Logic and Interactive RAtionality

Yearbook 2012

Volume II

Editors:
Zoé Christoff
Paolo Galeazzi
Nina Gierasimczuk
Alexandru Marcoci
Sonja Smets

Collecting Editors:
Alexandru Baltag (Europe)
Wesley Holliday (United States)
Fenrong Liu (China)

The printing of this book was supported by the UP fund of Johan van Benthem.
Cover design by Nina Gierasimczuk.

Contents of Volume II

Reasoning about Short Sight and Solutions in Extensive-Form Games

Chanjuan Liu, Fenrong Liu, and Kaile Su

Peking University
Tsinghua University
Griffith University
chanjuan.pkucs@gmail.com, fenrong@tsinghua.edu.cn, kailepku@gmail.com

Abstract

The notion of *short sight*, introduced by Grossi and Turrini, weakens the unrealistic assumption in traditional extensive games that every player is able to perceive the entire game structure. We develop new solution concepts for games with short sight and propose a new logic language for reasoning in such games. We then present an axiomatization for this logic. In addition, we show that the logic can formally characterize the solution concepts in games with short sight.

1 Introduction

The research direction to integrate logic and game theory has received considerable interest in recent years. Works in this line not only provide logical tools for reasoning about rationality and decision making, but also import game-theoretic notions into the realm of logic. Extensive games are those games allowing for sequencing players' possible moves, and their choices at every decision point. To characterize the structures and reason about the solution concepts of extensive games, much work has been done to provide the logical systems for such games. These logic systems focus on various perspectives of extensive games: Harrenstein et al. (2003) concentrated on describing equilibrium concepts and strategic reasoning. Lorini and Moisan (2011) proposed an epistemic logic to deal with epistemic aspects of extensive games. Van Otterloo et al.

(2004) introduced a logic reasoning about how information or assumptions about the preferences of other players can be used by agents in order to realize their own preferences. The work of Parikh (1985) on propositional game logic initiated the study of game structure using algebraic properties. Van Benthem(2002) used dynamic logic to describe games as well as strategies. Ramanujam and Simon (2008) studied a logic in which not only are games structured, but so also are strategies. Bonanno et al. (2003) worked on the relationship of branching time logic to extensive form games. However, these logics all work on traditional extensive game models, which explicitly assume that the entire structure of a game is common knowledge to all players.

The assumption of common knowledge on game structures in traditional extensive games is sometimes too strong and unrealistic. For instance, in a game like chess, the actual game space is exponential in the size of the game configuration, and may have a computation path too long to be effectively handled by most existing computers. So we often seek sub-optimal solutions by considering only limited information or bounded steps foreseeable by a player that has relatively small amount of computation resources. Grossi and Turrini proposed the concept of *games with short sight* (Grossi and Turrini 2012), in which players can only see part of the game tree. However, there is no work on the logical reasoning of the strategies and solution concepts in this game model.

Inspired by the previous logics for extensive games, this paper is devoted to the logical analysis of game-theoretical notions of the solutions concepts in games with short sight. In (Harrenstein et al. 2003), a logic for strategic reasoning and equilibrium concepts was developed, which is closest to ours in spirit. Whereas, what we present here is a new logical system LS for games with *short sight*. This logic deploy the additional modalities $[\prec]$, $[(\sigma_i)]$, $[\mathring{\sigma}^s]$, etc. to capture several new features such as restricted sight and limited steps. We also give an axiomatization for the logic. Further, we show that this logic can be used to characterize some properties of games with short sight.

The structure is as follows: The next section introduces the definition of traditional extensive games and solution concepts in such games. Then we study the model of games with short sight and analyze its solution concepts corresponding to that of general extensive games. After that, we present the logical system LS. Finally, we concludes the paper with further research issues.

2 Extensive games and the solution concepts

In this section, we introduce finite games in extensive form with perfect information, and three solution concepts including best response, Nash equilibrium and subgame perfect equilibrium. First, we recall the definition of extensive games with perfect information.

Definition 2.1 (Extensive game (with perfect information))**.** A finite extensive game (with perfect information) is a tuple G=$(N, V, A, t, \Sigma_i, \succeq_i)$, where (V, A) is a tree with V, a set of nodes or vertices including a root v_0, and $A \subseteq V^2$ a set of arcs. N is a non-empty set of the players, and $\succeq_i$ represents preference relation for each player i, which is a partial order over V. For any two nodes v and v', if $(v, v') \in A$, we call v' a *successor* of v, thus A is also regarded as the successor relation. Leaves are the nodes that have no successors, denoted by Z. t is turn function assigning a member of N to each non-terminal node. Σ_i is a non-empty set of strategies. A strategy of player i is a function $\sigma_i : \{v \in V \backslash Z |\ t(v) = i\} \to V$ which assigns a successor of v to each non-terminal node when it is i's turn to move.

As usual, $\sigma = (\sigma_i)_{i \in N}$ represents a strategy profile which is a combination of strategies from all players and Σ represents the set of all strategy profiles. For any $M \subseteq N$, σ_{-M} denotes the collection of strategies in σ excluding those for players in M. We define an outcome function $O : \Sigma \to Z$ assigning leaf nodes to strategy profiles, i.e., $O(\sigma)$ is the outcome if the strategy profile σ is followed by all players. $O(\sigma_{-M})$ is the set of outcomes players in M can enforce provided that the other players strictly follow σ. $O(\sigma'_i, \sigma_{-i})$ is the outcome if player i use strategy σ' while all other players employ σ.

Preference relation here is different from the conventional ones: In the literature the notion of preference is assumed to be a linear order over leaves, while in this paper it is a partial order over all nodes in V. We assume that players may not be able to precisely determine entire computation paths leading to leave nodes, and allow them to make estimations or even conjecture a preference between non-terminal nodes. This assumption also provides technical convenience for discussing games with short sight later.

Example 1. Consider the Tic-Tac-Toe game shown by Figure 1 (Part of the game tree is omitted). There are two players: player 1(×) and player 2(∘). The initial state is v_0. v_1,v_2,v_3 are all successors of v_0. v_{10},v_{11},v_{12} are the terminal nodes (leaves). The solid arrows show the moves of player 1 and dotted arrows for player 2. Formally, $N = \{1, 2\}$; $V = \{v_0, v_1, v_2 \cdots\}$; (v_0, v_1), (v_0, v_2),$(v_0, v_3) \in A$; $v_{10}, v_{11}, v_{12} \in Z$; $t(v_4) = t(v_5) = t(v_6) = 2$. Since player 1 wins the game in v_{12}, loses it in v_{10}, and gains a draw in v_{11}, it naturally follows that $v_{12} \succeq_1 v_{11} \succeq_1 v_{10}$ for her. There are different strategies for each player. For instance, a σ_1 such that $\sigma_1(v_0) = v_2$, $\sigma_1(v_5) = v_8$, etc. And a σ'_1 such that $\sigma'_1(v_0) = v_3$, etc. And thus there is a strategy profile $\sigma = (\sigma_1, \sigma_2)$ such that $O(\sigma) = v_{11}$.

Solution concept is a significant notion in game theory. Concerning different aspects, there are various solution concepts for extensive games. The following definition

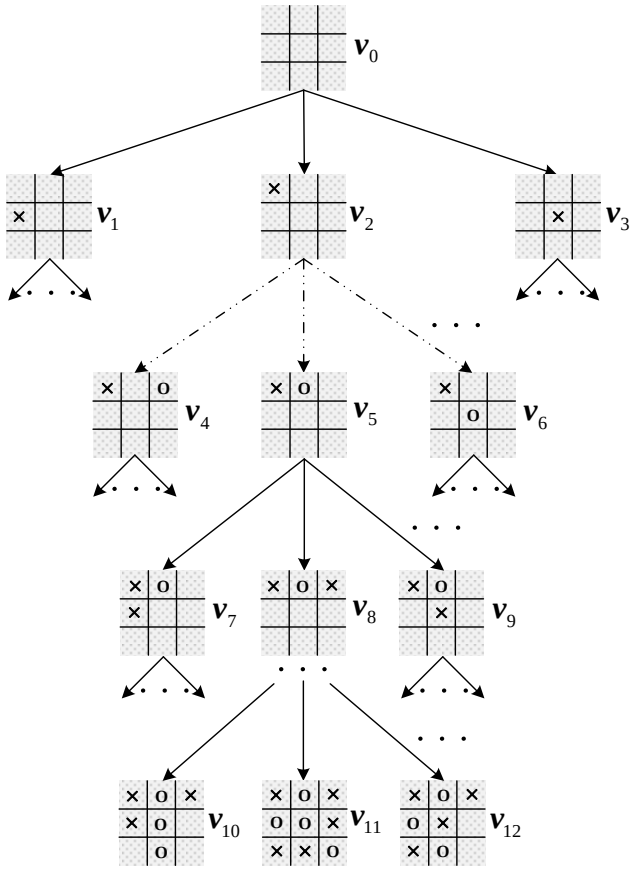


Figure 1: A Tic-Tac-Toe game

presents two of them: best response and Nash equilibrium, which are adapted from traditional notions (Fudenberg and Tirole 1991, Nash 1950). These two solutions ignore the sequential structure of games.

Definition 2.2 (Best response and Nash equilibrium)**.** A best response for player i of an extensive game is a strategy profile σ^* such that $O(\sigma_i^*, \sigma_{-i}^*) \succeq_i O(\sigma_i, \sigma_{-i}^*)$ for every strategy σ_i of player i. A strategy profile σ^* is a Nash equilibrium of an extensive game if it is a best response for every player i.

Another solution concepts for extensive games is the one taking the sequential structure of the game into account, i.e., subgame perfect equilibrium.

Definition 2.3 (Subgame perfect equilibrium)**.** Take a finite extensive game G. A strategy profile σ^* is a subgame perfect equilibrium (SPE) if for every player i, node v for which $t(v) = i$, it holds that $O|_v(\sigma_i^*|_v, \sigma_{-i}^*|_v) \succeq_i O|_v(\sigma_i, \sigma_{-i}^*|_v)$, for every strategy σ_i available to i in the subgame $G|_v$ of G that follows node v.

3 Games with Short Sight

In this section, we introduce games with short sight proposed by Grossi and Turrini (2012). In these games, players' sight is limited in the sense that they are not able to see the nodes in some branches of the game tree or have no access to some of the terminal nodes.

3.1 Short Sight

The following definition makes the notion of *short sight* mathematically precise.

Definition 3.1 (Sight function)**.** Let $G = (N, V, A, t, \Sigma_i, \succeq_i)$ be an *extensive game*. A short sight function for G is a function $s : V \backslash Z \to 2^{V|_v} \backslash \emptyset$, associating to each non-terminal node v a finite subset of all the available nodes at v, and satisfying:

$v' \in s(v)$ implies that $v'' \in s(v)$ for every $v'' \lhd v'$ with $v'' \in V|_v$, i.e. players' sight is closed under prefixes.

Intuitively, function s associates any choice point with vertices that each player can see.

Example 2. Consider the game in Example 1, one possible sight at v_0 is $s(v_0) = \{v_0, v_1, v_2, v_3, v_4, v_5\}$.

Definition 3.2 (Extensive game with short sight)**.** An extensive game with short sight (Egss) is a tuple $S = (G, s)$ where G is a finite extensive game and s a sight function for G.

3.2 Sight filtration and solution concepts of Egss

Each game with short sight yields a family of finite extensive games, one for each non-terminal node $v \in V \setminus Z$:

Definition 3.3 (Sight-filtrated extensive game)**.** Let S be an Egss given by (G, s) with $G{=}(N, V, A, t, \Sigma_i, \succeq_i)$. Given any non-terminal node v, a tuple $S\lceil_v$ is a finite extensive game by sight-filtration: $S\lceil_v{=}\ (N\lceil_v, V\lceil_v, A\lceil_v, t\lceil_v, \Sigma_i\lceil_v), \succeq_i\lceil_v)$ where

- $N\lceil_v{=}\ N$;
- $V\lceil_v{=}\ s_{t(v)}(v)$, which is the set of nodes within the sight of $t(v)$ from node v. The terminal nodes in $V\lceil_v$ are the nodes in $V\lceil_v$ of maximal distance, denoted by $Z\lceil_v$;
- $A\lceil_v{=}\ A \cap (V\lceil_v)^2$;
- $t\lceil_v{=}\ V\lceil_v \backslash Z\lceil_v \rightarrow N$ so that $t\lceil_v(v') = t(v, v')$;
- $\Sigma_i\lceil_v$ is the set of strategies for each player available at v and restricted to $s(v)$. It consists of elements $\sigma_i\lceil_v$ such that $\sigma_i\lceil_v(v') = \sigma_i(v, v')$ for each $v' \in V\lceil_v$ with $t\lceil_v(v') = i$;
- $\succeq_i\lceil_v{=}\ \succeq_i \cap (V\lceil_v)^2$.

Accordingly, we define the outcome function $O\lceil_v$: $\Sigma\lceil_v \rightarrow Z\lceil_v$ assigning leaf nodes of $S\lceil_v$ to strategy profiles.

Example 3. For the case of Example 2, the sight-filtrated extensive game $S\lceil_{v_0}$ could be shown in Figure 2. The set of players remain unchanged; $V\lceil_{v_0}{=}\ \{v_0, v_1,$
$v_2, v_3, v_4, v_5\}$, $Z\lceil_{v_0}{=}\ \{v_4, v_5\}$; $(v_2, v_4) \in A\lceil_{v_0}, (v_2, v_5) \in A\lceil_{v_0}$; $t\lceil_{v_0}(v)$ is consistent with t for any $v \in V\lceil_{v_0}$; $\Sigma_i\lceil_{v_0}$ and $\succeq_i\lceil_{v_0}$ are all restricted to the states that are within sight $s(v_0)$. E.g., for the strategy profile σ (defined in Example 1), $\sigma\lceil_{v_0}{=}\ (\sigma_1\lceil_{v_0}, \sigma_2\lceil_{v_0})$ such that $O\lceil_{v_0}(\sigma\lceil_{v_0}) = v_5$, with $\sigma_1\lceil_{v_0}(v_0) = v_2$ and $\sigma_2\lceil_{v_0}(v_2) = v_5$.

