to acquire increasingly complex reasoning strategies [18].
To this end, we studied first and second-order reasoning in playing Flobbe et al.'s strategic game in 5 to 12 years old children. Our approach is novel in analyzing the reasoning performance. We designed the reasoning task such that different items in the task distinguish between expected strategies in an optimal way. The strategies that could be expected are firstly, optimal strategies where children optimize absolute gain. Secondly, it can be expected that, as Flobbe et al. found, that some children optimize relative gain instead of absolute gain. Thirdly, young children might not master first or second-order reasoning and hence it is possible that zero-order reasoning is applied to the first-order task and first-order reasoning (that is only

played if first-order reasoning is mastered) is applied to the second-order task. Finally, children could also have a position bias.
We applied the statistical technique of latent class analysis (LCA) to model the strategies from the accuracy data. LCA (McCutcheon, 1987) provides a statistically reliable method to detect strategies from response patterns [10,19]. Hence, by the application of LCA one can establish which and how many strategies are actually applied. It is not required to fully define the expected strategies beforehand. After revealing the strategies for 5 to 12 years old children, for the 5 and 6 years old children we relate the use of strategies to age, IQ, Verbal Ability, working memory, and theory of mind. These children are expected to show the most variation in theory of mind and also the executive functioning that we measure.

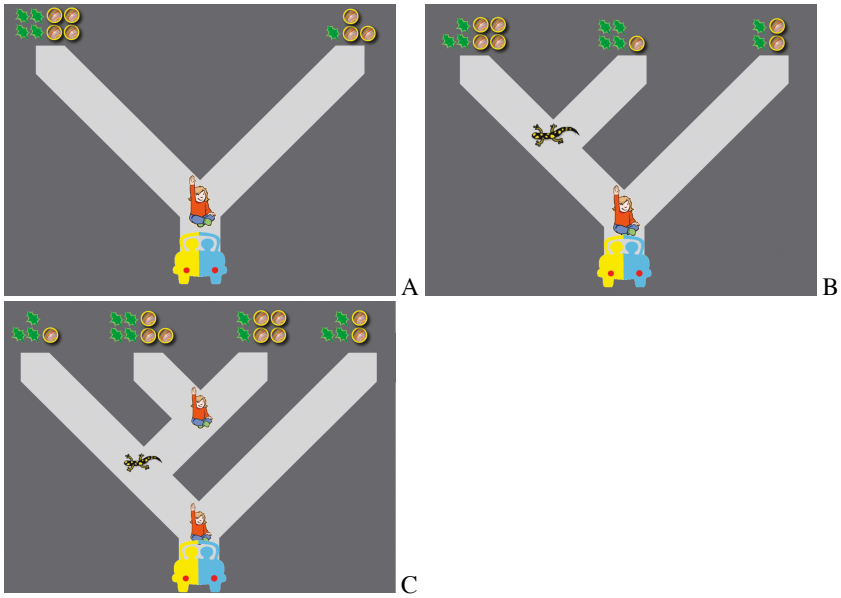


Fig. 1. Three items from the traveling game, which is based on Flobbe et al. (2008). 1A: a zero-order reasoning item, 1B: a first-order reasoning item; 1C: a second-order reasoning item. The child travels together with a lizard in a car. She/he has to acquire as many marbles as possible, but the lizard will try to gain as many leafs as possible. At each cross of the road either the child or the lizard (as is indicated) can decide where they go together, left or right. The player plays the child and the computer plays the lizard. The lizard has an optimal playing strategy, assuming the child uses an optimal strategy.

2.1 Overview

The main research question is whether inter-individual differences in playing a strategic game are due to using different strategies or to playing the same strategy with different accuracy. The next question is whether strategy use is related to the

developmental notion of theory of mind, as was suggested by [8], or to other cognitive abilities, specifically working memory after correction for age, verbal abilities and general IQ.
The reasoning task we applied is a traveling game as in Flobbe et al. [6], but the appearance is somewhat different (Figure 1). There exist three types of items, which were presented in three tasks: items that require zero-order (1A), first-order (1B), and second-order reasoning (1C). In the task, the child travels together with a lizard in a car. She/he has to acquire as many marbles as possible, but the lizard will try to gain as many leafs as possible. At each cross of the road either the child or the lizard can decide where to go, left or right. The player plays the child and the computer plays the lizard, which has an optimal playing strategy, assuming an optimal strategy of the child.

2 Method

Participants were 129 children in the age range of 5 to 12 years: 23 5-years old, 26 6-years old, 16 7-years old, 14 8-years old, 15-9 years old, 10 10-years old, 18 11-years old, 7 12 years old children. Children were tested at a middle-class primary school in Amsterdam, The Netherlands

2.1 Materials

Traveling game: The strategic game is briefly explained in Figure 1. The test consisted of three tasks: a task with 2 example and 9 zero-order test items, a task with 3 example and 15 first-order test items and a task with 3 example and 9 second-order test items. All items are listed in Appendix A.

For the 5 and 6 years old children we used the following battery of cognitive tests:
IQ test: The Raven's Progressive Matrices for fluid intelligence, part A, B, and C for which we calculate a sum score.
Verbal ability test: A Dutch test for sentence comprehension, TAK (Taaltoets voor Alle Kinderen; [21]).
Working Memory test: the digit span forward and backward task.
Theory of Mind test: For the false belief test, the participants heard two second-order false belief stories, accompanied by drawings by the hand of Flobbe. The first story was the 'Birthday Puppy Story' reported in [20], a standard second-order false belief task. The second story, the 'Chocolate Bar Story', was a second-order adaptation of a first-order story by Hogrefe and Wimmer [9]. Both stories had first- and second-order questions. These test were exactly the same tests as were used in [6] experiment 1.

2.2 Procedure

Children were tested in two sessions on two different days if they completed all tests (5 and 6 years old children), otherwise they were only tested in one session. The first

day they played the traveling game, the second day they completed the cognitive tests battery. The traveling game was explained and tested on a computer. Children started with the example zero-order items. The first item was used to explain the game. The child played the second item. Children responded by clicking an arrow on the road they wanted to go. In the example items children saw the animated car moving on the screen and they were presented on the screen the resulting marbles for her- or himself and the lizard was presented the leafs. If the second example item was made incorrect, the first item appeared again and the game was explained a second time. In this way, we also tested whether (the youngest) children could count. After the example items the child made the test items. Now, the child saw the animated car but did not get any direct feedback. Only after 3 items the cumulative gain was presented on the screen as a bag full of marbles.

The first-order items were explained to the children with 3 example items. Children were explained the task from the first item. Then, with the second item, which required first-order reasoning, the experimenter used instructional scaffolding to direct the child towards an optimal choice [16]. The child made the third example item, which could also be solved by a suboptimal strategy, by her-/himselves. After the example items the children made 15 test items. Again, only during the example items feedback occurred on the screen. After a choice on the test items only the first part of the animation was shown. After three items, the cumulative gain was shown on the screen.

After the first-order task, a total score for the first-order items was calculated. Only the children with 12 (out of 15) items correct, continued with the second-order items. The procedure for the second-order items was equivalent to the first-order items, again with three example items.

3 Results

All 129 children completed the experiment. However, only 55 children (43%) passed the first-order reasoning task and completed the second-order reasoning task. Their mean age is 9.78 years (sd = 1.96). Mean scores for the three tasks were above change level for the zero-order task ($t(128) = 40.1$, $p < .001$) and the first-order task ($t(128) = 10.6$, $p < .001$), but not for the second-order task ($t(54) = 1.6$, $p = .06$). Table 1 (last column) shows the scores for each task.

Strategy analysis was conducted for the three tasks separately. For each task latent class models with different number of classes were fitted to the accuracy scores of the items. The best fitting, most parsimonious model (according to the BIC, Schwartz, 1978) was selected for each task.

See table 2 for the resulting most parsimonious, best fitting models. Two strategies were found for zero-order reasoning (N = 129): The first strategy (12% of the participants) is suboptimal and has lower probability correct ($p = .62$) for the items for which the largest sum of leafs and marbles was not the optimal choice. Mean score for participants following this strategy is .66, which is above change level (sd = 12, $t(15) = 5.1$, $p < .001$; see Table 1). The second strategy (88%) is an optimal strategy.