Corresponding to the solution concepts in extensive games, we define sight-compatible solutions for games with short sight.

Definition 3.4 (Sight-compatible best response and Nash equilibrium)**.** Let $S = (G, s)$ be an Egss and $S\lceil_v$ be the sight-filtrated extensive game at v. A strategy profile σ^* is a sight-compatible best response for i if for every nonterminal node v, it holds that $O\lceil_v(\sigma_i^*\lceil_v, \sigma_{-i}^*\lceil_v) \succeq_i\lceil_v O\lceil_v(\sigma_i\lceil_v, \sigma_{-i}^*\lceil_v)$ for any strategy $\sigma_i\lceil_v$ available to i. σ^* is a *sight-compatible Nash equilibrium*(SCNE) of S if it is a sight-compatible best response for every player $i \in N$.

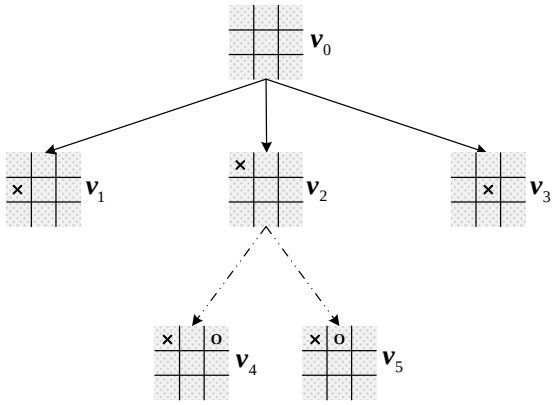


Figure 2: Sight-filtrated extensive game $S\lceil_{v_0}$

A sight-compatible best response for player i (Nash equilibrium) in Egss S is consistent with a best response for player i (Nash equilibrium) in each sight-filtrated extensive game $S\lceil_v$. Here is one aspect worth illustrating: at any point v, current player i is facing with a game $S\lceil_v$ determined by her sight. She attributes to her opponents the ability to see as much as she can see, supposing they are playing the same game $S\lceil_v$. Considering the actual process of playing games, this might be a conservative but realistic way for i to make decisions.

There is another solution for Egss, matching the notion of SPE for extensive games in Definition 2.3. The main idea lies in the following analysis: Given a Egss S, at each decision point v, player $t(v)$ is facing a sight-filtrated extensive game $S\lceil_v$. What he can achieve the best is to find a successor node of v maximizing his own profit within his current sight, which is the subgame perfect equilibrium of $S\lceil_v$. The players play in turns, choosing a best successor node at each point, until reaching a terminal node of S. Thus the crucial task for solving games with short-sighted players is in searching for the SPE of $S\lceil_v$ at each intermediate node v. The definition below is adapted from (Grossi and Turrini 2012).

Definition 3.5 (Sight-compatible subgame perfect equilibrium)**.** Let $S = (G, s)$ be an Egss and $S\lceil_v$ be the sight-filtrated extensive game at v. A strategy profile σ^* is a sight-compatible SPE of S if for every nonterminal node v, there exists a strategy profile $\sigma\lceil_v$ that is a subgame perfect equilibrium of $S\lceil_v$ and $\sigma_{t(v)}\lceil_v(v) = \sigma^*_{t(v)}(v)$.

Intuitively, sight-compatible SPE is the strategy profile that is in accordance with the SPE of the sight-filtrated game $S\lceil_v$ at each decision point v.

The lemma below paves the way for the result in Section 5, indicating that each sight-filtrated extensive game is a special extensive game with short sight, in which players can see the whole subtree that follows the current node.

Lemma 1. *Let* $G'=(N', V', t', \Sigma'_i, \succeq'_i)$[1] *be a sight-filtrated extensive game of* S *defined as* $(N, V, t, \Sigma_i, \succeq_i, s)$. *Then* G' *can be seen as a game with short sight* $G'_{s'}=(N', V', t', \Sigma'_i, \succeq'_i, s')$ *with* $G'_{s'}\lceil_u = G'|_u$ *for any* $u \in V'$.

4 A logic of extensive games with Short Sight

In this section we present a modal logic LS (Logic of Extensive Games with Short sight) in three steps. This logic supports reasoning about strategies and solutions in extensive games with short sight.

4.1 $\mathcal{L}$: The first step

A language for general extensive games is proposed in (Harrenstein et al. 2003), in which a strategy profile is taken as a modal operator, corresponding to an accessibility relation connecting a non-terminal node to leaf nodes. This language makes strategic reasoning simple, since one only needs to consider the outcome of this strategy without getting confused with all the actions at every choice point. To characterize what players can see in extensive games with short sight, we extend their language mainly by adding the modality $[\lessdot]$. Let P be the set of propositional variables, and Σ be the set of strategy profiles. The language $\mathcal{L}$ is given by the following BNF:

$$\varphi ::= p| \neg\varphi| \varphi_0 \wedge \varphi_1| \langle\leq_i\rangle\varphi| \langle\mathring{\sigma}\rangle\varphi| \langle\mathring{\sigma}_{-i}\rangle\varphi| \langle\lessdot\rangle\varphi$$

where $p \in P$, $\sigma \in \Sigma$. As usual, The dual of $\langle . \rangle\varphi$ is $[.]\varphi$. We begin with a brief explanation of the intuition behind the logic.

- The label $\leq_i$ denotes player i's preference relation.
- The label $\mathring{\sigma}$ stands for the outcomes of strategy profiles. $(v, v') \in R_{\mathring{\sigma}}$ iff v' is the terminal node reached from v by following σ.
- $(v, v') \in R_{\mathring{\sigma}_{-i}}$ iff v' is one of the leaf nodes extending v that player i can enforce provided that the other players strictly follow their strategies in σ.

[1] Here we use G' to denote sight-filtrated extensive game $S\lceil_v$, to avoid the complicated symbol $S\lceil_v|_u$.

- The label $\lessdot$ is sight function for the current player, and $\langle\lessdot\rangle\varphi$ means "φ holds in some node within the player $t(v)$'s sight at the present node v."

Let $S = (N, V, A, t, \Sigma_i, \succeq_i, s)$ be an Egss. The tuple of $(V, R_{\leq_i}, R_{\mathring{\sigma}}, R_{\mathring{\sigma}_{-i}}, R_{\lessdot})$ is defined as the frame F_S for $\mathcal{L}$, where for each player i, strategy profile σ, nodes v, v', the accessibility relations are given as follows.

$$\begin{array}{lll} vR_{\leq_i}v' & \text{iff} & v' \succeq_i v \\ vR_{\mathring{\sigma}}v' & \text{iff} & v' = O|_v(\sigma|_v) \\ vR_{\mathring{\sigma}_{-i}}v' & \text{iff} & v' \in O|_v(\sigma_{-i}|_v) \\ vR_{\lessdot}v' & \text{iff} & v' \in s_{t(v)}(v) \end{array}$$

A model M for $\mathcal{L}$ is a pair (F, π) where F is a frame for $\mathcal{L}$ and π a function assigning to each proposition p in P a subset of V, i.e., $\pi : P \to 2^V$. The interpretation for $\mathcal{L}$ formulas in model M are defined as follows:

$$\begin{array}{lll} M, v \models p & \text{iff} & v \in \pi(p). \\ M, v \models \neg\varphi & \text{iff} & \text{not } M, v \models \varphi. \\ M, v \models \varphi \wedge \psi & \text{iff} & M, v \models \varphi \text{ and } M, v \models \psi. \\ M, v \models \langle\leq_i\rangle\varphi & \text{iff} & M, u \models \varphi \text{ for some } u \in V \text{ with } vR_{\leq_i}u. \\ M, v \models \langle\mathring{\sigma}\rangle\varphi & \text{iff} & M, u \models \varphi \text{ for some } u \in V \text{ with } vR_{\mathring{\sigma}}u. \\ M, v \models \langle\mathring{\sigma}_{-i}\rangle\varphi & \text{iff} & M, u \models \varphi \text{ for some } u \in V \text{ with } vR_{\mathring{\sigma}_{-i}}u. \\ M, v \models \langle\lessdot\rangle\varphi & \text{iff} & M, u \models \varphi \text{ for some } u \in V \text{ with } vR_{\lessdot}u. \end{array}$$

Example 4. To illustrate the language, take S as the game in Example 2. Suppose $O(\sigma) = v_{11}$ and $O(\sigma_{-1}, \sigma_1^*) = v_{10}$. Let M be the model for S in which $\pi(p) = \{v_4, v_{11}, v_{12}\}$. Then we have the following:

- $M, v_8 \models \langle\mathring{\sigma}_{-1}\rangle\neg p$. I.e., p is false at one of the outcomes $O(\sigma_{-1})$ (namely v_{10}) that extends v_8 .
- $M, v_5 \models \langle\mathring{\sigma}\rangle\langle\leq_1\rangle p$. I.e., p is true at some node (namely, v_{12}) which is preferred by player 1 to the terminal node $O(\sigma)$ (namely, v_{11}) that extends v_5.
- $M, v_0 \models \langle\lessdot\rangle\langle\mathring{\sigma}\rangle p$. I.e., there is a node v (exactly, v_2), that can be seen at v_0 and the terminal node $O(\sigma)$ (namely, v_{11}) that extends v satisfies p.

The validities of a formula φ in models and frames are the same as the standard definitions (van Benthem 2010, Blackburn et al. 2001).

We now present the valid principles of the logic L. First, we have the following standard axioms.

(A_0) *Taut*, any classical tautology.

(A_1) K axiom for modalities $[\leq_i]$, $[\mathring{\sigma}]$, $[\mathring{\sigma}_{-i}]$, $[\lessdot]$.

Table 1 lists the other valid principles of L. The first column (N) is the *name* of the principle. The second column denotes the *modalities* that each principle is applied to.

The third column shows the formula *schema*. The fourth column describes the *property* of the corresponding accessibility relation R.

N	Modality	Schema	Property
T	$[\leq_i]$ $[\lessdot]$	$[\leq_i]\varphi \to \varphi$ $[\lessdot]\varphi \to \varphi$	reflexivity
4	$\leq_i]$	$[\leq_i]\varphi \to [\leq_i][\leq_i]\varphi$	transtivity
D	$[\mathring{\sigma}]$	$[\mathring{\sigma}]\varphi \leftrightarrow \langle\mathring{\sigma}\rangle\varphi$	determinism
I	$([\mathring{\sigma}], [\mathring{\sigma}_{-i}])$	$[\mathring{\sigma}_{-i}]\varphi \to [\mathring{\sigma}]\varphi$	inclusiveness
M	$[\mathring{\sigma}]$ $[\mathring{\sigma}_{-i}]$	$[\mathring{\sigma}]([\mathring{\sigma}']\varphi \leftrightarrow \varphi)$ $[\mathring{\sigma}_{-i}]([\mathring{\sigma}'_{-i}]\varphi \leftrightarrow \varphi)$	terminating

Table 1: Valid principles of L

K is used in all variants of the standard modal logic. T and 4 determine the preference of players to be *reflexive* and *transitive*. The sight of a player is reflexive. D ensures that a node reachable by a strategy profile σ from a node v is *determined*. I says that every outcome of strategy σ is *included* in the sets of outcomes by letting i free, and the other players sticking to σ. M guarantees the final outcome vertices to be *terminated*.

The inference rules for L are Modus Ponens (*MP*) and Necessitation (*Nec*) for operators $\mathring{\sigma}, \mathring{\sigma}_{-i}, \lessdot$ and $\leq_i$.

4.2 $\mathcal{L}(\mathring{\sigma}^s, \mathring{\sigma}^s_{-i})$: further extension

Now we enrich language $\mathcal{L}$ to also describe the outcomes in sight-filtrated extensive games. We add two modalities into the language $\mathcal{L}$. Intuitively, $\langle\mathring{\sigma}^s\rangle\varphi$ means "φ holds in some state v' in $S\lceil_v$, which is the terminal node of $S\lceil_v$ that is reachable from the starting point v when all players adopt the strategy profile σ, i.e., $v' = O\lceil_v(\sigma\lceil_v)$." The interpretation for $\langle\mathring{\sigma}^s_{-i}\rangle\varphi$ is similar. To show the reason of introducing these two modalities as new operators, we prove that they are undefinable by $\mathcal{L}$ through bisimulation:

Example 5. Consider the following (Figure 3) two games S_1, S_2, and $\mathcal{L}$-models M_1, M_2 for them respectively (Solid arrow represents R_s, while dotted arrow represents $R_{\mathring{\sigma}}$). p is true at w. Then obviously, M_1 and M_2 are bisimilar with respect to the $\mathcal{L}$-models.[2] Suppose we could define $\langle\mathring{\sigma}^s\rangle$. Then we could write down an expres-

[2]We can easily extend the standard definition of bisimulation (van Benthem 2010) for the case of $\mathcal{L}$-models. In this example, dotted lines show the links between two bisimilar states in M_1 and M_2.

sion $\alpha(p)$ containing symbols from $\mathcal{L}$ such that for every model M, $M, v \models \alpha(p)$ iff $M, v' \models p$ with $v' = O\lceil_v(\sigma\lceil_v)$. Then assume $s(u) = \{u, w, x\}$, $s(w) = \{w\}$, $s(x) = \{x\}$, it holds that $M_2, u \models \alpha(p)$. Then by bisimulation, it follows that $M_1, u \models \alpha(p)$. Then we have $M_1, x \models p$. Contradicts with the fact that p is only true at w.

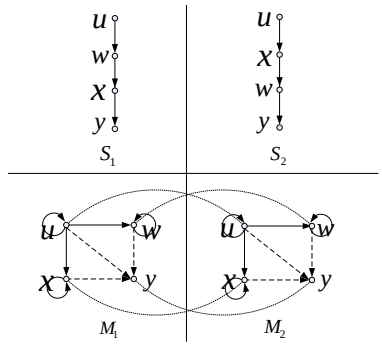


Figure 3: Undefinable modality

Accordingly, the frame structure for $\mathcal{L}(\mathring{\sigma}^s, \mathring{\sigma}^s_{-i})$ is obtained by incorporating the two accessibility relations into the frames for $\mathcal{L}$. That is:

$vR_{\mathring{\sigma}^s}v'$ iff $v' = O\lceil_v(\sigma\lceil_v)$

$vR_{\mathring{\sigma}^s_{-i}}v'$ iff $v' \in O\lceil_v(\sigma_{-i}\lceil_v)$

The truth conditions are:

$M, v \models \langle\mathring{\sigma}^s\rangle\varphi$ iff $M, u \models \varphi$ for some $u \in V$ with $vR_{\mathring{\sigma}^s}u$.

$M, v \models \langle\mathring{\sigma}_{-i}{}^s\rangle\varphi$ iff $M, u \models \varphi$ for some $u \in V$ with $vR_{\mathring{\sigma}^s_{-i}}u$.

Example 6. In the context of Example 4. We have the following:

• $M, v_0 \models \langle\mathring{\sigma}^s_{-2}\rangle p$. I.e., there is a terminal node (actually, v_4) in $S\lceil_{v_0}$ that satisfies p, and that can be reached from v_0 when player 2 being free while other players adhere to σ.

• $M, v_0 \models \langle\mathring{\sigma}^s\rangle\neg p$. I.e., p does not hold at the terminal node (actually, at v_5) of $S\lceil_{v_0}$ that can be arrived at when all players adopt strategy profile σ.

K axiom naturally holds for $\mathring{\sigma}^s$, $\mathring{\sigma}^s_{-i}$. Other valid principles concerning the two modalities are listed in Table 2:

Y shows the *visibility* of all the nodes that can be reached from the current node v in sight-filtrated game $S\lceil_v$. D and I are the same as that for $\mathring{\sigma}$ and $\mathring{\sigma}_{-i}$.

N	Modality	Schema	Property
D	$[\mathring{\sigma}^s]$	$[.]\varphi \leftrightarrow \langle . \rangle \varphi$	determinism
I	$([\mathring{\sigma}^s], [\mathring{\sigma}^s_{-i}])$	$[\mathring{\sigma}_{-i}]\varphi \rightarrow [\mathring{\sigma}]\varphi$	inclusiveness
Y	$([\lessdot], [\mathring{\sigma}^s])$ $([\lessdot], [\mathring{\sigma}^s_{-i}])$	$[\lessdot]\varphi \rightarrow [\mathring{\sigma}^s]\varphi$ $[\lessdot]\varphi \rightarrow [\mathring{\sigma}^s_{-i}]\varphi$	visibility

Table 2: Valid principles of $L(\mathring{\sigma}^s, \mathring{\sigma}^s_{-i})$

4.3 $\mathcal{L}(\mathring{\sigma}^s, \mathring{\sigma}^s_{-i}, \sigma_i)$: the complete language

The existing modalities can not justify the intermediate nodes before reaching a leaf node. We add another modality $\langle \sigma_i \rangle$ to $\mathcal{L}(\mathring{\sigma}^s, \mathring{\sigma}^s_{-i})$. R_{σ_i} determines the node that is reachable from the current node v after player i carrying on strategy σ_i. Formula $\langle \sigma_i \rangle \varphi$ indicates that "φ holds in the successor node determined by strategy σ_i".

Thus, the frames for $\mathcal{L}(\mathring{\sigma}^s, \mathring{\sigma}^s_{-i}, \sigma_i)$ are defined by extending the frames for $\mathcal{L}(\mathring{\sigma}^s, \mathring{\sigma}^s_{-i})$ with R_{σ_i}. That is,

$vR_{\sigma_i}v'$ iff $v' = \sigma_i(v)$

And the truth condition is:

$M, v \models \langle \sigma_i \rangle \varphi$ iff $M, u \models \varphi$ for some $u \in V$ with $vR_{\sigma_i}u$.

Example 7. Consider the case in Example 4. We have the following:

• $M, v_0 \models \langle \lessdot \rangle \langle \sigma'_2 \rangle p$. I.e., there is a node (actually, v_4) that satisfies p, and that can be reached by following σ'_2 from a node (actually, v_2) that is within players sight at v_0.

• $M, v_0 \models \langle \sigma_2 \rangle \neg p \wedge \langle \mathring{\sigma}^s \rangle \neg p$. I.e., there is a node (exactly, v_5) falsifies p and can be reached from v_0 by following σ_2. Moreover, there is also a terminal node (actually, v_5) of $O\lceil_{v_0}(\sigma\lceil_{v_0})$ in $S\lceil_{v_0}$ that falsifies p.

With the approach in Example 5, we can also prove the reason of introducing $\langle \sigma_i \rangle$ as new operator by showing that σ_i is undefinable in $\mathcal{L}(\mathring{\sigma}^s, \mathring{\sigma}^s_{-i})$.

K axiom holds for $[\sigma_i]$.

For convenience, we use $\mathcal{LS}$ to represent the language $\mathcal{L}(\mathring{\sigma}^s, \mathring{\sigma}^s_{-i}, \sigma_i)$. Then logic LS is a set of formulas that contains all tautologies, $K, T, 4, D, I, M$, and Y introduced in the above three steps and that is closed under Modus Ponens and Necessitation for the modalities in LS.

5 Expressing properties of games with Short Sight

Language $\mathcal{LS}$ can be used to characterize the solution concepts of extensive games with short sight. To show this, we first define the subframes of F_S.

Given an extensive game with short sight S, and any non-terminal node v in S, we can obtain from a frame F_S a subframe $F_{S\lceil v}$, where $S\lceil v=(N\lceil_v, V\lceil_v, A\lceil_v, t\lceil_v, \Sigma_i\lceil_v, \succeq_i\lceil_v)$ is the sight-filtrated extensive game of S at v:

The tuple $(V\lceil_v, R_{\leq_i\lceil_v}, R_{\chi\lceil_v}, R_{\lhd\lceil_v}, R_{\chi^s\lceil_v}, R_{\sigma_i\lceil_v})$ is defined as a sight-filtrated subframe $F_{S\lceil v}$ of F_S, where $\chi=\{\mathring{\sigma}, \mathring{\sigma}_{-i}\}$.

For any two nodes u, u' in the game $S\lceil_v$, i.e., $u, u' \in V\lceil_v$, the accessibility relations in $F_{S\lceil v}$ are defined as follows (recall Definition 3.3 and Lemma 1), where $O\lceil_v(\sigma\lceil_v)(u)$ represents the terminal node that can be reached from u by adopting σ in the game $S\lceil_v$:

$uR_{\leq_i\lceil_v}u'$ *iff* $uR_{\leq_i}u'$.

$$uR_{\chi\lceil_v}u' \quad \textit{iff} \quad \begin{cases} u' = O\lceil_v(\sigma\lceil_v)(u), & \chi = \mathring{\sigma} \\ u' \in O\lceil_v(\sigma_{-i}\lceil_v)(u), & \chi = \mathring{\sigma}_{-i} \end{cases}$$

$uR_{\lhd\lceil_v}u'$ *iff* $u \lhd u'$.

$$uR_{\chi^s\lceil_v}u' \quad \textit{iff} \quad \begin{cases} (u,u') \in R_{\chi\lceil_v}, & \chi = \mathring{\sigma} \\ (u,u') \in R_{\chi\lceil_v}, & \chi = \mathring{\sigma}_{-i} \end{cases}$$

$uR_{\sigma_i\lceil_v}u'$ *iff* $(u,u') \in R_{\sigma_i}$.

A model $M_{S\lceil v}$ is a pair $(F_{S\lceil v}, \pi)$ where $F_{S\lceil v}$ is a sight-filtrated subframe for $\mathcal{LS}$ and π an assignment function $\pi : P \to 2^{V\lceil_v}$.

The truth conditions of formulas in $M_{S\lceil v}$ are :

$M_{S\lceil v}, u \models p$ iff $u \in \pi(p)$.
$M_{S\lceil v}, u \models \neg\varphi$ iff not $M_{S\lceil v}, u \models \varphi$.
$M_{S\lceil v}, u \models \varphi \wedge \psi$ iff $M_{S\lceil v}, u \models \varphi$ and $M_{S\lceil v}, u \models \psi$.
$M_{S\lceil v}, u \models \langle\alpha\rangle\varphi$ iff $M_{S\lceil v}, w \models \varphi$ for some w with $uR_\alpha w$.
(α represents any modal opertor in $\mathcal{LS}$)

Theorem 1. *Let S be an Egss given by $(N, V, A, t, \Sigma_i, \succeq_i, s)$. Then for any player i, any strategy profiles σ in S and any formulas φ of $\mathcal{LS}$:*

(*a*) *σ is a sight-compatible best response* (SCBR) *of S for i iff $\mathcal{F}_S \models [\mathring{\sigma}^s]\varphi \to [\mathring{\sigma}^s_{-i}]\langle\leq_i\rangle\varphi$.*

(*b*) *σ is a sight-compatible Nash equilibrium* (SCNE) *of S iff $F_S \models \bigwedge_{i\in N}([\mathring{\sigma}^s]\varphi \to [\mathring{\sigma}^s_{-i}]\langle\leq_i\rangle\varphi)$.*

(*c*) *σ is a subgame perfect equilibrium* (SPE) *of $S\lceil v$ iff for any $u \in V\lceil_v \backslash Z\lceil_v$, $F_{S\lceil_v}, u \models \bigwedge_{i\in N}([\mathring{\sigma}]\varphi \to [\mathring{\sigma}_{-i}]\langle\leq_i\rangle\varphi)$.*

(*d*) *A strategy profile* σ *is a sight-compatible* SPE *of* S *iff for all* $v \in V \backslash Z$, $F_{S\lceil_v}, v \models [\prec](\bigwedge_{i\in N}([\mathring{\sigma}]\varphi \rightarrow [\mathring{\sigma}_{-i}]\langle\leq_i\rangle\varphi))$.

Proof. (*a*) For the direction ($\Rightarrow$) , assume $[\mathring{\sigma}^s]\varphi \rightarrow [\mathring{\sigma}^s_{-i}]\langle\leq_i\rangle\varphi$ is invalid in F_S. Then there exist nodes v, v', v'', s.t. $vR_{\mathring{\sigma}^s}v'$, $vR_{\mathring{\sigma}^s_{-i}}v''$ and $(v'', v') \notin R_{\leq_i}$. By the Definition of F_S, we have: $v' \in O\lceil_v(\sigma\lceil_v)$, $v'' \in O\lceil_v(\sigma_{-i}\lceil_v)$ and $v'' >_i v'$. By Definition 3.4, we can get that $\sigma\lceil_v$ is not a SCBR of $S\lceil_v$ for i.

For ($\Leftarrow$), assume σ is not a sight-compatible best response (SCBR) of S for i. Then there exist a nonterminal node v, a strategy $\sigma_i^*\lceil_v$ available to i, such that $O\lceil_v(\sigma_i^*\lceil_v, \sigma^*_{-i}\lceil_v) \geq_i\lceil_v O\lceil_v(\sigma_i\lceil_v, \sigma^*_{-i}\lceil_v)$. This is to say that there exist nodes v', v'', s.t. $vR_{\mathring{\sigma}^s}v'$, $vR_{\mathring{\sigma}^s_{-i}}v''$ and $v'' \geq_i v'$, I.e., $\neg(v''pref_i\ v')$. Then it follows that $F_S \not\models [\mathring{\sigma}^s]\varphi \rightarrow [\mathring{\sigma}^s_{-i}]\langle\leq_i\rangle\varphi$.

(*b*) Proof of (*b*) would be trivial given (*a*) (by definition 3.4).

(*c*) For ($\Rightarrow$), assume $\bigwedge_{i\in N}([\mathring{\sigma}]\varphi\rightarrow[\mathring{\sigma}_{-i}]\langle\leq_i\rangle\varphi)$ is invalid at some state u in $F_{S\lceil_v}$. Then for some player i, $[\mathring{\sigma}]\varphi \rightarrow [\mathring{\sigma}_{-i}]\langle\leq_i\rangle\varphi$ is invalid at u in $F_{S\lceil_v}$. Consequently, $[\mathring{\sigma}]\varphi \wedge \langle\mathring{\sigma}_{-i}\rangle[\leq_i]\neg\varphi$ is valid at u. It follows that σ is not a subgame perfect equilibrium of $S\lceil_v$ (by Definition 2.3). ($\Leftarrow$) can be proved similarly.