The mean score for participants following this strategy is .98 (sd = .05). Strategies for the zero-order task were related to age (p = .002; Figure 2).

Table 1. Mean scores for the three reasoning tasks per strategy

Task	S1	S2	S3	S4	All
0-order	0.66 (0.12)	0.98 (0.05)			0.94 (0.12)
1st-order	0.58 (0.16)	0.54 (0.05)	0.51 (0.08)	0.94 (0.06)	0.70 (0.22)
2nd-order	0.42 (0.16)	0.60 (0.17)			0.54 (0.19)

Note. Columns S1 - S4 denote the mean (sd) proportion correct for the different tasks for participants responding according to strategies S1 – S4 respectively. Column All shows the mean (sd) proportion correct for all participants. Note that cell S2 for 0-order relates to different subjects than cell S2 for 1st-order task, etc. The strategies are listed in the same order as in Fig. 1.

Four strategies were found for the first-order reasoning task (N = 129): The first strategy (39%) cannot be distinguished from guessing. The second strategy (19%) is a zero-order strategy. The third strategy (5%) is to go right, which avoids a choice by the lizard. Although the third group is very small (6 children) the class does contribute to a better, parsimonious description of the data. The fourth strategy (37%) has a high probability of responding optimally for all items. The children in this group have an optimal strategy. Table 1 shows the mean scores per strategy. Strategies for the first-order task were related to age (p < .001; Figure 2).

Table 2. Resulting models from latent class analysis.

	prior	conditional probabilities			
zero-order items		type 1	type 2		
bias to sum	17%	.62	.75		
optimal	83%	.98	.99		
first-order items		type 1	type 2	type 3	
guess	39%	.62	.47	.79	
0-order	19%	.96	.04	.96	
go right	4%	.03	.97	.03	
optimal	38%	.94	.94	.94	
second-order items		type 1	type 2	type 3	type 4
guess	44%	.5	.5	.5	.5
first-order	56%	.5	1	.5	.5

Note. The estimated parameters from the best fitting, most parsimonious models of the accuracy scores of each task. The priors show the prior probability of belonging to that class (in percentages). The conditional probabilities are the probabilities of responding correctly for the corresponding item type given the strategy. For the zero-order task items 1, 4, 7, 9 are type 1, items 3, 5, 6 are type 2. For the first-order task items 1, 4, 6, 9, 15 are type 1, items 2, 5, 7, 8, 10, 12, 13 are type 2, items 3, 14 are type 3. For the second order task items 4, 7, 9 are type 1, items 2, 3 are type 2, item 6 is type 3, item 8 is type 4.

For the second-order reasoning task (N = 55) two strategies were found. The first strategy (33%) is not distinguishable from guessing. The second strategy (67%) resembles most a first-order strategy where the final choice by the child is neglected in the decision. Table 1 shows the mean scores per strategy. The mean score for participants following a kind of first-order strategy is .60, which is above chance level (sd = .17, $t(36) = 3.5$, $p < .001$). Strategies for the second-order task were related to age ($p = .005$). Figure 2 shows per task the distribution of strategies (in percentages) for each age group. The relation between strategy-use and age is apparent from this figure.

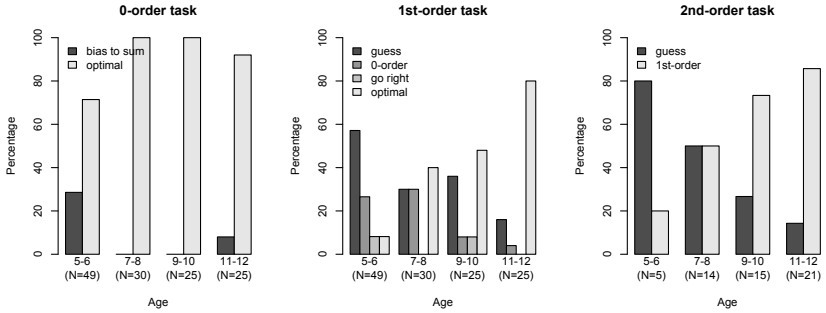


Fig 2. The distribution of strategies (in percentages) for each age group. Below the bar the number of participants within the age group is depicted. A) zero=order task, B) first-order task, C) second-order task.

Table 3. Summary data cognitive tests

Task	5 years		6 years	
	mean	sd	mean	sd
ToM	4.80	1.96	5.92	1.44
ToM1	2.65	1.18	2.92	0.74
ToM2	2.15	1.14	3.00	1.02
DS	5.90	2.05	7.42	1.58
RPM	13.30	4.50	15.88	4.23
Tak	21.85	3.69	24.77	2.39

Note. ToM is the Theory of Mind test, the sum of scores to first-order (ToM1) and second-order (ToM2) questions. DS is the Digit Span test, backwards and forwards DS summed. RPM is the Raven's Progressive Matrices, part A, B, C. Tak is the Sentence Comprehension test.

The cognitive abilities were tested only for the 5 and 6 years old children (mean age = 6.0 years, sd = 0.6). Summary statistics are shown in Table 3. For all measures we have a considerable variation, which is important to detect a relationship with strategy use. After correcting for age and verbal ability, there was a correlation between theory of mind and working memory ($r = .32$, $p = .02$) and theory of mind and IQ ($r = .16$, $p = .005$). For the zero-order task, in this age group, only age had a unique relation to

strategy use and not to the other abilities that were measured (logistic regression: coeff. = .1, p = .047). For the first order task, in a logistic regression analysis with all cognitive abilities and age as predictors, only working memory had a unique relation to strategy use (coeff. = .59, p = .017) and not the other abilities or age.

4 Conclusion

Strategy analysis of playing a strategic game gives interesting insights into children's reasoning. For the zero-order tasks all children play with the same strategy, but with different accuracies. For the first- and second-order tasks children play with different strategies. The subgroups of children with different accuracies and strategies for the zero-, first and second order task were revealed by careful construction of items and with latent class analysis. On average children have high scores on the zero-order task, but nevertheless they show two types of performances. A subgroup of the children is making more mistakes and is distracted by a large amount of total gains (the sum of marbles and leafs), which results in a suboptimal choice for type 1 items. This strategy is more frequent in younger children. For the first-order task, there is one group of children with an optimal strategy. The other children have different ways of being suboptimal: guessing, location bias or zero-order reasoning. Zero-order reasoning means that the choice of the lizard is not taken into account.
For the 5 and 6 years of age children, strategy use is not related to theory of mind (in addition to age and other abilities), as was suggested in the literature, but only related to working memory. Although the age range is small, the variation in theory of mind scores is quite large. It can be questioned whether the strategic-game tasks and the theory of mind tasks have something specific in common at all. The fact that we do find a specific relation between the strategic games and working-memory task indicates that the reliability of the strategic games are large enough to find relationships with other cognitive abilities.
Finally, for the second-order task, we find one subgroup who's choices could not be distinguished from guessing. The other group seems to apply a kind of first-order strategy, combined with guessing. Although the participants in this subgroup do not use a second-order strategy, the scores of this subgroup are above chance level. This shows that from the fact that a participants have above chance performance, one cannot conclude that the participants master the task and/or the correct strategy. First, it could only be a subgroup mastering the task. Second, it could be that only a partially correct strategy was applies, which is considerably more easy. Hence, sum scores of age groups are not always very indicative for their cognitive abilities.