(*d*) For ($\Rightarrow$), assume $[\prec](\bigwedge_{i\in N}([\mathring{\sigma}]\varphi \rightarrow [\mathring{\sigma}_{-i}]\langle\leq_i\rangle\varphi))$ is invalid at the starting point v of $F_{S\lceil_v}$. Then there exists a node v' such that (1) $v' \in s(v)$ and (2) $F_{S\lceil_v}, v' \not\models \bigwedge_{i\in N}([\mathring{\sigma}]\varphi \rightarrow [\mathring{\sigma}_{-i}]\langle\leq_i\rangle\varphi))$. By (2) and (*c*), we have that σ is not a SPE of $S\lceil_v$. Then by (1) it follows that σ is not a sight-compatible subgame perfect equilibrium of S. For ($\Leftarrow$), assume σ is not a sight-compatible subgame perfect equilibrium of S. Then there exists a state $v \in V$ such that for any SPE σ^* of $S\lceil_v$, $\sigma \neq \sigma^*$. Then σ is not a SPE of $S\lceil_v$. By (c), we have: $\exists u \in V\lceil_v$ such that $F_{S\lceil_v}, u \models \neg \bigwedge_{i\in N}([\mathring{\sigma}]\varphi \rightarrow [\mathring{\sigma}_{-i}]\langle\leq_i\rangle\varphi)$. It follows that $F_{S\lceil_v}, v \models \langle\prec\rangle\neg \bigwedge_{i\in N}([\mathring{\sigma}]\varphi \rightarrow [\mathring{\sigma}_{-i}]\langle\leq_i\rangle\varphi)$. By Dual, $F_{S\lceil_v}, v \not\models [\prec](\bigwedge_{i\in N}([\mathring{\sigma}]\varphi \rightarrow [\mathring{\sigma}_{-i}]\langle\leq_i\rangle\varphi))$.

□

By now, we hope we have illustrated the expressive power of the new language by formally characterizing the solution concepts for Egss.

6 Conclusions

We proposed a new logic for strategic reasoning about games with short sight. We then presented an axiomatization for the logic. Finally, we showed that the logic can

formally characterize the solution concepts, e.g. sight-compatible best response, in Egss.

Since our focus has been to capture the strategies and solution concepts in games, our logic took strategy profiles as primitive modality in this paper. Yet in case the internal structure of strategies is required, a logic considering players' moves (rather than strategy profiles consisting of a sequence of moves) as atomic actions (like PDL) may be appropriate. We would like to incorporate such a perspective and extend our current results. Halpern and Rêgo (2006) studied games with awareness, and Grossi and Turrini (2012) then proved a representation of games with awareness by games with short sight. Our language can be easily extended with *awareness* modality, then we can provide similar representation results and study other interesting phenomena in games. Finally, we would like to look into the model checking problem, especially, comparing the complexity of the problem in the standard game model and that in games with short sight.

Acknowledgements This work is partially supported by 973 Program (No. 2010CB328103). We would like to thank Johan van Benthem for his helpful comments.

References

J. van Benthem. Extensive games as process models. *Journal of Logic, Language and Information*, 11(3):289–313, June 2002.

J. van Benthem. *Modal Logic for Open Minds*. Center for the Study of Language and Information Lecture Notes, Stanford University, Feb. 2010.

P. Blackburn, M. de Rijke, and Y. Venema. *Modal logic*. Cambridge University Press, 2001.

G. Bonanno, M. Magill, and K. van Gaasback. Branching time logic, perfect information games and backward induction. Working Papers 9813, University of California, Davis, Department of Economics, 2003.

D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1991.

D. Grossi and P. Turrini. Short sight in extensive games. In *AAMAS*, pages 805–812, 2012.

J. Y. Halpern and L. C. Rêgo. Extensive games with possibly unaware players. In *AAMAS*, pages 744–751, 2006.

P. Harrenstein, W. van der Hoek, J.-J. C. Meyer, and C. Witteveen. A modal characterization of nash equilibrium. *Fundamenta Informaticae*, 57(2-4):281–321, 2003.

E. Lorini and F. Moisan. An epistemic logic of extensive games. *Electronic Notes in Theoretical Computer Science*, 278:245–260, 2011.

J. F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1):48–49, 1950.

S. van Otterloo, W. Van der Hoek, and M. Wooldridge. Preferences in game logics. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '04, pages 152–159, Washington, DC, USA, 2004. IEEE Computer Society.

R. Parikh. The logic of games and its applications. In *Annals of Discrete Mathematics*, pages 111–140. Elsevier, 1985.

R. Ramanujam and S. E. Simon. Dynamic logic on games with structured strategies. In G. Brewka and J. Lang, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference, KR 2008, Sydney, Australia, September 16-19, 2008*, pages 49–58. AAAI Press, 2008.

Probabilistic Semantics for Natural Language

Jan van Eijck and Shalom Lappin

CWI and ILLC Amsterdam
King's College London
jve@cwi.nl, shalom.lappin@kcl.ac.uk

Abstract

Probabilistic and stochastic methods have been fruitfully applied to a wide variety of problems in grammar induction, natural language processing, and cognitive modeling. In this paper we explore the possibility of developing a class of combinatorial semantic representations for natural languages that compute the semantic value of a (declarative) sentence as a probability value which expresses the likelihood of competent speakers of the language accepting the sentence as true in a given model, relative to a specification of the world. Such an approach to semantic representation treats the pervasive gradience of semantic properties as intrinsic to speakers' linguistic knowledge, rather than the result of the interference of performance factors in processing and interpretation. In order for this research program to succeed, it must solve three central problems. First, it needs to formulate a type system that computes the probability value of a sentence from the semantic values of its syntactic constituents. Second, it must incorporate a viable probabilistic logic into the representation of semantic knowledge in order to model meaning entailment. Finally, it must show how the specified class of semantic representations can be efficiently learned. We construct a probabilistic semantic fragment and consider how the approach that the fragment instantiates addresses each of these three issues.

1 Introduction

A formal semantic theory recursively defines the denotation of an expression in terms of the denotations of its syntactic constituents. It computes the semantic values of a sentence as a function of the values of its syntactic constituents. Within such a theory the meaning of an expression is identified with a function from indices (the expressions themselves, worlds, situations, times, etc.), to denotations in a model. The meaning of a sentence is a function from indices to truth-values.

Formal semantic theories model both lexical and phrasal meaning through categorical rules and algebraic systems that cannot accommodate gradience effects. This approach is common to theories which sustain compositionality and those with employ underspecified representations.[1] It effectively invokes the same strong version of the competence-performance distinction that categorical models of syntax assume. This view of linguistic knowledge has dominated linguistic theory for the past fifty years.

Gradient effects in representation are ubiquitous throughout linguistic and other cognitive domains. Appeal to performance factors to explain gradience has no explanatory content unless it is supported by a precise account of how the interaction of competence and performance generates these effects in each case. By contrast, gradience is intrinsic to the formal models that information theoretic methods use to represent events and processes.

Bach (1986) identifies two theses on the character of natural language.

(a) **Chomsky's thesis:** Natural languages can be described as formal systems.

(b) **Montague's thesis:** Natural languages can be described as *interpreted* formal systems.

Recent work in computational linguistics and cognitive modeling suggests a third proposal.

(c) **The Harris-Jelinek thesis:** Natural languages can be described as information theoretic systems, using stochastic models that express the distributional properties of their elements.

The Harris-Jelinek thesis implies the *The Language Model Hypothesis* (LMH) for syntax, which holds that grammatical knowledge is represented as a stochastic language model.[2] On the LMH, a speaker acquires a probability distribution

[1] See, *inter alia*, Reyle 1993, Bos 1995, Blackburn and Bos 2005, Copestake et al. 2006, Koller et al. 2008, Fox and Lappin 2010 for discussions of underspecified semantics.

[2] See (Clark and Lappin 2011) for a discussions of the merits and problems of the LMH. An obvious difficulty with the LMH is that in the primary linguistic data for language acquisition short, ill formed sentences

$D : \Sigma^* \rightarrow [0, 1]$, over the strings $s \in \Sigma^*$, where Σ is a set of words (morphemes, etc.) of the language, and $\sum p_D(s) = 1$. This distribution is generated by a probabilistic automaton or a probabilistic grammar, which assigns a structure to a string with a probability that is the product of the rules applied in the derivation of that string. The probability of the string itself is the sum of the parses that the grammar generates for it. This probability represents the likelihood of a sentence's occurrence in a corpus.[3]

Representing linguistic knowledge stochastically does not eliminate the competence – performance distinction. It is still necessary to distinguish between a probabilistic grammar or automaton that generates a language model, and the parsing algorithm that implements it. However, a probabilistic characterization of linguistic knowledge does alter the nature of this distinction. The gradience of linguistic judgements and the defeasibility of grammatical constraints are now intrinsic to linguistic competence, rather than distorting factors contributed by performance mechanisms.

Lexically mediated relations like synonymy, antinomy, polysemy, and hyponymy are notoriously prone to clustering and overlap effects. They hold for pairs of expressions over a continuum of degrees [0,1], rather than Boolean values {1,0}. Moreover, the denotations of major semantic types, like the predicates corresponding to CNs, AdjPs, and VPs, can rarely, if ever, be identified as sets with determinate membership. The case for abandoning the categorical view of competence and adopting a probabilistic model is at least as strong in semantics as it is in syntax (as well as in other parts of the grammar)

A probabilistic semantics needs to express the probability of a different property than occurrence in a corpus. Knowing the meaning of a declarative sentence consists largely in being able to estimate the probability that competent speakers of the language would take it to be true across different states of the world (different worlds). This view is a probabilistic extension of a classical truth-conditional view of meaning. It can be extended to non-declarative sentences by formulating fulfillment conditions for

consisting of high frequency lexical items may have higher probability than longer, complex, well formed sentences containing low frequency words. A possible solution to this problem is to model grammatical acceptability in stochastic terms by imposing a lower bound on the probability of an acceptable string s that is dependent on properties of s, like its length, and features of the distribution for Σ^*. So, for example, a three word string like *You is here* is likely to have lower probability than the average probability of three word strings consisting of the word class sequence $\langle N, V, ADV \rangle$. By contrast, the string *Trading in complex instruments like mortgage backed derivatives and credit default swaps remains opaque and inexplicably under-regulated, which continues to be a major cause of instability in the financial markets* can be expected to have at least the average probability of strings of the same length and word class sequence. This approach to modeling acceptability uses the idea that one's expectation for the likelihood of occurrence of a string in a corpus depends, in part, on its properties and those of the distribution for its string set. It is derived from the stochastic model of indirect negative evidence that Clark and Lappin (2011) propose.

[3] See Manning and Schütze 1999, Collins 2003, Jurafsky and Martin 2009, Chelba 2010, Clark and Lappin 2010, Clark 2010 for discussions of statistical parsing and probabilistic grammars.

them and identifying the meaning of a sentence with the function that determines the probability that speakers of the language construe it as fulfilled (a question answered, an imperative obeyed, a request satisfied, etc.) in any given state of affairs.[4]

As in the case of parsing, adopting a probabilistic view of semantic knowledge does not entail the eradication of the distinction between competence and performance. We still need to separate the semantic representation that generates a probability distribution for sentences in relation to states of affairs from the application of this representation in interpreting sentences. But like probabilistic grammars, these models incorporate gradience as an intrinsic feature of the objects that they characterize.

In this paper we argue that by replacing truth-conditions with probability conditions we can capture at least some of the pervasive gradience effects in semantic judgements. This allows us to reduce a number of important varieties of vagueness to the sort of uncertainty of belief (in this case, semantic belief) that probabilistic theories are designed to model. We are also able to account for several important kinds of semantic learning as a process of updating a learner's probability distributions over the worlds (which encode possible knowledge states) in which he/she evaluates the predicates whose meanings he/she is acquiring. This approach is consistent with the Harris-Jelinek thesis in that it represents semantic knowledge as a probability distribution over worlds that is generated by a probabilistic model for interpreting expressions of a language.

In Section 2 we present definitions of a model, a basic type theory, and a recursive definition of an interpretation function for a fragment of a formal representation language. In Section 3 we propose the outline of an account of semantic learning in which learners acquire the interpretation of new predicates, treated as probabilistic classifiers, in their language. We compare our approach to distributional treatments of meaning, particularly vector space models (VSMs), in Section 4. VSMs have emerged as highly efficient procedures for learning semantic relations among lexical items in corpora. Recent work has focussed on extending these methods to sentences. We discuss the complex connections among probability, gradience, and semantic vagueness in Section 5. Finally, in Section 6 we draw conclusions from our proposals and indicate directions for future work.

2 Probabilistic models for a semantic fragment

Classical probabilistic logic (Carnap 1950, Nilsson 1986, Fagin and Halpern 1991, Paris 2010) models uncertainty in our knowledge of the facts about the world. Probability distributions are specified over a set of possible states of the world (possible

[4]Lappin (1982) offers an early proposal for characterizing truth conditions as an instance of fulfillment conditions.

worlds), and the probabilities for the elements of this set sum to 1. A proposition φ is assigned truth-values across worlds, and φ's probability is computed as $\sum_{w \in W} p(w)$ for $\{w : |\varphi|^w = t\}$.

In characterizing meaning probabilistically, we can talk of uncertainty about the truth-value of a sentence, given some probability distribution over possible states of affairs. The probability of a sentence expresses the likelihood that (semantically) competent speakers of the language assign to the truth of the sentence, given the state of their knowledge about the world. We can then represent the meaning of a sentence as a function that maps intensions to functions from knowledge states to probabilities (probability conditions). The semantic value of a sentence S is of type $I \to K \to [0, 1]$, where I is the set of intensions, K is the set of knowledge representations, and $[0, 1]$ is the set of reals p with $0 \leq p \leq 1$.

Let a propositional language over a set of basic predications be given, as follows:

$$\begin{array}{rcl} t & ::= & x \mid a_1 \mid a_2 \mid \cdots \mid a_m \\ Q & ::= & Q_1 \mid Q_2 \mid \cdots \mid Q_n \\ \varphi & ::= & Qt \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi. \end{array}$$

Here we assume a single variable x, a finite number of proper names $a_1, a_2, \ldots, a_m$ and a finite number of basic unary predicates $Q_1, Q_2, \ldots, Q_n$.

Any φ that contains occurrences of x is called a *predication*. Use $\varphi(x)$ for predications, and $\varphi(a/x)$ for the result of replacing x by a everywhere in a predication.