The overall performance for the first-order and second-order tasks is poor compared to performance on theory of mind tasks. Only 50% of the 9 and 10 years olds show a true first-order strategy, which agrees with the percentage of children that passes the criterion in [6]. However, none of the children shows a proper second order strategy. The poor performance might be due to the instruction by scaffolding instead of learning by feedback, which was used by Flobbe et al. [6]. Note that for the theory of mind tasks, instructions are mostly very limited. The reason that we have chosen for

the scaffolding explanation is that we want to have optimal performance for all ages. Since learning by feedback differs importantly between 5 and 12 years of age (eg., [3]), we avoided learning by feedback. Moreover, for a strategy analysis one should test stable performance. Feedback will result in changing performance if people are not performing in an optimal way from the start (as was found by [8]). For future research it would be interesting to train children extensively on these strategic game items in an adaptive training system over a time frame of weeks, to reveal the optimal performance children gain after extensive deliberate practice [4]. An adaptive test and training system was developed as the Mathsgarden.com (rekentuin.nl; [11]). For a different complex reasoning game, static MasterMind, we see very high performance for primary school children after extensive deliberate practice on a large item bank.

There is a second possible reason why we found few children responding optimally on first and second-order reasoning items, as compared to theory of mind tasks. It is important to note that responding with a non-optimal strategy is not necessarily resulting in non-optimal choices for all items. This is not only true for the items that we designed but also for some of the items that were included in the Flobbe et al. task [6]. The result is that sum scores might end up above chance level unless children are not following a true first- or second-order strategy. Hence, strategy analysis is important for fully understanding performance on complex reasoning tasks and its development.

References

1 Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. Infant and Child Development, 11, 73–92.

2 Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. Journal of experimental child psychology, 87(4), 299-319.

3 Duijvenvoorde, A. C. K. van, Zanolie, K., Rombouts, S. a R. B., Raijmakers, M. E. J., & Crone, E. a. (2008). Evaluating the negative or valuing the positive? Neural mechanisms supporting feedback-based learning across development. The Journal of neuroscience : the official journal of the Society for Neuroscience, 28(38), 9495-9503.

4 Ericsson, K. A. (2006). The Cambridge handbook of expertise and expert performance. Chapter: The Influence of experience and Deliberate Practice on the Development of Superior Expert Performance. Cambridge University Press.

5 Flavell, J. H., Green, F. L., & Flavell, E. R. (1986). Development of knowledge about the appearance– reality distinction. Monographs of the Society for Research in Child Development, 51 (1).

6 Flobbe, L., Verbrugge, R., Hendriks, P., & Krämer, I. (2008). Children's Application of Theory of Mind in Reasoning and Language. Journal of Logic, Language and Information, 17(4), 417-442.

7 Gopnik, A., & Astington, J. W. (1988). Children_s understanding of representational change and its relation to the understanding of false belief and the appearance–reality distinction. Child Development, 59, 26–37.

8 Hedden, T., & Zhang, J. (2002). What do you think I think you think? Strategic reasoning in matrix games. Cognition, 85, 1–36.

9 Hogrefe, G., & Wimmer, H. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. Child Development, 57, 567.
10 Jansen, B.R.J. and Maas, H.L.J. van der (1997). Statistical Test of the Rule Assessment Methodology by Latent Class Analysis. Developmental Review, 17, 321–357.
11 Klinkenberg, S., Straatemeier, M., & Maas, H. L. J. van der. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. Computers & Education, 1-12.
12 McCutcheon, A. L. (1987). Latent class analysis. Newbury Park: Sage.
13 Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological bulletin, 135*, 749-773.
14 Muris, P., Steerneman, P., Meesters, C., Merckelbach, H., Horselenberg, R., van den Hogen, T., & van Dongen, L. (1999). The TOM test: A new instrument for assessing theory of mind in normal children and children with pervasive developmental disorders. Journal of Autism and Developmental Disorders, 29, 67–80.
15 Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds_ difficulty understanding false belief: Representational limitation, lack of knowledge, or pragmatic misunderstanding? British Journal of Developmental Psychology, 5, 125–137.
16 Rodgers, E. M. (2004). Interactions that scaffold reading performance. Journal of Literacy Research, 36(4), 501-532
17 Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464.
18 Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. Cognitive Psychology, 28, 225-273.
19 Straatemeier, M., Van der Maas, H.L.J., & Jansen, B.R.J. (2008). Children's Knowledge of the Earth: A New Methodological and Statistical Approach. Journal of Experimental Child Psychology, 100, 276–296.
20 Tager-Flusberg, H., & Sullivan, K. (1994). A second look at second-order belief attribution in autism. Journal of Autism and Developmental Disorders, 24, 577–586.
21 Verhoeven, L., & Vermeer, A. (2006). Verantwoording Taaltoets Alle Kinderen (TAK). Arnhem: Centraal Instituut voor Toetsontwikkeling.
22 Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. Child Development, 72, 655-684.
23 Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13, 103–128.

Appendix A

In Appendix A, all items are enumerated (Tables A1, A3, and A5) and the expected accuracy patterns according to different strategies (Tables A2, A4, and A6). In the latent class analysis not all items are used (see note Table 2), because these items did not correlate well with the items of the same type for unclear reasons. Items of the same type have the same expected scores for all strategies.

Table A1. Zero-order items

Items	B1		B2		Optimal
	L	M	L	M	Response
A	3	3	2	1	L
B	3	1	2	4	R
1	3	3	1	4	R
2	4	4	1	3	L
3	1	2	3	3	R
4	1	4	4	2	L
5	4	3	1	2	L
6	1	3	3	4	R
7	1	3	4	2	L
8	4	3	2	4	R
9	3	3	1	4	R

Note. Items A and B are example items, 1 – 9 are test items. Items are coded by an enumeration of leafs (L) and marbles (M) from the left branch (B1) to the right branch (B2). See Figure 1 for the configuration of leafs and marbles. The optimal choice is left (L) or right (R).

Table A2. Expected accuracy patterns for different potential strategies

	Strategies		
Items	0-A	0-B	0-C
1, 8, 9	1	0	1
4, 7	1	0	1
2, 5	1	1	0
3, 6	1	1	0

Note. The potential strategies are 0-A, the optimal strategy, 0-B, the choice for largest sum of leafs and marbles, 0-C, the choice for largest relative gain. 1 is correct, 0 is incorrect

Table A4. Expected accuracy patterns for different potential strategies

	Strategies					
Items	1-A	1-B	0-A	0-B	0-C	0-D
1, 4, 6, 9, 15, 11	1	0	1	1	0	0
2, 5, 7, 8, 10, 12, 13	1	1	0	0	0	1
3, 14	1	1	1	1	1	0

Note. The potential strategies are 1-A, the optimal strategy, 1-B, the choice for largest relative gain, 0-A, a zero-order strategy with largest gain, 0-B, zero-order strategy with largest sum of leafs and marbles, 0-C zero-order strategy with largest relative gain, 0-D, go to the right. 1 is correct, 0 is incorrect.

Table A5. Second-order items

Item	B1		B2		B3		B4		Optimal
	L	M	L	M	L	M	L	M	Response
A	2	3	4	4	3	1	1	2	L
B	2	2	1	4	4	1	1	3	R
C	4	3	1	1	2	2	2	1	L
1	3	3	4	1	2	4	2	2	L
2	3	2	2	4	4	1	2	3	R
3	3	2	2	4	4	1	2	3	R
4	1	3	2	4	4	2	3	3	L
5	3	1	4	3	1	4	1	2	R
6	3	1	1	4	4	3	2	2	R
7	2	3	1	3	4	1	2	2	L
8	3	1	4	3	2	4	3	2	R
9	1	2	4	2	2	4	2	3	L

Note. See note Table A1

Table A6. Expected accuracy patterns for different potential strategies

	Strategies		
Items	2-A	1-A	1-B
1, 9,4,7	1	1	0
2, 3	1	0	1
5, 6, 8	1	0	0

Note. The potential strategies are 2-A, the optimal strategy, 1-A, a first-order strategy with a second choice for the child, 1-B, a first-order strategy without a second choice for the child. 1 is correct, 0 is incorrect.

The Advantage of Higher-Order Theory of Mind in the Game of Limited Bidding

Harmen de Weerd, Bart Verheij

Institute of Artificial Intelligence, University of Groningen, P.O. Box 407, 9700 AK, Groningen

Abstract. Higher-order theory of mind is the ability to recursively model mental states of other agents. It is known that adults in general can reason adequately at the second order (covering attributions like "Alice knows that Bob knows that she wrote a novel under pseudonym"), but there are cognitive limits on higher-order theory of mind. For example, children under the age of around 6 cannot correctly apply second-order theory of mind, and it seems to be a uniquely human ability. In this paper, we make use of agent-based models to investigate the advantage of applying a higher-order theory of mind among agents with bounded rationality. We present a model of computational agents in the competitive setting of the limited bidding game, and describe how agents achieve theory of mind by simulating the decision making process of their opponent as if it were their own. Based on the results of a tournament held between these agents, we find diminishing returns on making use of increasingly higher orders of theory of mind.

1 Introduction

Humans, in many aspects, are extraordinary within the animal kingdom. They show an impressive ability to reason about the world around them, as well as about unobservable mental content of others, such as others' knowledge, beliefs and plans. This so-called *theory of mind* [1] is said to be unique to humans [2]. Humans use theory of mind beyond its first-order application, concerning other's propositional attitudes with respect to world facts. They take this ability to a second-order theory of mind, in which they reason about the way others reason about mental content. For example, suppose that Alice is throwing Bob a surprise party. Bob engages in second-order theory of mind when he knows about the party, but is playing along with Alice so she won't find out that he already knows; he may make the second-order attribution "Alice doesn't know that I know".

Although the ability to use higher-order (i.e., at least second-order) theory of mind is well established for humans, both through the attribution of second-order false belief [3] as well as in strategic games [4–6], the use of theory of mind of any kind by non-human species is a controversial matter [see for example 2, 7, 8]. Also, research shows that even human adults have difficulty applying higher-order theory of mind correctly [4, 5, 9].

In this paper, we consider agent-based computational models [10, 11] to investigate the advantages of making use of higher-order theory of mind. The use of computational agents allows us to precisely control and monitor the mental content, including application of theory of mind, of our test subjects. This allows us to investigate the conditions under which a theory of mind would present individuals with an evolutionary advantage over individuals without such abilities. Following the Machiavellian intelligence hypothesis [12, 13], the main driving force behind the evolution of social cognition, such as theory of mind, would be the competitive ability within the species (for a discussion of alternative hypotheses, see [9]). We therefore simulate computational agents in a competitive game, and determine the extent to which higher-order theory of mind provides individuals with an advantage over competitors that are more restricted in their use of theory of mind. In particular, we consider whether the ability to use second-order theory of mind provides individuals with advantages beyond the use of first-order theory of mind.

The setting in which we compare the performance of the computational agents is a newly designed competitive game of limited bidding, which is explained in Section 2. Section 3 gives a detailed description of the way agents that are limited in their ability to explicitly represent mental content are implemented for the limited bidding game. These agents are placed in competition with one another, the results of which are presented in Section 4. We compare the advantages of using second-order theory of mind to those obtained using first-order theory of mind. Finally, Section 5 provides discussion and gives directions for future research.

2 Limited bidding

2.1 Game outline

The limited bidding game (adapted from a game in [14]) is a competitive game played by two players. At the start of the game, each player receives an identical set of N tokens, valued 1 to N. Over the course of N rounds, players simultaneously choose one of their own tokens to use as a 'bid' for the round. Once both players have made their choice, the tokens selected by the players are revealed and compared, and the round is won by the player that selected the highest value token. In case of a draw, there are no winners. The object of the game is to win as many rounds as possible while losing as few rounds as possible. However, each token may be used only once per game. This forces players to plan ahead and strategically choose which of the still available tokens to place as the bid. For example, a player that selects the token with the highest value (N) in the first round will ensure that the first round will not result in a win for his opponent. However, this also means that during the remaining $N - 1$ rounds, the token with value N will not be available to this player.

After each round in the limited bidding game, the tokens that were played are announced to each player. That is, at the end of each round, every player not only knows who won the round, but also which tokens were used. This allows

players to keep track of the tokens that may still be played by their opponent. Since our computational agents are limited in their ability to make use of theory of mind, these agents do not have the capacity to explicitly represent common knowledge. However, we assume that players do not hold beliefs that would be inconsistent with common knowledge of the rules and dynamics of the limited bidding game.

2.2 A scenario for agents with bounded rationality

Under the assumption of common knowledge of rationality, rational agents play the limited bidding game randomly (see Appendix A), such that during each round, a rational agent randomly plays one of the still available tokens. However, experiments with human subjects have shown contexts in which humans regularly fail to behave as predicted by game theory [e.g. 15–19]. In reality, agents may not be fully rational, or consider their opponent to be fully rational. When agents repeatedly interact with the same opponent, they may show patterns of behaviour that deviate from random play, which may be used to their opponent's advantage. In this section, we tentatively describe the process of playing the limited bidding game by agents that are limited in their application of theory of mind. In the remainder, we will speak of a ToM_i agent to indicate an agent that has the ability to use theory of mind up to and including the i-th order. Also, to avoid confusion, we will refer to agents as if they were male, and opponents as if they were female.

Consider the situation in which a ToM_0 agent meets a ToM_1 opponent for the second time in the setting of the limited bidding game. During the first round of the game, suppose that the ToM_0 agent recalls that his opponent played token 1 in the first round of the last game. When deciding what token to play, a ToM_0 agent cannot make use of any theory of mind. In particular, a ToM_0 agent cannot consider the possibility that his opponent has goals that are competitive to his own. The only information available to the agent is that his opponent sometimes plays token 1 in the first round of the game. Against token 1, the best response is token 2, and thus the ToM_0 agent chooses to play token 2.

The ToM_1 opponent, on the other hand, forms beliefs about what the ToM_0 agent believes. She remembers that the last time she played against the ToM_0 agent, she selected token 1 in the first round. She reasons that if the situation were reversed, and she had been in the ToM_0 agent's position, she would conclude that the best response against token 1 is playing token 2. From this, the ToM_1 opponent concludes that the ToM_0 agent will be playing token 2. Against token 2, the best response is token 3, which is the token that the ToM_1 agent will select to play.

In our setup, none of the agents is aware of the abilities of his opponent. Through repeated interaction, a ToM_1 agent may come to believe that his opponent is not a ToM_0 agent, but that she does not have any beliefs at all, and plays according to some unchanging strategy. Based on this belief, a ToM_1 agents can choose to play as if he were a ToM_0 agent himself. Each agent forms and updates his beliefs through repeated interaction, in an attempt to uncover what order

of theory of mind he should use to win the game. The need for such learning becomes apparent for agents that make use of higher-order theory of mind. A ToM_2 agent, for example, engages in second-order theory of mind by forming beliefs about what his opponent believes him to believe. The implicit assumption in this modeling is that his opponent is a ToM_1 opponent that is able to form beliefs about what he believes. When in reality she is a ToM_0 opponent, the ToM_2 agent therefore attributes beliefs to his opponent that she cannot represent.

3 A mathematical model of theory of mind agents

In this section, we discuss the implementation of computational agents that are limited in their ability to make use of theory of mind while playing the limited bidding game, similar to the agents described in Section 2.2.

Computational agents in the limited bidding game represent the game situation by its observable features, that is, the set of tokens T that is still available to the agent and the set of tokens S that is available to his opponent. Based on this representation (T, S), an agent has beliefs in the form of a probability distribution $b^{(0)}$, such that $b^{(0)}(s; T, S)$ represents what the agent believes to be the probability that his opponent will play the token with value s in situation (T, S). A ToM_1 agent furthermore attributes beliefs to his opponent in the form of a distribution $b^{(1)}$, such that he believes his opponent to assign probability $b^{(1)}(t; S, T)$ to the event that he will play token t in situation (T, S). A ToM_2 agent maintains an additional belief structure $b^{(2)}$, such that he believes his opponent to believe that he assigns probability $b^{(2)}(s; T, S)$ to the event of her playing token s in situation (T, S).

Since an agent's beliefs $b^{(i)}$ represent probability distributions, we assume that they are non-negative and normalized such that $\sum_{s \in S} b^{(i)}(s; T, S) = 1$ for all $S \neq \emptyset$ and all orders of theory of mind i. Besides these beliefs, agents are governed by their confidence in the predictions based on application of first- and second-order theory of mind, c_1 and c_2 respectively, as well as learning speed λ and discounting rate δ. Unlike the beliefs $b^{(i)}$ and confidences c_i, an agent's learning speed λ and discounting rate δ, to be discussed later in this section, are fixed and agent-specific traits that are beyond the agent's ability to control.