Call this language L_n^m. If we extend L_n^m with one name a_{m+1}, the new language is called L_n^{m+1}. If we extend L_n^m with one new predicate Q_{n+1}, the new language is called L_{n+1}^m

For convenience, we identify names and objects, so we assume a domain $D_m = \{a_1, a_2, \ldots, a_m\}$. The type of a (restricted) world w is given by $w : \{Q_1, \ldots, Q_n\} \to \mathcal{P}(D_m)$. $w(Q_i)$ is the interpretation of Q_i in w.

A *probabilistic model* M is a tuple $\langle D, W, P\rangle$ with D a domain, W a set of worlds for that domain (predicate interpretations in that domain), and P a probability function over W, i.e., for all $w \in W$, $p(w) \in [0, 1]$, and $\sum_{w \in W} p(w) = 1$.

An interpretation of L_n^m in an L_n^m-model $M = \langle D, W, P\rangle$ is given in terms of the standard notion $w \models \varphi$, as follows:

$$[\![\varphi]\!]^M \quad := \quad \sum\{P(w) \mid w \in W, w \models \varphi\}.$$

It is straightforward to verify that this yields $[\![\neg\varphi]\!]^M = 1 - [\![\varphi]\!]^M$. Also, if $\varphi \models \neg\psi$, i.e., if $W_\varphi \cap W_\psi = \emptyset$, then $[\![\varphi \vee \psi]\!]^M = \sum_{w \in W_{\varphi\vee\psi}} P(w) = \sum_{w \in W_\varphi} P(w) + \sum_{w \in W_\psi} P(w) = [\![\varphi]\!]^M + [\![\psi]\!]^M$, as required by the axioms of Kolmogorov (1950)'s probability calculus.

2.1 A toy fragment

Basic types are e (entities), s (worlds), t (truth values), d (domains) and $[0, 1]$ (the space of probabilities). Abbreviate $d \to s \to t$ as i (intensions). Types for S, N, VP, NP, DET are lifted to the level of intensions, by substituting i for t in all types. This gives, e.g., $DET = (e \to i) \to (e \to i) \to i$.

The lifting rules for the interpretation functions are completely straightforward:

$$I(Some) = \lambda p \lambda q \lambda dom \lambda w.some(\lambda x.p\ x\ dom\ w)(\lambda y.q\ y\ dom\ w).$$

Here *some* is the familiar constant function for existential quantification, of type $(e \to t) \to (e \to t) \to t$.

This type system gives sentences an interpretation of type i, i.e., $d \to s \to t$. Such intensions can be mapped to probabilities by means of a function *prob* of type $i \to m \to [0, 1]$, where m is the type of models with their domains, i.e., objects of the shape $\langle D, W, P\rangle$.

The function *prob* on sentences f and models $M = \langle D, W, P\rangle$ is given by:

$$prob\ f\ \langle D, W, P\rangle = \sum\{P(w) \mid w \in W, f\ D\ w\}.$$

This function assigns to every sentence of the fragment a probability, on the basis of the prior probabilities encoded by $\langle D, W, P\rangle$.

2.2 Semantic priors

The probabilities in a model M are the prior of a target semantic representation. We can take this prior to encode the knowledge representation that competent speakers converge upon as they acquire the meanings of the predicates of their language. Learners start out with different priors (probability distributions over models) than mature speakers, and update them through semantic learning. The prior that a learner brings to the learning task constitutes his/her initial assumptions about the state of the world, and, in a sense, it is the basis for semantic learning

Kemp et al. (2007) propose a hierarchical Bayesian learning framework in which observational classifiers and the learning priors that express expectations concerning the distribution of observations categorized by these classifiers can be acquired simultaneously from the same data. The priors are themselves derived from more general higher-order priors.

3 Semantic learning

Classical semantic theories characterize a class of representations for the set of meanings of expressions in natural language. However, it is unclear how these representations could be learned from the primary linguistic data (PLD) of language acquisition. The problem of developing a plausible account of efficient learnability of appropriate target representations is as important for semantics as it is for other types of linguistic knowledge. Most work in formal learning for natural languages has focussed on syntax (grammar induction), morphology, and phonology.

3.1 Simple cases of learning

Example 1

Assume there are just two predicates Q_1 and Q_2, and two objects a, b. Complete ignorance about how the predicates are applied is represented by a model with 16 worlds, because for each object x and each predicate Q there are two cases: Q applies to x or not. If the likelihood of each of the cases is completely unknown, each of these worlds has probability $\frac{1}{16}$.

Example 2

Suppose again there are two objects a, b and two predicates Q_1, Q_2. Assume that it is known that a has Q_1, and the probability that b has Q_1 is taken to be $\frac{2}{3}$. Suppose it is known that no object has Q_2. Then $W = \{w_1, w_2\}$ with $w_1(Q_1) = \{a, b\}$, $w_2(Q_1) = \{a\}$, $w_1(Q_2) = \emptyset$, $w_2(Q_2) = \emptyset$. P is given by $P(w_1) = \frac{2}{3}$, $P(w_2) = \frac{1}{3}$. In this example $\neg Q_1(b)$ is true in w_2 and not in w_1. Therefore $[\![\neg Q_1(b)]\!] = \frac{1}{3}$.

Learning new definable predicates

Learning a new semantic concept Q_{n+1} is learning how (or to what extent) predicate Q_{n+1} applies to the objects one knows about. The simplest way to model such a learning event is as a pair $\langle Q_{n+1}, \varphi(x)\rangle$ where $\varphi(x)$ is an L_n^m predication. The effect of the learning event could then be modeled in a way that is very similar to the manner in which factual change is modeled in an epistemic update logic.

The result of updating a model $M = \langle D, W, P\rangle$ with concept learning event $\langle Q_{n+1}, \varphi(x)\rangle$ is the model that is like M except for the fact that the interpretation in each world of Q_{n+1} is given by

$$w(Q_{n+1}) := \{a \mid a \in D_m, w \models \varphi(a/x)\}$$

Note that the probability function P of the model does not change in this case.

Let's return to example 1. This is the model where there are two objects and two predicates, and nothing is known about the properties of the objects. Take the learning event $(Q_3, Q_1x \wedge \neg Q_2x)$. This defines Q_3 as the difference of Q_1 and Q_2. The resulting model will again have 16 worlds, and in each world w_i, $w_i(Q_3)$ is given by $w_i(Q_1) \cap (D - w_i(Q_2))$. Again, the probabilities of the worlds remain unchanged.

3.2 Adjusting the meaning of a predicate

To allow adjustment of the meaning of a classifier by means of a learning event, we can use probabilistic updating (following van Benthem et al. (2009)). A classifier learning event now is a tuple $\langle Q, \varphi, \psi(x), q\rangle$ where φ is a sentence, $\psi(x)$ is a predication, and q is a probability. φ expresses the observational circumstances of the revision. q expresses the observational certainty of the new information.

The result of updating $M = \langle D, W, P\rangle$ with $\langle Q, \varphi, \psi(x), q\rangle$ is a new model $M = \langle D, W', P'\rangle$. W' is given by changing the interpretation of Q in members w of W_φ to $\{a \mid w \models \psi(a/x)\}$, while leaving the interpretation of Q in members of $W_{\neg\varphi}$ unchanged.

P' is given by $P'(w) = \frac{P(w)\times q}{X}$ for members of W_φ, and by $P'(w) = \frac{P(w)\times(1-q)}{X}$ for members of $W_{\neg\varphi}$. $\frac{1}{X}$ (the normalization factor) is given by

$$X = \sum_{w\in W_\varphi} P(w) \times q + \sum_{w\in W_{\neg\varphi}} P(w) \times (1 - q).$$

Learning classifiers by example

Consider again the example with the two objects and the two properties, where new information concerning the application of the predicates to objects in the domain is acquired. A learning event for this could be $\langle Q_2, \neg Q_1b, Q_1x \vee Q_2x, \frac{2}{3}\rangle$. Then the resulting model has again 2 worlds, but now the probability of w_2 has gone up from $\frac{1}{3}$ to $\frac{\frac{2}{3}\times\frac{1}{3}}{\frac{1}{3}\times\frac{2}{3}+\frac{2}{3}\times\frac{1}{3}} = \frac{1}{2}$. The probability of w_1 has gone down from $\frac{2}{3}$ to $\frac{\frac{1}{3}\times\frac{2}{3}}{\frac{4}{9}} = \frac{1}{2}$.

You are given something of which you are told that it is called a "rose", and you observe that it is thorny, red and a flower. A learning example is an encounter with a new object a_{m+1}. Suppose you learn that predicate Q applies to a_{m+1}. The properties you observe of a_{m+1} are given by $\theta(a_{m+1})$, where $\theta(a_{m+1})$ is a conjunction of $\pm Q_i(a_{m+1})$ for all known predicates. The update event is $\langle a_{m+1}, Q, \theta(a_{m+1})\rangle$. You learn that a_{m+1} is called a Q, and you observe that a_{m+1} satisfies the properties $\theta(a_{m+1})$.

Updating a model $M = \langle D, W, P\rangle$ for L_n^m with this event creates a new model $M' = \langle D \cup \{a_{m+1}\}, W', P\rangle$ for L_n^{m+1}. The new model has domain $\{a_1, \ldots, a_{m+1}\}$. W' is given

by assigning, in each w, to a_{m+1} the properties specified by $\theta(a_{m+1})$. The interpretation of Q is given by setting $w(Q) = \{a \mid w \models \theta(a/a_{m+1})\}$. This resets the interpretation Q on the basis of the new observation. The probability distribution remains unchanged.

We can refine this account of learning to accommodate cases where an observation is less precise. Let the learning event be:

$$\langle a_{m+1}, Q, \{(\theta_1(a_{m+1}), q_1), \ldots, (\theta_k(a_{m+1}), q_k)\}\rangle.$$

Here q_i gives the observational probability that the new object satisfies θ_i. The probabilities should satisfy $\sum_{i=1}^{k} q_i = 1$. The update can be defined so that the probability of the new predicate applying to the old objects will be recomputed.

3.3 Semantic knowledge and knowledge of the world

Our specification of the class of probabilistic models and our treatment of learning raise the question of how to distinguish between semantic knowledge and knowledge of the world. It might seem that the distinction disappears entirely in our framework, and we are simply modeling epistemic update. In fact this is not the case. In a probabilistic account of epistemic update one seeks to express the effect of new information about the actual world on a belief agent's probability distribution over possible worlds. In our system of semantic representation we specify the meaning of a sentence as the likelihood that competent speakers of the language will assess it as true, given the distribution over worlds that sustains the interpretation of the expressions of their language. We are, then, seeking to model the probability that speakers assign to sentences across possible states of affairs, where these probability conditions are derived from the prior that speakers specify for worlds as a condition for sharing the meanings of their predicates. Semantic learning is a process of converging on the target model that generates this distribution by forming hypotheses on the intensions of predicates (the classifiers that they encode) on the basis of the PLD.

The notion of a semantic prior in terms of which the probability value of a sentence is computed allows us to identify semantic knowledge as distinct from general epistemic commitment. It is, however, the case that the distinction between semantic and extra-linguistic knowledge is not absolute. In learning a predicate one is acquiring a classifier that sorts objects on the basis of their properties. One could not apply such a classifier without recognizing these properties and making predictions concerning the likelihood that unobserved objects with similar properties satisfy (fail to satisfy) the classifier. It seems reasonable to assume that learners starting out with a semantic prior that is radically divergent from the target representation in most respects may find it difficult or impossible to acquire this representation from the PLD. If this does, in fact,

	context 1	context 2	context 3	context 4
financial	0	6	4	8
market	1	0	15	9
share	5	0	0	4
economic	0	1	26	12
chip	7	8	0	0
distributed	11	15	0	0
sequential	10	31	0	1
algorithm	14	22	2	1

Figure 1: Word Type-Context Matrix

turn out to be the case, then we can conclude that semantic learning depends on a core of shared beliefs about the nature of the world.

4 Distributional treatments of meaning

4.1 Lexical Vector Space Models

Vector Space Models (VSMs, Turney and Pantel 2010) offer a fine-grained distributional method for identifying a range of semantic relations among words and phrases. They are constructed from matrices in which words are listed vertically on the left, and the environments in which they appear are given horizontally along the top. These environments specify the dimensions of the model, corresponding to words, phrases, documents, units of discourse, or any other objects for tracking the occurrence of words. They can also include data structures encoding extra-linguistic elements, like visual scenes and events.

The integers in the cells of the matrix give the frequency of the word in an environment. A vector for a word is the row of values across the dimension columns of the matrix. Figure 1 gives a schematic example of such a word-context matrix, with made up vector values. In this matrix the vectors for *chip* and *algorithm* are [7 8 0 0] and [14 22 2 1], respectively.

A pair of vectors from a matrix can be projected as lines from a common point on a plane. The smaller the angle between the lines, the greater the similarity of the terms, as measured by their co-occurrence across the dimensions of the matrix. Computing the *cosine* of this angle is a convenient way of measuring the angles between vector pairs. If $\vec{x} = \langle x_1, x_2, ..., x_n \rangle$ and $\vec{y} = \langle y_1, y_2, ..., y_n \rangle$ are two vectors, then:

$$cos(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}}.$$

The cosine of $\vec{x}$ and $\vec{y}$ is their internal product, formed by summing the products of the corresponding elements of the two vectors, and normalizing the result relative to the lengths of the vectors. In computing $cos(\vec{x}, \vec{y})$ it may be desirable to apply a smoothing function to the raw frequency counts in each vector to compensate for sparse data, or to filter out the effects of high frequency terms. A higher value for $cos(\vec{x}, \vec{y})$ correlates with greater semantic relatedness of the terms associated with the $\vec{x}$ and $\vec{y}$ vectors.

VSMs provide highly successful methods for identifying a variety of lexical semantic relations, including synonymy, antinomy, polysemy, and hypernym classes. They also perform very well in unsupervised sense disambiguation tasks. VSMs offer a distributional view of lexical semantic learning. On this approach speakers acquire lexical meaning by estimating the environments (linguistic and non-linguistic) in which the words of their language appear.