To decide what token to use, agents make use of three basic functionalities: a value function Φ, a decision function t^* and a belief updating function Δ. The value function Φ is used to obtain a measure of the expected outcome of the game when playing token t in situation (T, S). This is achieved through

$$\Phi_{T,S}(t, b^{(i)}) = \begin{cases} \sum_{s \in S} b^{(i)}(s; T, S) \cdot sgn(t - s) & \text{if } |T| = 1 \\ \sum_{s \in S} b^{(i)}(s; T, S) \Big(sgn(t - s) + \delta \max_{t' \in T \setminus \{t\}} \Phi_{T \setminus \{t\}, S \setminus \{s\}}(t', b^{(i)}) \Big) & \text{if } |T| > 1, \end{cases} \quad (1)$$

where sgn is the signum function. Note that the value function Φ makes use of exponential time discounting with parameter $0 \leq \delta \leq 1$ [20, 21]. A higher value

of time discounting δ indicates that the agent is more patient, and more willing to lose the next round if it means winning the game.

Agents use the value function Φ to weigh the likelihood of winning the current round by playing token t against the value of the situation that results from losing token t for the remainder of the game. Based on beliefs $b^{(i)}$, agents decide what token to use according to the decision function

$$t^*_{T,S}(b^{(i)}) = \arg\max_{t \in T} \Phi_{T,S}(t, b^{(i)}). \tag{2}$$

Through application of theory of mind, agents come to believe that their opponent will be playing some token $\hat{s}$. The extent to which ith-order theory of mind governs the decisions of the agent's actions is determined by his confidence $0 \leq c_i \leq 1$ that ith-order theory of mind accurately predicts his opponent's behaviour. For every order of theory of mind available to the agent, he therefore adjusts his beliefs using the belief adjustment function Δ, given by

$$\Delta(b^{(i)}, \hat{s}, c_i)(s; T, S) = \begin{cases} (1 - c_i) \cdot b^{(i)}(s; T, S) & \text{if } s \neq \hat{s} \\ c_i + (1 - c_i) \cdot b^{(i)}(s; T, S) & \text{if } s = \hat{s}. \end{cases} \tag{3}$$

The functions Φ, t^* and Δ are shared by all agents, but the type of agent determines how these functions are used. A ToM_0 agent selects what token to play by using Equation (2) directly. That is, given discounting rate δ and zeroth-order beliefs $b^{(0)}$, a ToM_0 agent faced with situation (T, S) will play token $t^*_{T,S}(b^{(0)})$.

In contrast, ToM_1 agents consider the possibility that their opponent is playing as a ToM_0 agent. A ToM_1 agent makes use of this by determining what token he would play if the situation were reversed. To do so, a ToM_1 agent maintains first-order beliefs $b^{(1)}$ that describe what he would believe in his opponent's situation, and thus what he believes his opponent to believe. Using these beliefs, a ToM_1 agent can estimate what token his opponent will believe him to be playing by calculating $\hat{s}^{(1)} = t^*_{S,T}(b^{(1)})$.

Once a ToM_1 agent has derived what token $\hat{s}^{(1)}$ he would play in his opponent's situation, he adjusts his own beliefs $b^{(0)}$ to represent that he believes his opponent to play $\hat{s}^{(1)}$. That is, using the belief adjustment function Δ, a ToM_1 agent decides what token to use by calculating

$$t^*_{T,S}\Big(\Delta\big(b^{(0)}, \hat{s}^{(1)}, c_1\big)\Big) = t^*_{T,S}\left(\Delta\Big(b^{(0)}, t^*_{S,T}\big(b^{(1)}\big), c_1\Big)\right). \tag{4}$$

Note that in this sense, the computational agents described here represent their theory of mind according to *simulation-theory of mind* [22–24]. That is, rather than forming a *theory-theory of mind* [1, 25] that relates observable features of the world to unobservable mental states of their opponent through explicit hypotheses, agents simulate the mental content of their opponent in their own mind. A ToM_1 agent thus considers the mental states of his opponent by considering her viewpoint as if it were his own, implicitly assuming that this accurately describes her thought process. In this particular setting, this means that a ToM_1 agent makes use of his own discounting rate δ in determining $\hat{s}^{(1)}$, and therefore assumes his opponent to have the same rate of impatience he has.

Similar to the way a ToM_1 agent models his opponent as a ToM_0 agent, a ToM_2 agent determines what token he would play if he were in the position of his opponent, playing as a ToM_1 agent. In order to do so, a ToM_2 agent needs to specify his opponent's confidence in first-order theory of mind. In our experiments, we have assumed that all ToM_2 agents use a value of 0.8 to determine their opponent's behaviour playing as a ToM_1 agent, resulting in the estimate $\hat{s}^{(2)} = t^*_{S,T}\Big(\Delta\Big[b^{(1)}, t^*_{T,S}\big(b^{(2)}\big), 0.8\Big]\Big)$. This estimate is then used to update the ToM_2 agent's beliefs a second time before he makes his choice of what token to use. This choice can therefore be represented as

$$t^*_{T,S}\Big(\Delta\Big[\Delta\Big(b^{(0)}, \underbrace{t^*_{S,T}\big(b^{(1)}\big)}_{\hat{s}^{(1)}}, c_1\Big), \underbrace{t^*_{S,T}\Big(\Delta\Big[b^{(1)}, t^*_{T,S}\big(b^{(2)}\big), 0.8\Big]\Big)}_{\hat{s}^{(2)}}, c_2\Big]\Big). \tag{5}$$

To arrive at his decision of what token to play, an agent makes use of beliefs $b^{(i)}$, which are initialized randomly, and confidence levels c_i, which are initialized at zero. After each round, the actual choices of the agent $\tilde{t}$ and his opponent $\tilde{s}$ are revealed. At this moment, an agent updates his confidence in theory of mind based on the accuracy of its predictions. That is, given his agent-specific learning speed $0 \leq \lambda \leq 1$, a ToM_1 agent updates his confidence in first-order theory of mind c_1 according to

$$c_1 := \begin{cases} (1-\lambda)\cdot c_1 & \text{if } \tilde{s} \neq \hat{s}^{(1)} \\ \lambda + (1-\lambda)\cdot c_1 & \text{if } \tilde{s} = \hat{s}^{(1)}. \end{cases} \tag{6}$$

A ToM_1 agent thus increases his confidence in the use of first-order theory of mind if it yields accurate predictions, and lowers his confidence if predictions are inaccurate. A ToM_2 agent additionally adjusts his confidence in the use of second-order theory of mind c_2 according to

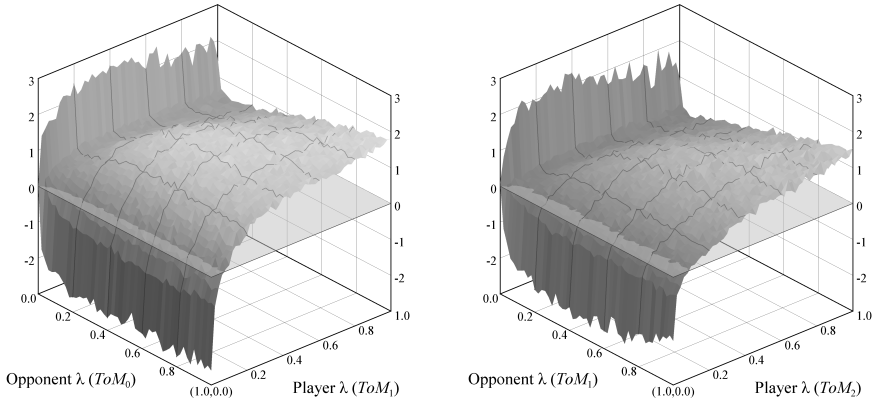
$$c_2 := \begin{cases} (1-\lambda)\cdot c_2 & \text{if } \tilde{s} \neq \hat{s}^{(2)} \\ c_2 & \text{if } \tilde{s} = \hat{s}^{(1)} = \hat{s}^{(2)} \\ \lambda + (1-\lambda)\cdot c_2 & \text{if } \tilde{s} = \hat{s}^{(2)} \neq \hat{s}^{(1)}. \end{cases} \tag{7}$$

This update is similar to the updating of the confidence in first-order theory of mind, except that a ToM_2 agent does not change his confidence in second-order theory of mind when first- and second-order theory of mind both yield correct predictions. That is, a ToM_2 agent only grows more confident in the use of second-order theory of mind when this results in accurate predictions that could not have been made with first-order theory of mind.

Finally, the agent also updates his beliefs $b^{(i)}$. For zeroth- and second-order beliefs $b^{(0)}$ and $b^{(2)}$, an agent updates his beliefs using his opponent's choice $\tilde{s}$, while first-order beliefs $b^{(1)}$ are updated using his own choice $\tilde{t}$, such that

$$b^{(i)}(s;T,S) := \Delta\Big(b^{(i)}, \tilde{s}, \lambda\Big)(s;T,S) \text{ for } i = 0, 2, \text{ and} \tag{8}$$

$$b^{(1)}(t;S,T) := \Delta\Big(b^{(1)}, \tilde{t}, \lambda\Big)(t;S,T). \tag{9}$$



(a) Average performance of a focal ToM_1 agent playing against a ToM_0 opponent.

(b) Average performance of a focal ToM_2 agent playing against a ToM_1 opponent.

Fig. 1: Effects of learning speed λ on average performance in a game of 5 tokens. Performance was determined as the average score over 50 trials, for every 0.02 increase of λ in the range $0 \leq \lambda \leq 1$. Discounting rate δ was fixed at 0.9.

The agents described above implicitly assume that their opponents update their beliefs using the same learning speed $0 \leq \lambda \leq 1$ as themselves. Furthermore, equations (8) and (9) maintain the normalization and non-negativity of beliefs, while the confidences c_1 and c_2 remain limited to the range $[0, 1]$. Finally, agents do not update their beliefs and confidence levels after the last round, in which they make the degenerate choice of playing the only token still available to them.

4 Results

The agents described in Section 3 have been implemented in Java and their performance has been tested in competition in a limited bidding game of five tokens. Performance per game was measured as the difference between the number of rounds an agent won and the number of rounds won by his opponent. Note that since it is not possible for an agent to win more than four out of five rounds[1], an agent's game score ranges from -3 to 3. Agents play against each other in trials that consist of 50 consecutive games. An agent's trial score is the average of the agent's game scores over all 50 games in the trial.

Figure 1 shows the advantage of making use of theory of mind as a function of the learning speed of the focal agent (λ_f) and his opponent (λ_o). Higher and lighter areas represent that the focal agent performed better than his opponent, while lower and darker areas show that his opponent obtained a higher average

[1] If an agent wins the first four rounds, the final round will be won by his opponent.

score. To emphasize the shape of the surface, the grid that appears on the bottom plane has been projected onto the surface.

Both figures show that an agent with learning speed $\lambda = 0$ cannot successfully compete with his opponent, and obtains a negative score. Note that in this case, the agent does not learn at all. Instead, he plays according to a fixed strategy, irrespective of his ability to use theory of mind.

Figure 1a shows that ToM_1 agents predominantly obtain a positive score when playing against ToM_0 opponents. The bright area along the line $\lambda_f = \lambda_o$ indicates that this advantage is again particularly high when learning speeds are equal. In this case, the ToM_1 agent's implicit assumption that his opponent has the same learning speed as himself is correct. Surprisingly, Figure 1a shows that even when the ToM_1 agent fails to accurately model his opponent, he will on average obtain a positive score for any learning speed $\lambda_f > 0.08$.

Figure 1b shows that ToM_2 agents obtain an advantage over ToM_1 opponents. However, although Figure 1b shows many of the same features as Figure 1a, such as the brighter area along the line $\lambda_f = \lambda_o$, ToM_2 agents playing against ToM_1 agents obtain a score that is on average 0.5 lower than the score of ToM_1 agents playing against ToM_0 agents. As a result, a ToM_2 agent needs a learning speed of at least $\lambda_f > 0.12$ in order to obtain, on average, a positive score when playing against a ToM_1 agent.

5 Discussion and future research

By making use of agent-based models, we have shown that in the competitive setting of the limited bidding game, the ability to make use of theory of mind presents individuals with an advantage over opponents that lack such an ability. This advantage presents itself even when an agent fails to model his opponent correctly, although an agent that accurately models his opponent obtains more of an advantage than an agent that over- or underestimates the speed at which his opponent learns from past behaviour. In competitive settings like the limited bidding game, there may therefore be an evolutionary incentive that justifies the application of higher-order theory of mind.

Our results also show diminishing returns on higher orders of theory of mind. Concretely, although second-order theory of mind agents outperform first-order theory of mind opponents, the advantage is not as high as for first-order theory of mind agents playing against zeroth-order theory of mind agents. Further evidence suggests that the advantage diminishes quickly for even higher orders of theory of mind (see Appendix B). This could help explain why humans have difficulty applying higher-order theory of mind correctly.

One possible direction for future research presents itself in the form of variable-frame level-n theory [16]. Variable-frame level-n theory expresses theory of mind as levels of bounded rationality, which an agent uses to model the behaviour of his co-player in the setting of a coordination game. An agent makes use of salience to determine what he believes his co-player to believe to be the best course of action, and selects his own action accordingly. In our competitive set-

ting, variable-frame level-n theory could be used to shape an agent's initial beliefs based on the salience of the tokens with the highest and lowest values. This could provide theory of mind agents with additional advantages early in the game.

Although we have shown that the use of theory of mind benefits individuals in the setting of the limited bidding game, in order to represent the beliefs they attribute to others, the higher-order theory of mind agents we describe need additional memory capacity. Based on additional experiments, it seems that the explicit attribution of mental content to competitors presents individuals with advantages beyond those of an increase in memory capacity (see Appendix C). That is, it seems that the advantage obtained by the application of theory of mind cannot be fully explained by an increase in memory capacity.

Acknowledgments

This work was supported by the Netherlands Organisation for Scientific Research (NWO) Vici grant NWO 277-80-001. We would also like to thank Rineke Verbrugge for her valuable comments and advice.

References

[1] Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences **1** (1978) 515–526
[2] Penn, D., Povinelli, D.: On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. Philosophical Transactions of the Royal Society B: Biological Sciences **362** (2007) 731
[3] Perner, J., Wimmer, H.: "John thinks that Mary thinks that...". Attribution of second-order beliefs by 5 to 10 year old children. Journal of Experimental Child Psychology **39** (1985) 437–71
[4] Hedden, T., Zhang, J.: What do you think I think you think?: Strategic reasoning in matrix games. Cognition **85** (2002) 1–36
[5] Flobbe, L., Verbrugge, R., Hendriks, P., Krämer, I.: Children's application of theory of mind in reasoning and language. Journal of Logic, Language and Information **17** (2008) 417–442
[6] Meijering, B., Van Maanen, L., Van Rijn, H., Verbrugge, R.: The facilitative effect of context on second-order social reasoning. In: Proceedings of the 32nd Annual Conference of the Cognitive Science Society, Cognitive Science Society (2010)
[7] Burkart, J., Heschl, A.: Understanding visual access in common marmosets, Callithrix jacchus: Perspective taking or behaviour reading? Animal Behaviour **73** (2007) 457–469
[8] Carruthers, P.: Meta-cognition in animals: A skeptical look. Mind & Language **23** (2008) 58–89
[9] Verbrugge, R.: Logic and social cognition: The facts matter, and so do computational models. Journal of Philosophical Logic **38** (2009) 649–680
[10] Axelrod, R.: The Evolution of Cooperation. Basic Books, New York (1984)