4.2 Compositional VSMs

Lexical VSMs measure semantic distances and relations among words independently of syntactic structure. They apply a "bag of words" approach to meaning. Recent work has sought both to integrate syntactic information into the dimensions of the vector matrices (Padó and Lapata 2007), and to extend VSM semantic spaces to the compositional meanings of sentences. Mitchell and Lapata (2008) compare additive and multiplicative models for computing the vectors of complex syntactic constituents, and they demonstrate better results (as measured by human annotator judgements) with the latter for sentential semantic similarity tasks. These models use simple functions for combining constituent vectors, and they do not represent the dependence of composite vectors on syntactic structure.

Coecke et al. (2010), Grefenstette et al. (2011) propose a procedure for computing vector values for sentences that specifies a correspondence between the vectors and the syntactic structures of their constituents. This procedure relies upon a category theoretic representation of the types of a pregroup grammar (PGG, Lambek 2008a;b), which builds up complex syntactic categories through direction-marked function application in a manner similar to a basic categorial grammar. All sentences receive vectors in the same vector space, and so they can be compared for semantic similarity using measures like cosine.

A PGG compositional VSM (CVSM) determines the values of a complex syntactic structure through a function that computes the tensor product of the vectors of its constituents, while encoding the correspondence between their grammatical types and their semantic vectors. For two (finite) vector spaces A, B, their tensor product $A \otimes B$ is constructed from the Cartesian product of the vectors in A and B. For any two vectors $v \in A$, $w \in B$, $v \otimes w$ is the vector consisting of all possible products $v_{i \in v} \times w_{j \in w}$. Smolensky (1990) uses tensor products of vector spaces to construct representations of complex structures (strings and trees) from the distributed variables and values of the units in a connectionist network.

PGGs are modeled as *compact closed categories*. A sentence vector is computed by a linear map f on the tensor product for the vectors of its main constituents, where f stores the type categorial structure of the string determined by its PGG representation. The vector for a sentence headed by a transitive verb, for example, is computed according to the equation

$$\overrightarrow{subj\ V_{tr}\ obj} = f(\overrightarrow{subj} \otimes \overrightarrow{V_{tr}} \otimes \overrightarrow{obj}).$$

The vector of a transitive verb V_{tr} could be taken to be an element of the tensor product of the vector spaces for the two noun bases corresponding to its possible subject and object arguments $\overrightarrow{V_{tr}} \in N \otimes N$. Then the vector for a sentence headed by a transitive verb could be computed as the point-wise product of the verb's vector, and the tensor product of its subject and its object

$$\overrightarrow{subj\ V_{tr}\ obj} = \overrightarrow{V_{tr}} \odot (\overrightarrow{subj} \otimes \overrightarrow{obj}).$$

PGG CVSMs offer a formally grounded and computationally efficient method for obtaining vectors for complex expressions from their syntactic constituents. They permit the same kind of measurement for relations of semantic similarity among sentences that lexical VSMs give for word pairs. They can be trained on a (PGG parsed) corpus, and their performance evaluated against human annotators' semantic judgements for phrases and sentences. Grefenstette and Sadrzadeh (2011) report that their system outperforms Mitchell and Lapata (2008)'s multiplicative CVSM in a small scale corpus experiment on predicting semantic distance for pairs of simple transitive VP sentences.

The PGG CVSM raises at least two major difficulties First, while the vector of a complex expression is the value of a linear map on the vectors of its parts, it is not obvious what independent property this vector represents. Sentential vectors do not correspond to the distributional properties of these sentences, as the data in the primary

linguistic data (PLD) from which children learn their language is too sparse to estimate distributional vectors for all but a few sentences, across most dimensions.

Coecke et al. (2010) show that it is possible to encode a classical model theoretic semantics in their system by using vectors to express sets, relations, and truth-values. But this simply demonstrates the formal power of PGG CVSMs as semantic coding devices. CVSMs are empirically interesting to the extent that the sentential vectors that they assign are derived from lexical vectors that represent the actual distributional properties of these expressions.

In classical formal semantic theories the functions that drive semantic composition are supplied by the type theory, where the type of each expression specifies the formal character of its denotation in a model. The sequence of functions that determines the semantic value of a sentence exhibits at each point a value that directly corresponds to an independently motivated semantic property of the expression to which it is assigned. Types of denotation provide non-arbitrary formal relations between types of expressions and classes of entities specified relative to a model. The sentential vectors obtained from distributional vectors of lexical items lack this sort of independent status. In our fragment we have specified a conservative extension of a classical type system for computing probabilistic values for sentences and predicates. An important advantage of our approach is that we sustain the independently motivated denotations that a classical type system assigns to syntactically complex expressions within a probabilistic framework designed to capture the gradience and relative uncertainty of lexical semantic relations.

The second major problem is as follows. An important part of the interpretation of a sentence involves knowing its truth (more generally, its satisfaction or fulfillment) conditions. We have exchanged truth conditions for probability conditions formulated in terms of the likelihood of a sentence being accepted by competent speakers of the language as true, given certain states of affairs in the world. It is not obvious how we can extract either classical truth conditions, expressed in Boolean terms, or probability conditions, from sentential vector values, when these are computed from vectors expressing the distributional properties of their constituent lexical items. By contrast, our fragment offers a recursive specification of the meaning of a sentence which yields its probability conditions.

5 Probability, gradience, and vagueness

5.1 Two views of semantic vagueness

The fact that sentences receive probability conditions that express the likelihood that competent speakers would accept them as true relative to states of affairs permits us

to model the uncertainty that characterizes some of these speakers' judgements concerning the semantic relations and predications that hold for their language. This sort of uncertainty accounts for an important element of gradience in semantic knowledge. It captures the defeasibility of implications, and the graded nature of synonymy (co-intensionality) and meaning intersection. However, it remains unclear whether all species of semantic vagueness can be subsumed by the uncertainty that probabilistic judgements express. Consider, in particular, the case of degree adjectives and adverbs. If a door is slightly ajar, there is a sense in which it fully satisfies neither *open* nor *closed*.[5]

Two views (*inter alia*) have been proposed for determining the relation between probability and semantic vagueness. On one of these, vagueness can be characterized in terms of the truth of judgements that predicates apply to objects, modifiers to states or events, etc. The epistemicist account of vagueness (Williamson 1994) provides a prominent instance of this approach. It takes vagueness to consist in the same sort of uncertainty that attaches to epistemic claims about the world. This view is attractive to the extent that it can be used to support the idea that one models the gradience of semantic properties as a probability distribution over the applicability of expressions of different functional types to their arguments. However, it has the unattractive consequence that it assumes the existence of sharp boundaries on the extensions of predicates, but takes these to be epistemically opaque (essentially unknowable) to speakers of the language. Applying a predicate to an entity is, in many cases then, analogous to making a bet on the existence of a state of affairs, where one cannot identify the situation that decides the outcome of the wager. There appears to be no independent motivation for such unknowable limits on the extensions of terms. Therefore, it looks like an ad hoc device which the theory requires in order to explain the fact that vagueness, unlike epistemic uncertainty, cannot be eliminated by additional information about either language or the world.

Lassiter (2011) offers a refined alternative version of the view that vagueness is the expression of probability judgements. He avoids the epistemicist assumption of unknowable determinate predicate extensions, by replacing these with a set of possible languages all of whose expressions receive non-vague interpretations. Vagueness is the result of a probability distribution over these languages (their predicates) in different worlds. Speakers assign probabilities to language-world pairs, seeking to maximize the probability of pairs that converge on the observed linguistic and non-linguistic facts. This analysis characterizes a vague predicate as ambiguous among a large disjunction of semantically determinate variants over which probability is distributed. In order to express the gradient nature of vagueness it would seem to be necessary to proliferate

[5] We are grateful to Peter Sutton for helpful discussion of the issues that we deal with in this section.

a large (possibly unbounded) number of determinate readings for vague predicates to range over. This looks like an awkward result. Vagueness is naturally thought of an alternative to ambiguity rather than a consequence of it.

Edgington (1997) proposes the second view. She uses a Bayesian probability logic to model semantic vagueness, but she argues that vagueness and epistemic uncertainty are distinct. The problem with this approach is that it leaves the formal isomorphism between the two phenomena unexplained. If they really are different in the way that she suggests, then why should a calculus for computing the probability of statements under uncertainty provide a more accurate system for representing the vagueness of predicates than fuzzy or supervaluational logics, as she shows to be the case? The success of probabilistic models in expressing vagueness suggests that there is, in fact, a non-accidental connection between reasoning under conditions of epistemic uncertainty and the vagueness of predication. However, it may not be as direct or straightforward as the epistemicists hold it to be.

5.2 Semantic vagueness as an effect of learning

It might be possible to develop a third view by mediating the relation between probability and vagueness through learning. Speakers learn predicates by generalizing from paradigm instances where their applications to an object are valued as 0 or 1 in worlds of high probability. Extending the application of these predicates to new objects with different property sets will produce an update in the probability function of the model that estimates the likelihood of competent speakers assenting to the predications as intermediary or low. In the absence of additional disambiguating evidence, this probability distribution over worlds for a range of predicate applications will survive learning to be incorporated into the model of mature speakers. In this way uncertainty in learning becomes vagueness in the intensions of predicates in the target representation.

This approach treats epistemic uncertainty as a central element of semantic learning. The concern to converge on the classifiers that competent speakers apply drives the learner to update his/her probability distributions for the application of predicates (and other terms) in light of new linguistic and extra-linguistic evidence. But once the target representation is (more or less) achieved, many terms of the language remain under determined for objects in their domain. Vagueness is, then, the residue of probabilistic learning. It cannot be resolved by additional facts, linguistic or extra-linguistic, as it has been incorporated into the adult language itself. Therefore, it has its origin in probabilistic judgements on the truth of predication during the learning process, but it becomes an independent feature of the semantics of the language.

We offer this suggestion as the sketch of an alternative account of vagueness that seeks to account for it in probabilistic terms, but does not reduce it to epistemic uncer-

tainty in the competent speakers of the language. In order to be viable it is necessary to work out a detailed formal theory of semantic learning and the target language that it converges on. This is a research project that this paper is intended to introduce, rather than complete.

6 Conclusions and future work

Compositional VSMs can represent gradience in semantic relations among words, phrases, and sentences, and they offer a viable account of lexical semantic learning. However, the vectors that CVSMs assign to complex syntactic structures do not have clear interpretations, and they do not express sentential meaning as probability conditions.

We propose a fragment of a probabilistic semantic theory that uses a conservative extension of classical type theory to compute the probability value of a sentence on the basis of a model for the knowledge of a semantic learner. Our approach offers a framework for developing a probabilistic account of semantic learning that is consonant with current Bayesian approaches to classifier acquisition.

We suggest a view of vagueness that treats it as originating in the probabilistic judgements of semantic learning, but which develops into an independent non-epistemic variety of uncertainty in the mature target representation language.

Acknowledgements The initial research for this paper was done when the second author visited the first at the CWI for a month in the summer of 2011. The second author expresses his gratitude to the CWI for its generous hospitality and for the stimulating working environment that it provided during this time. Earlier versions of this paper were presented to a joint ILLC colloquium of the Computational Linguistics group and the DIP at the University of Amsterdam in September 2011, the King's College London Philosophy Colloquium in December 2011, The Oxford Advanced Seminar on Informatic Structures in December 2011, and the Hebrew University Logic, Language, and Cognition Center Colloquium in January 2012. We thank the participants of these meetings for their useful feedback and helpful suggestions. The second author is also grateful to Peter Sutton for helpful comments on an earlier draft of this paper, and for illuminating discussions of the relation between probability and semantic vagueness. These discussions have caused him to develop and refine some of the ideas proposed here. Of course we bear sole responsibility for any errors in the paper.

References

E. Bach. Natural languate metaphysics. In R. Barcan-Marcus, G. Dorn, and P. Weingartner, editors, *Logic, Methodology, and Philosophy of Science*, volume VII, pages 573-595. North Holland, Amsterdam, 1986.

J. van Benthem, J., Gerbrandy, and B. Kooi. Dynamic update with probabilities. Studia Logica, 93:67–96.

P. Blackburn and J. Bos. *Representation and Inference for Natural Language*. CSLI, Stanford, 2005.

J. Bos. Predicate logic unplugged. In *Proceedings of the Tenth Amsterdam Colloquium*. Amsterdam, Holland, 1995.

R. Carnap. *Logical Foundations of Probability*. Uniiversity of Chicago Press, Chicago, 1950.

C. Chelba. Statistical language modeling. In A. Clark, C. Fox, and S. Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*, pages 74–104. Wiley-Blackwell, Chichester, West Sussex and Malden, MA, 2010.

A. Clark and S. Lappin. Unsupervised learning and grammar induction. In A. Clark, C. Fox, and S. Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell, Oxford, 2010.

A. Clark and S. Lappin. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell, Chichester, West Sussex, and Malden, MA, 2011.

S. Clark. Statistical parsing. In A. Clark, C. Fox, and S. Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*, pages 333–363. Wiley-Blackwell, Chichester, West Sussex and Malden, MA, 2010.

B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis, Festschrift for Joachim Lambek*, 36:345–384, 2010.

M. Collins. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, 2003.

A. Copestake, D. Flickinger, C. Pollard, and I. Sag. Minimal recursion semantics. *Research on Language and Computation*, 3:281–332, 2006.

D. Edgington. Vagueness by degrees. In R. Keefe and P. Smith (Eds.) *Vagueness: A reader*, pp. 294–316. Cambridge, MA: MIT Press, 1997.

R. Fagin and J. Halpern. Uncertainty, belief, and probability. *Computational Intelligence*, 7:160–173, 1991.

C. Fox and S. Lappin. Expressiveness and complexity in underspecified semantics. *Linguistic Analysis, Festschrift for Joachim Lambek*, 36:385–417, 2010.

E. Grefenstette and M. Sadrzadeh. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, 2011.

E. Grefenstette, M. Sadrzadeh, S. Clark, B. Coecke, and S. Pulman. Concrete sentence spaces for compositional distributional models of meaning. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS-11)*, pages 125–134, Oxford, UK, 2011.

D. Jurafsky and J. Martin. *Speech and Language Processing*. Second Edition, Prentice Hall, Upper Saddle River, NJ, 2009.

C. Kemp, A. Perfors, and J. Tenenbaum. Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, 103:307–317, 2007.

A. Koller, M. Regneri, and S. Thater. Regular tree grammars as a formalism for scope underspecification. In *Proceedings the 46th Annual Meeting of the ACL*. Columbus, OH, 2008.

A. Kolmogorov. *Foundations of Probability*. Chelsea Publishing, New York, 1950.

J. Lambek. Pregroup grammars and chomsky's earliest examples. *Journal of Logic, Language and Information*, 17(2):141–160, 2008a.

J. Lambek. *From Word to Sentence*. Polimetrica, Milan, 2008b.

S. Lappin. On the pragmatics of mood. *Linguistics and Philosophy*, 4:559–578, 1982.

D. Lassiter. Vagueness as probabilistic linguistic knowledge. In R. Nouwen, R. van Rooij, U. Sauerland, and H.-C. Schmitz (Eds.), *Vagueness in communication*, pp. 127–150). Berlin: Springer, 2011.

C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

J. Mitchell and M. Lapata. Vector-based models of semantic composition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2008*, pages 236–244, 2008.

N. Nilsson. Probabilisitic logic. *Artificial Intelligence*, 28:71–87, 1986.

S. Padó and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):177–199, 2007.

J. Paris. Pure inductive logic. Winter School in Logic, Guangzhou, China, 2010.

U. Reyle. Dealing with ambiguities by underspecification: Construction, representation and deduction. *Journal of Semantics*, 10:123–179, 1993.

P. Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46:159–216, 1990.

P. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

T. Williamson. *Vagueness*. Routledge, London, 1994.

Quantum Probabilistic Dyadic Second-Order Logic

Alexandru Baltag, Jort M. Bergfeld, Kohei Kishida, Joshua Sack, Sonja J. L. Smets, and Shengyang Zhong

Institute for Logic, Language and Computation, Universiteit van Amsterdam
thealexandrubaltag@gmail.com

Abstract

We propose an expressive but decidable logic for reasoning about quantum systems. The logic is endowed with tensor operators to capture properties of composite systems, and with probabilistic predication formulas $P^{\geq r}(s)$, saying that a quantum system in state s will yield the answer 'yes' (i.e. it will collapse to a state satisfying property P) with a probability at least r whenever a binary measurement of property P is performed. Besides first-order quantifiers ranging over quantum states, we have two second-order quantifiers, one ranging over quantum-testable properties, the other over quantum "actions". We use this formalism to express the correctness of some quantum programs. We prove decidability, via translation into the first-order logic of real numbers.

1 Introduction

This paper introduces a powerful new logic for reasoning about quantum computation. Our *Quantum Probabilistic Dyadic Second-Order Logic* (**QPDSOL**) is *expressive enough* to capture superpositions, entanglements, measurements, quantum-logical gates and probabilistic features; it can express the correctness of a wide range of complex quantum protocols and algorithms; but at the same time it is logically tractable, in the sense of being *decidable*.

It is well-known that "classical" First-Order Logic is undecidable, and moreover that "classical" Second-Order Logic, as well as its monadic and dyadic fragments[1] are not even axiomatizable. By moving to the quantum world, it is natural to *extend* the range of first-order quantifiers to *quantum "states"* (i.e. superpositions of classical states), while at the same time it is natural to *restrict* the range of monadic second-order quantifiers to *quantum-testable properties* (closed linear subspaces of the state space), and to similarly restrict the range of dyadic second-order quantifiers to *quantum "actions"* (linear maps between state spaces). Indeed, it is widely accepted in the literature on Quantum Logic and on Foundations of Quantum Mechanics that quantum-testable properties are *the only* experimentally meaningful properties of a quantum system: any other (non-testable, non-linear) properties have no physical/experimental meaning in a quantum setting. Similarly, it is widely accepted in quantum computation that all meaningful quantum programs are obtainable by composing quantum gates (unitary maps) and quantum tests (measurements), and thus are quantum "actions" in the above sense.[2] So restricting the interpretations of the unary and binary predicates as above is a natural thing to do in a quantum setting: it only restricts the second-order quantifiers to properties/actions that are *physically meaningful*. The resulting logic *is indeed the natural "quantum analogue"* of classical (dyadic) second-order logic!

Surprisingly, this quantum analogue turns out to be much more tractable than its classical counterpart: the above well-justified and natural restrictions of range are enough to restore full decidability, even after the addition of "exotic" features such as probabilistic predication and tensors!

In a sense, this is not as surprising as it may first appear. Our semantics for second-order logic is "non-standard": not all sets of states (whose existence is guaranteed by the standard axioms of Set Theory) are accepted as "predicates". The second-order quantifiers are thus restricted to a limited range of predicates. Such non-standard variations of second-order logic have been studied before. Henkin's weak semantics for second-order logic (Henkin 1950) involves a restriction on the range of the second-order quantifiers (to some model-dependent class of admissible predicates), that re-

[1] *Monadic* Second-Order Logic is the fragment allowing quantification only over *unary* predicates, while the Dyadic fragment allows quantification only over *unary and binary* predicates.

[2] The converse is not obvious, and may even fail in practice. But from a theoretical perspective, one can argue that the converse is true in a sense: for any quantum action (linear map) f between systems $\mathcal{H}$ and $\mathcal{H}'$ there exists an entangled state s_f in $\mathcal{H} \otimes \mathcal{H}'$ with the property that, if a local measurement performed on the $\mathcal{H}$-subsystem of (a system in state) s_f yields state x, then after that a local measurement on the $\mathcal{H}'$-subsystem will yield the result $f(x)$. In this way, any such action f can be physically computed, in principle: first, prepare a large number of entangled states s_f; then perform local measurements on the $\mathcal{H}$-subsystem until one of them yields the desired input value x; and then perform a measurement on the $\mathcal{H}'$-subsystem, yielding the output-value $f(x)$.

stores the axiomatizability of the logic. Some variants of monadic second-order logic (for very restricted models) are even decidable (Rabin 1969).

But these classical results are conceptually very different from ours: none of these weaker logics can be considered to be a genuine and natural variant of second-order logic. In particular, Henkin's semantics (restricting second-order quantifiers to some arbitrary collections of subsets of the state space) is not a independently-justifiable restriction. It does not even provide a unique, canonical way to restrict the quantifiers (but a model-dependent one). In contrast, our restriction of quantifiers to quantum-testable properties (and quantum-performable operations) is natural, canonical (providing a unique collection for each dimension) and amply justified on independent grounds by a whole body of literature in Quantum Logic, Foundations of Quantum Mechanics and Quantum Computation.

Indeed, seen from the perspective of the quantum world, our "non-standard" semantics *looks like the "true" semantics* of second-order logic: it only eliminates the predicates that are "physically meaningless". Moreover, while in a sense being a restriction of the classical (standard) semantics, in a different sense this can be thought of as *an extension of the classical semantics*! Indeed, one can argue that, if we restrict ourselves to *classical states* (i.e., n-long tuples of bits $|0\rangle$ or $|1\rangle$, for any dimension n) then *all the standard predicates of such classical states are realized as quantum-testable predicates* (and hence fall within the range of our second-order quantifiers): for *every* set $A \subseteq \{|0\rangle, |1\rangle\}^n$, there exists a unique quantum-testable predicate (linear subspace[3]) $P_A \subseteq \mathcal{H}_2^{\otimes n}$ such that a classical n-state $s \in \{|0\rangle, |1\rangle\}^n$ satisfies P_A iff it belongs to the set A. So, insofar as *classical* states are concerned, our range restriction for second-order quantifiers *is not a restriction at all*: their range really includes (quantum counterparts of) *every set* of classical states. It is only when we look at non-classical (superposed) states that we see that the quantifier range is restricted (though in a natural way).

In conclusion, regardless of whether one considers it as a natural restriction of the classical semantics for (predicates of) quantum states, or as a huge extension of the classical semantics for (predicates of) classical states, we can still safely claim that *our logic really is the correct quantum (and probabilistic) counterpart of the classical (dyadic) second-order logic*.

As a consequence, we think that our decidability result is a significant contribution to the logical understanding of quantum mechanics: it shows in essence that, whereas the natural formulation of (dyadic) second-order logic in the *classical* world is undecidable, *the natural formulation of (dyadic) second-order logic for the quantum world is decidable.*

[3] In fact, this is the linear subspace P_A generated by A.

The fundamental reason for this tractability is the one severe constraint put by quantum mechanics on the "meaningful" properties and actions: *linearity*.[4] Once again, this does not really restrict the predicates/actions as far as classical states are concerned (since any two classical states of the same space are orthogonal to each other, a classical state cannot be written as a linear combination of other classical states). But linearity *does* constrain the behavior of "meaningful" predicates/actions on *superposed* states. And, in the end, linearity allows the reduction of all the "meaningful" second-order objects (predicates/actions) to their underlying linear expressions: matrices of (complex) numbers.

So this natural (and physically-imposed) linearity constraint reduces thus our quantum version of second-order logic to the *first-order theory* of complex numbers. And now, a classical result comes to our help: while first-order logic is in general undecidable (and the first-order theories of many useful structures, such as the ring of natural numbers, are not even axiomatizable), *the first-order theory of complex numbers is decidable*. This was pointed out by Alfred Tarski (1948) as a corollary to the analogue result for the field of real numbers (proved in the same paper by quantifier-elimination).

Our decidability proof makes essential use of Tarski's decidability result, as well as of the finite dimensionality; it translates effectively the probabilistic dyadic second-order logic of finite-dimensional quantum systems into the decidable first-order theory of reals. This proof method is inspired by the one given in (Dunn et al. 2005), where the traditional (propositional) quantum logic of any finite-dimensional Hilbert space was proved to be decidable. However, the result in (Dunn et al. 2005) required that we first fix a particular Hilbert space (model of a quantum system) of a finite dimension, so as to translate the logic of the space into the finitary language of reals, thus limiting the scope of application by fixing a finite dimension (and hence the number of *quantum bits* or *qubits*) throughout the discourse. In contrast, our logic overcomes this limitation by using types and tensors in the language, thus accommodating *an unbounded number of qubits*, while preserving the logical tractability.

Our results in this paper can be seen as part of a wider on-going effort towards bridging the gap between traditional quantum logic and the theory of quantum computation. On the one hand, traditional quantum logic (as originated in Birkhoff and von Neumann 1936) has focused on axiomatics and logical properties of the lattice of closed linear subspaces of an *infinite-dimensional* Hilbert space, with the goal being "to discover the logical structure one may hope to find in physical theories which, like QM, do not conform to classical logic" (Birkhoff and von Neumann 1936). Quantum computation, on the other hand, concerns encoding and describing computations

[4] For unary predicates: having a linear subspace (not an arbitrary subset) as their extension; for actions: being induced by a linear map.

on the basis of quantum systems, and involves quantum ingredients such as superposition and entanglement, in order to perform certain tasks much faster than classical computers. The underlying theoretical framework for quantum computation is given by *finite-dimensional* Hilbert spaces. Among the few treatments of such finite-dimensional quantum logics and their decidability are the work of Chadha et al. (2009), Dunn et al. (2005).

Another contrast between quantum logic and quantum computation lies in the treatment of "quantum entanglement". In traditional quantum logic, entanglement has been viewed as a problem-child, posing difficulties to the lattice-theoretic setting (Aerts 1981, Randall and Foulis 1979) (though naturally treatable in a category-theoretical setting (Abramsky and Coecke 2004, Selinger 2004)). In quantum computing, however, entanglement is viewed as a *computational resource*, that allows us to go beyond the world of classical computing. Among the papers that address this part of the gap between quantum logic and quantum computation are (Baltag et al. 2013, Chadha et al. 2009), and (Dalla Chiara et al. 2004, Chapter 17). Our work strengthens the connection further. The logic we propose in the following sections—dyadic second-order quantum logic—is fit to deal with multi-partite systems that exhibit quantum entanglement. Equipped with an explicitly typed language, with types for states, predicates, and actions, with tensor operators connecting them, as well as with probabilistic predication, our logic allows us to capture all the essential computational properties of composite quantum systems, and in particular it can encode the correctness of a wide range of quantum algorithms.

The design of dyadic second-order quantum logic in this paper builds further on the earlier work of Baltag and Smets (2005; 2006) on propositional dynamic quantum logics. It is well known that standard Propositional Dynamic Logic (PDL), as well as its fragment called the Hoare Logic, plays an important role in classical computing and programming. In particular, PDL and Hoare Logic are among the main logical formalisms used for classical program verification. The quantum version of PDL extends the area of applicability to the verification of quantum programs and quantum protocols. In (Baltag and Smets 2006), a quantum dynamic logic was designed that was expressive enough to prove the correctness of basic non-probabilistic quantum protocols such as teleportation and quantum secret sharing. The work of Baltag et al. (2012) used the tools of Dunn et al. (2005) to prove the decidability of such a propositional dynamic quantum logical system. While these results are important, note that the logic in (Baltag et al. 2012) was unable yet to capture the correctness of any probabilistic quantum protocols. In this paper, we overcome this limitation and equip our logic with a *probabilistic predication operator*, indicating that a state of a quantum system will collapse to a state having property P with probability at least r whenever a measurement of property P is performed. This operator allows us to express the correctness of

those quantum algorithms (such as quantum search) that make essential use of quantum probabilities.