[11] Epstein, J.: Generative Social Science: Studies in Agent-based Computational Modeling. Princeton University Press, Princeton (NJ) (2006)
[12] Byrne, R., Whiten, A.: Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans. Oxford University Press, USA (1988)
[13] Whiten, A., Byrne, R.: Machiavellian Intelligence II: Extensions and Evaluations. Cambridge University Press, Cambridge (1997)
[14] De Bono, E.: Edward de Bono's Super Mind Pack: Expand Your Thinking Powers with Strategic Games & Mental Exercises. Dorling Kindersley Publishers Ltd, London, UK (1998)
[15] McKelvey, R., Palfrey, T.: An experimental study of the centipede game. Econometrica **60** (1992) 803–836
[16] Bacharach, M., Stahl, D.O.: Variable-frame level-n theory. Games and Economic Behavior **32** (2000) 220–246
[17] Stahl, D., Wilson, P.: On players' models of other players: Theory and experimental evidence. Games and Economic Behavior **10** (1995) 218–254
[18] Fehr, E., Gächter, S.: Cooperation and punishment in public goods experiments. American Economic Review (2000) 980–994
[19] Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R.: In search of homo economicus: Behavioral experiments in 15 small-scale societies. The American Economic Review **91** (2001) 73–78
[20] Green, L., Myerson, J.: Exponential versus hyperbolic discounting of delayed outcomes: Risk and waiting time. Integrative and Comparative Biology **36** (1996) 496
[21] Dimitri, N., van Eijck, J.: Time discounting and time consistency. In van Eijck, J., Verbrugge, R., eds.: Games, Actions and Social Software. Texts in Logic and Games (FOLLI subseries of LNCS). Springer Verlag, Berlin (2011)
[22] Davies, M.: The mental simulation debate. Philosophical Issues **5** (1994) 189–218
[23] Nichols, S., Stich, S.: Mindreading: An Integrated Account of Pretence, Self-awareness, and Understanding Other Minds. Oxford University Press, USA (2003)
[24] Hurley, S.: The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. Behavioral and Brain Sciences **31** (2008) 1–22
[25] Gopnik, A., Wellman, H.: Why the child's theory of mind really is a theory. Mind & Language **7** (1992) 145–171
[26] Osborne, M., Rubinstein, A.: A Course in Game Theory. The MIT press, Cambridge (MA) (1994)
[27] Bicchieri, C.: Common knowledge and backward induction: A solution to the paradox. In: Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge, Morgan Kaufmann Publishers Inc. (1988) 381–393
[28] Von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press, Princeton (NJ) (1944) Commemorative edition 2007.

Appendix A Rational agents in the limited bidding game

In game theory, it is common to make the assumption of common knowledge of rationality [26, 27]. In terms of theory of mind, this means that rational agents possess the ability to make use of theory of mind of any depth or order. In this section, we will explain how rational agents play the limited bidding game under the assumption of common knowledge of rationality.

For simplicity, we consider a limited bidding game of three tokens. In such a game, players decide what token to play at two moments: once at the start of the game, and again once the result of the first round has been announced. Although new information also becomes available after the second round, the choice of which token to play in the third round is a degenerate one; at the start of the third round both players only have one token left. Since both players have the choice of three tokens to play in the first round, there are nine variations of the subgame the agents play at the second round of the game. We first consider what a rational agent will choose to do at the start of the second round.

		Player 2					
		123	132	213	231	312	321
Player 1	123	(0,0)	(0,0)	(0,0)	(-1,1)	(1,-1)	(0,0)
	132	(0,0)	(0,0)	(-1,1)	(0,0)	(0,0)	(1,-1)
	213	(0,0)	(1,-1)	(0,0)	(0,0)	(0,0)	(-1,1)
	231	(1,-1)	(0,0)	(0,0)	(0,0)	(-1,1)	(0,0)
	312	(-1,1)	(0,0)	(0,0)	(1,-1)	(0,0)	(0,0)
	321	(0,0)	(-1,1)	(1,-1)	(0,0)	(0,0)	(0,0)

Table 1: Payoff table for the limited bidding game of three tokens. Each outcome of the game corresponds to a tuple in the table. The first value of the tuple is the payoff for player one, the second is the payoff for player two.

Since every player tries to maximize the number of rounds won and minimize the numbers of rounds lost, at the end of each game, each player receives a payoff equal to the difference between the two. Table 1 lists the payoffs for both players for each possible outcome of the game, where each outcome is represented as the concatenation of the tokens in the order in which the player has played them. Each payoff structure is presented as a tuple (x, y), such that player 1 receives payoff x and player 2 receives payoff y. The subgames that are played at the beginning of the second round are represented as 2-by-2 submatrices, highlighted by alternating background color in Table 1.

Note that whenever the first round of the game ends in a draw, the resulting subgame is a degenerate one. In this case, both players receive zero payoff irrespective of the final outcome. When the first round does not end in a draw, the

resulting subgame is a variation on the matching pennies game [28]. This game is known to have no pure-strategy Nash equilibrium. That is, there is no combination of pure strategies such that each player maximizes his payoff given the strategy of its opponent. However, there is a unique mixed-strategy Nash equilibrium in which each player plays each possible strategy with equal probability. If both players play either one of their remaining tokens with 50% probability, neither one of them has an incentive to switch strategies: given that its opponent is playing randomly, a rational agent has no strategy available that will yield a better expected payoff than playing randomly as well.

		Player 2		
		1	2	3
Player 1	1	(0.0,0.0)	(-0.5,0.5)	(0.5,-0.5)
	2	(0.5,-0.5)	(0.0,0.0)	(-0.5,0.5)
	3	(-0.5,0.5)	(0.5,-0.5)	(0.0,0.0)

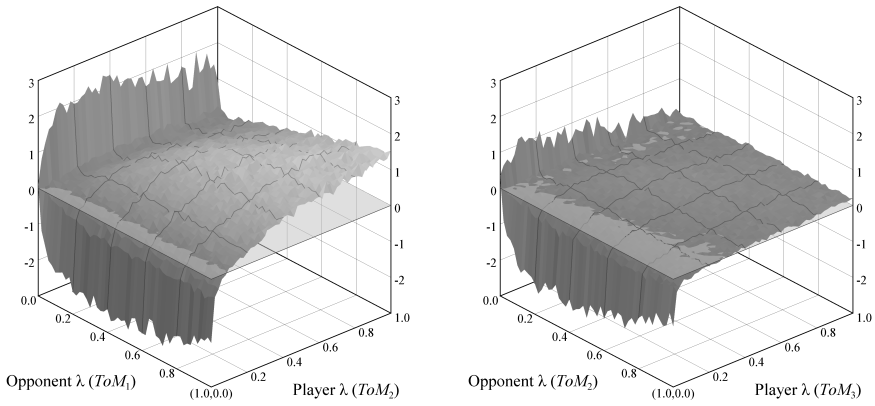
Table 2: Payoff table for the limited bidding game of three tokens once the players have derived that after the first round, both players will play randomly.

Due to the common knowledge of rationality, each player knows that both of them have reached the conclusion that after the first round, they will both play randomly. This means we can rewrite the payoff matrix to reflect the results of each of the subgames, as shown in table 2. Note that this is another variation of the matching pennies game with three strategies, also known as a stone-paper-scissors game [28]. As before, there is no pure-strategy Nash equilibrium, but the unique mixed-strategy Nash equilibrium is reached when both players play each strategy with equal probability. That is, rational agents, under the assumption of common knowledge of rationality, solve the limited bidding game by playing randomly at each round.

This result also holds when the game is played using more than three tokens. That is, to prevent their opponent from taking advantage of any regularity in their strategy, rational agents play the limited bidding game randomly.

Appendix B Limits of the advantage of theory of mind

In Section 4, we have shown that ToM_2 agents can obtain advantages that go beyond those obtained by ToM_1 agents. In this section, we extend the model of Section 3 to allow for ToM_3 agents. These agents possess an additional distribution $b^{(3)}$, such that a ToM_3 agent believes that his opponent believes that he believes her to assign probability $b^{(3)}(t; S, T)$ to him playing token t in situation (T, S). These beliefs are used to determine what token the ToM_3 agent would play if he were in the position of his opponent, and playing as a ToM_2 agent.



(a) Average performance of a focal ToM_2 agent playing against a ToM_1 opponent.

(b) Average performance of a focal ToM_3 agent playing against a ToM_2 opponent.

Fig. 2: Effects of learning speed λ on average performance in a game of 5 tokens. Performance was determined as the average score over 50 trials, for every 0.02 increase of λ in the range $0 \leq \lambda \leq 1$. Discounting rate δ was fixed at 0.9.

The ToM_3 agent considers a 'pure' ToM_2 agent, such that he specifies $c_1 = 0.0$ and $c_2 = 0.8$ for his opponent. The confidence in first-order theory of mind that he believes her to assign to him is $c_1 = 0.8$. This results in the estimate

$$\hat{s}^{(3)} = t^*_{S,T}\Big(\Delta\Big[b^{(1)}, t^*_{T,S}\Big(\Delta\Big[b^{(2)}, t^*_{S,T}\big(b^{(3)}\big), 0.8\Big]\Big), 0.8\Big]\Big). \tag{10}$$

This estimate is then used to update the ToM_3 agent's beliefs a third time before he makes his choice of what token to use. This choice therefore is

$$t^*_{T,S}\Big(\Delta\Big\{\Delta\Big[\Delta\Big(b^{(0)}, \underbrace{t^*_{S,T}\big(b^{(1)}\big)}_{\hat{s}^{(1)}}, c_1\Big), \underbrace{t^*_{S,T}\Big(\Delta\Big[b^{(1)}, t^*_{T,S}\big(b^{(2)}\big), 0.8\Big]\Big)}_{\hat{s}^{(2)}}, c_2\Big], \underbrace{t^*_{S,T}\Big(\Delta\Big[b^{(1)}, t^*_{T,S}\Big(\Delta\Big[b^{(2)}, t^*_{S,T}\big(b^{(3)}\big), 0.8\Big]\Big), 0.8\Big]\Big)}_{\hat{s}^{(3)}}, c_3\Big\}\Big). \tag{11}$$

This agent has been implemented in Java and placed in competition with the ToM_2 agent described in Section 3. The results are shown in Figure 2b. For convenience, the average performance of a ToM_2 agent playing against a ToM_1 opponent has been repeated in Figure 2a. As Figure 2b shows, a ToM_3 agent barely outperforms a ToM_2 agent. The average score only exceeds 0.3 when the ToM_2 opponent has zero learning speed. Although it appears as if a ToM_3 agent can still on average obtain a positive score when his learning speed is at least $\lambda > 0.32$, Figure 2b shows that when the ToM_2 opponent has learning speed $0 < \lambda < 0.1$, performance of the ToM_3 agent may still fall below zero.

Interestingly, the poor performance of ToM_3 agents playing against ToM_2 opponents is partially caused by the model that the ToM_2 opponent holds of the ToM_3 agent. Note that since confidence levels c_i are initialized at zero, all agents start out by playing as ToM_0 agents. When a focal ToM_3 agent is in competition with a ToM_2 opponent, both of them will notice that their predictions based on first-order theory of mind $\hat{s}^{(1)}$ are correct. Through Equation (6), this causes both agents to grow more confident in application of first-order theory of mind. As a result, they both gradually start playing more as a ToM_1 agent. When this happens, predictions based on first-order theory of mind $\hat{s}^{(1)}$ will become less accurate, but predictions based on second-order theory of mind $\hat{s}^{(2)}$ become increasingly accurate, increasing confidence in the application of second-order theory of mind through Equation (7). Both the focal agent and his opponent will therefore start playing as a ToM_2 agent. At this point, the opponent can no longer model the focal agent. That is, she will notice that none of her predictions are correct and start to play as a ToM_0 agent again. However, when the focal agent tries to take advantage of this by playing as a ToM_1 agent, the opponent recognizes this and once again grows more confident in her predictions based on second-order theory of mind. This causes the ToM_2 opponent to constantly keep changing her strategy, which hinders the ToM_3 agent's efforts of trying to model her behaviour.

Appendix C Theory of mind is more than increased memory

The results in Section 4 show that the use of a theory of mind benefits individuals in the setting of the limited bidding game. However, in order to represent the beliefs they attribute to others, the higher-order theory of mind agents we described in Section 3 need additional memory capacity; for every additional order of theory of mind available to the agent, it maintains another belief structure $b^{(i)}$. In this section, we consider the high-memory ToM_0 agent, which has the ability to remember what token $t_{T,S}^{(-1)}$ he played the last time in any game situation (T, S). The high-memory ToM_0 agent makes use of this by representing beliefs of the form $b_{Mem}^{(0)}$, such that he believes that the probability of his opponent playing token s in situation (T, S) is $b_{Mem}^{(0)}(s; T, S, t_{T,S}^{(-1)})$. That is, the high-memory ToM_0 agent has different beliefs concerning what his opponent will play in situation (T, S) based on the last token he played in the same situation.

To determine whether the contribution of theory of mind to an agent's performance can be explained by additional memory alone, we placed the high-memory ToM_0 agent in competition with the ToM_2 agent described in Section 3, both of which have similar demands on memory capacity. The number of game situations in which a player makes a non-trivial choice of what token to play is $\sum_{i=0}^{N-2} \binom{N}{i}^2$. For a game of five tokens, there are 226 such situations. Since a ToM_2 agent needs to maintain three belief structures $b^{(i)}$, a ToM_2 agent needs enough

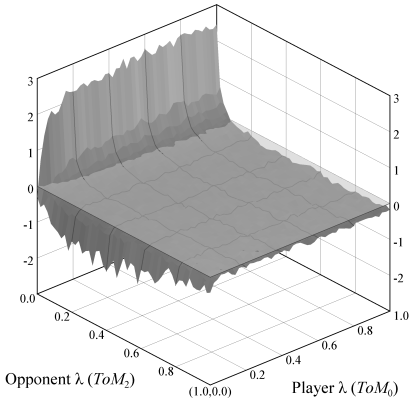


Fig. 3: Effects of learning speed λ on the average performance of a high-memory ToM_0 agent playing against a (low-memory) ToM_2 opponent in a game of 5 tokens. Performance was determined as the average score over 50 trials, for every 0.02 increase of λ in the range $0 \leq \lambda \leq 1$. Discounting rate δ was fixed at 0.9.

memory to represent 678 beliefs. A high-memory ToM_0 agent has a richer representation of the game, which causes him to consider $\sum_{i=0}^{N-2}(N-i)\binom{N}{i}^2$ game situations in which he makes a non-trivial choice of what token to play. In addition to remembering his last choice $t_{T,S}^{(-1)}$ in 226 situations, the high-memory ToM_0 agent therefore needs enough memory to represent 605 beliefs to maintain his belief structure $b_{Mem}^{(0)}$.

Note that the high-memory ToM_0 agent represents an unpredictable opponent for the ToM_2 agent. A ToM_2 agent models the behaviour of his opponent by considering his own actions in her situation. However, the representation of the game situation held by a high-memory ToM_0 agent differs from that of his low-memory ToM_2 opponent. That is, the ToM_2 opponent fails to accurately model the high-memory ToM_0 agent.

Figure 3 show the average performance of a high-memory ToM_0 agent when playing against a low-memory ToM_2 opponent. Surprisingly, even though the ToM_2 opponent is unable to effectively use her theory of mind, she outperforms the high-memory ToM_0 agent whenever her learning speed $\lambda > 0.12$. On average, a high-memory ToM_0 agent scores -0.20 when playing against a ToM_2 opponent.

A possible reason for the negative score of the high-memory ToM_0 agents, even though their ToM_2 opponent is unable to accurately model them, may be the length of the trials. In our setup, trials consist of 50 consecutive games, which may not provide a high-memory ToM_0 agent with sufficient information to gain an advantage over his ToM_2 opponent. In contrast, although the ToM_2 opponent incorrectly models the high-memory ToM_0 agent, her model is accurate enough to obtain a reliable advantage.