A remark is in order regarding the fact that each given program in our syntax, and so each given sentence, uses only a given number of qubits (and thus it refers to a Hilbert space with a given finite number of dimensions). We would like to stress that our result is much more significant than, say, the decidability of checking the correctness of a classical circuit of a given size applied to a problem of given input size. This is because *we do not fix the size of the input, but only the dimension*. This point is important, since for a given fixed dimension (greater than 1) there are *infinitely* (in fact *uncountably*) many non-equivalent quantum states of that dimension (while typically there are only finitely many inputs of a given size). Hence, the algorithm for deciding satisfiability (on states of a space of given dimension) *cannot* simply proceed by exhaustive search over a finite domain (as in the case of models of bounded size). The correctness statements presented in this paper really capture the correctness of a program for uncountably many non-equivalent inputs![5]

2 Preliminaries

According to quantum theory (see, e.g. Nielsen and Chuang 2011), any quantum system can be described by a Hilbert space $\mathcal{H}$ of appropriate dimension. Similar to the tradition of Piron (1976), we identify *(pure) states* of the system with the "rays" in $\mathcal{H}$ (i.e. the one-dimensional linear subspaces of $\mathcal{H}$) and the "impossible state" (zero-dimensional subspace, which we include as it allows us to discuss only total functions without loss of generality). Given a vector $|\psi\rangle \in \mathcal{H}$, we will write $\widehat{|\psi\rangle}$ for the state generated by $|\psi\rangle$. Given a state space $\mathcal{H}$ of some quantum system, we write $\Sigma_{\mathcal{H}}$ for the set of all states, i.e. the set of all one-dimensional linear subspaces of $\mathcal{H}$ and $\widehat{\mathbf{0}_{\mathcal{H}}}$ (where $\mathbf{0}_{\mathcal{H}}$ is the zero vector).

Any change of the state of a quantum system can be described by a linear map on $\mathcal{H}$. There are two important kinds of linear maps: unitary operators and projectors. A *unitary operator* U is such that both $U^\dagger U$ and $UU^\dagger$ are the identity operator, where $(\cdot)^\dagger$ is the adjoint operation on linear maps. In quantum computation, unitary operators are the counterpart of logical gates in classical computation. An operator A is a *projector*, if it is bounded, idempotent, i.e. $AA = A$, and self-adjoint, i.e $A^\dagger = A$. Projectors

[5] Moreover, these correctness statements, even when translated back into the arithmetic of real numbers, do *not* boil down to simple equations involving addition and multiplication of *specific* real numbers and/or matrices. Instead, they reduce to complex first-order statements in the theory of real numbers, that involve in an essential way quantification over uncountably many objects. It just happens that (due to Tarski's theorem) this theory is still decidable!

are essential in describing quantum measurements, which are the only way we extract information from a quantum system. In this paper, our level of abstraction allows us to consider not only linear maps on a Hilbert space but also those between different Hilbert spaces. Every linear map $A : \mathcal{H} \to \mathcal{H}'$ from a quantum system $\mathcal{H}$ to a possibly different system $\mathcal{H}'$ naturally induces a unique function (also denoted by A) from the set of states $\Sigma_{\mathcal{H}}$ to the set of set of states $\Sigma_{\mathcal{H}'}$, given by $A(\widehat{|\psi\rangle}) := \widehat{A(|\psi\rangle)}$ for every $|\psi\rangle \in \mathcal{H}$. An *action* is any such function $A : \mathcal{H} \to \mathcal{H}'$ induced on state spaces by some linear map $A : \Sigma_{\mathcal{H}} \to \Sigma_{\mathcal{H}'}$. We can also define composition, tensor product and adjoint of actions in a natural way via composition, tensor product and adjoint of linear maps which induce the actions[6]. We will use the same symbols for operations on actions as those for linear maps.

In this paper, a *property* of a quantum system with state space $\mathcal{H}$ is just a subset of $\Sigma_{\mathcal{H}}$. However, according to quantum theory, not any subset of $\Sigma_{\mathcal{H}}$ represents a property of the system that can be tested. A property is *testable* iff it corresponds to a closed linear subspace W of $\mathcal{H}$ in such a way that the states in the property are exactly those generated by vectors in W. Since this correspondence is one-to-one and natural, we will always use the same symbol to denote a testable property and its corresponding closed linear subspace. Moreover, according to linear algebra, closed linear subspaces lie in one-to-one correspondence with projectors in the following sense:

1. For every projector A on $\mathcal{H}$, $ran(A)$ (the range of A) is a closed linear subspace of $\mathcal{H}$, and for every vector $|\psi\rangle \in \mathcal{H}$, $|\psi\rangle \in ran(A)$ iff $A(|\psi\rangle) = |\psi\rangle$.

2. For every closed linear subspace W of $\mathcal{H}$, there is a *unique* projector on $\mathcal{H}$, called *the projector onto* W and denoted by $?^{\mathcal{H}}(W)$, such that for every vector $|\psi\rangle \in \mathcal{H}$, $|\psi\rangle \in W$ iff $?^{\mathcal{H}}(W)(|\psi\rangle) = |\psi\rangle$.

The state space of a qubit, the unit of quantum information, is of dimension 2. Given a fixed orthonormal basis $\{|0\rangle, |1\rangle\}$ of the state space of a qubit, the two states generated by $|0\rangle$ and $|1\rangle$, respectively, correspond to the values 0 and 1 of a classical bit. Given several qubits indexed by elements in a finite set I, they form a compound quantum system, and the state space for I is the tensor product $\bigotimes_{i\in I} \mathcal{H}_i$ of the state space $\mathcal{H}_i$ for each qubit $i \in I$. A standard way of obtaining an orthonormal basis of this state space is to take tensor products of vectors in the fixed orthonormal bases of each $\mathcal{H}_i$. It is easy to see that there are $2^{|I|}$ vectors in this basis, and we will index them by elements in ${}^I\mathbf{2}$, the set of all functions from I to $\mathbf{2} = \{0, 1\}$, in such a way that $|f\rangle = \otimes_{i\in I} |f(i)\rangle_i$, for each $f \in {}^I\mathbf{2}$. We call a state of a compound system *classical* if it is generated by a vector in this basis. Moreover, we write $|0\rangle_I$ for $\otimes_{i\in I} |0\rangle_i$.

[6]Note that different linear maps could induce the same action, but the operations on actions are still well-defined according to linear algebra.

It is well known that an n-dimensional Hilbert space is isomorphic to $\mathbb{C}^n$. In this case, every linear subspace is closed and every operator is bounded. Moreover, every state can be represented by n complex numbers if we pick a vector in the state as its representative. Every property, identified with its corresponding projector, can be represented by an $n \times n$ matrix of complex numbers. Every linear map from an n-dimensional Hilbert space to an m-dimensional one can be represented by an $m \times n$ matrix of complex numbers.

In this paper, for generality, we assume that we are supplied with countably infinitely many qubits indexed by elements in ω, the set of all natural numbers, which we take to be non-negative integers. We further assume that an orthonormal basis has been fixed for each qubit, and we obtain an orthonormal basis for compound systems consisting of a finite number of qubits by applying the tensor product in the way just described. Finally, we use $\mathcal{P}_{\text{fin}}(\omega)$ to denote the set of all finite, *non-empty* subsets of ω. For each $\tau \in \mathcal{P}_{\text{fin}}(\omega)$, by *$\tau$-system* we mean the quantum system consisting of qubits indexed by elements of τ. Whenever $\mathcal{H}_\tau$, the state space of the τ-system, appears as a superscript or subscript in a symbol, we simply write τ; for example, we write simply Σ_τ for $\Sigma_{\mathcal{H}_\tau}$.

Moreover, for each $\tau, \rho \in \mathcal{P}_{\text{fin}}(\omega)$ s.t. $\tau \subseteq \rho$, we know from linear algebra that $\mathcal{H}_\tau$ can be canonically embedded into $\mathcal{H}_\rho$, by "padding" all the vectors with $|0\rangle$'s for all the extra dimensions. Hence in this paper we write $\Theta_{\tau\to\rho} : \mathcal{H}_\tau \to \mathcal{H}_\rho$ for this canonical embedding

$$\Theta_{\tau\to\rho} = \textstyle\sum_{f\in\tau\mathbf{2}}(|f\rangle \otimes |0\rangle_{\rho\setminus\tau})\,\langle f|.$$

We also write $\Theta_{\rho\to\tau} : \mathcal{H}_\rho \to \mathcal{H}_\tau$ for the canonical projection that reverses the above embedding:

$$\Theta_{\rho\to\tau} = \textstyle\sum_{f\in\tau\mathbf{2}}|f\rangle\,(\langle f| \otimes \langle 0|_{\rho\setminus\tau}).$$

Using the canonical embeddings and projections, one can *generalize projectors to arbitrary dimensions*: For every space $\mathcal{H}_\tau$ and every closed linear subspace W_ρ of some other space $\mathcal{H}_\rho$, we can define *the generalized projector of $\mathcal{H}_\tau$ onto W_ρ*, denoted by $?^\tau(W_\rho)$, by putting:

$$?^\tau(W_\rho) = \Theta_{\rho\cup\tau\to\rho} \circ \left(?^\rho(W_\rho) \otimes |0\rangle_{\tau\setminus\rho}\,\langle 0|_{\tau\setminus\rho}\right) \circ \Theta_{\tau\to\rho\cup\tau}$$

This is a linear map that takes a vector in $\mathcal{H}_\tau$ and "projects" it onto W_ρ. Physically, this action corresponds to *a successful measurement of a ρ-property performed on a τ-system.*

We introduce some notation. Given a binary relation R and a set $A \subseteq dom(R) = \{x \mid \exists y.\,(x,y) \in R\}$, let $R[A] \stackrel{\text{def}}{=} \{b \mid \exists a{\in}A.\,(a,b){\in}R\}$ be the direct image of A under R. Given

a set $B \subseteq ran(R) = \{y \mid \exists x.(x,y) \in R\}$, we let $[R]B \stackrel{\text{def}}{=} \{a \mid \forall b.(a,b) \in R \Rightarrow b \in B\}$ be the so-called weakest precondition of B under R. Note that when R is a function instead of a relation in general, $[R]B$ is sometimes called the inverse image of B under R. In general, given two sets A and B, we write ${}^{A}B$ for the set of functions from A to B. Given a positive number N, let $\mathbf{N} = \{0, 1, \ldots, N-1\}$. Given a linear map T, let $\mathbf{T}$ be its matrix representation under the fixed bases.

3 Quantum Probabilistic Dyadic Second-Order Logic

Syntax of QPDSOL Our language consists of terms (for quantum states), predicates symbols (for quantum testable properties), and function symbols (for actions). The language is *typed*: each of these symbols comes with a type, which is an element of $\mathcal{P}_{\text{fin}}(\omega)$, indicating the underlying set of qubits involved in that state, property or action. E.g. terms of type τ refer to the possible (pure) states of the τ-system; predicate symbols of type τ are unary predicates referring to *quantum-testable properties* of the τ-system; function symbols of type $\tau \to \rho$ are dyadic predicates (restricted to functions) referring to *actions*. As the types range over all of $\mathcal{P}_{\text{fin}}(\omega)$, the entire domain of discourse involves infinitely many qubits; but each formula involves only finitely many types, each involving only finitely many qubits, so that a formula can only talk about finitely many qubits.

For each pair of elements $\tau, \rho \in \mathcal{P}_{\text{fin}}(\omega)$, we include in the language a countable set of *state variables* x_τ of type τ, a countable set of *state constants* c_τ of type τ, a countable set of *predicate variables* p_τ of type τ, a countable set of *predicate constants* T_τ of type τ, a countable set of *action variables* $a_{\tau\to\rho}$ of type $\tau \to \rho$, and a countable set of *action constants* $C_{\tau\to\rho}$ of type $\tau \to \rho$. It is assumed that these sets are pairwise disjoint, and that each of them is indexed by elements in ω without repetition.

Definition 3.1. For any $\tau, \rho \in \mathcal{P}_{\text{fin}}(\omega)$, we define by (triple) mutual recursion the following sets of syntactic expressions: the set $\mathcal{T}\rceil\nabla\Updownarrow_\tau$ of *terms of type* τ

$$t_\tau ::= \; x_\tau \mid c_\tau \mid t_{\tau_1} \otimes t_{\tau_2} \mid \alpha_{\rho\to\tau}(t_\rho)$$

where $\tau_1, \tau_2 \in \mathcal{P}_{\text{fin}}(\omega)$ are such that $\tau_1 \cup \tau_2 = \tau$, $\tau_1 \cap \tau_2 = \emptyset$), the set $\mathcal{P}_\tau$ of *(unary) predicate symbols of type* τ

$$P_\tau ::= p_\tau \mid T_\tau \mid t_\tau \mid {\sim}P_\tau \mid P_\tau \cap P_\tau \mid P_{\tau_1} \otimes P_{\tau_2} \mid \alpha_{\rho\to\tau}[P_\rho] \mid [\alpha_{\tau\to\rho}]P_\rho$$

where $\tau_1, \tau_2 \in \mathcal{P}_{\text{fin}}(\omega)$ are such that $\tau_1 \cup \tau_2 = \tau$, $\tau_1 \cap \tau_2 = \emptyset$, and the set $\mathcal{A}_{\tau\to\rho}$ of *function symbols of type* $\tau \to \rho$

$$\alpha_{\tau\to\rho} ::= a_{\tau\to\rho} \mid C_{\tau\to\rho} \mid ?^\tau P_\rho \mid \alpha^\dagger_{\rho\to\tau} \mid \alpha_{\tau\to\mu}; \alpha_{\mu\to\rho} \mid \alpha_{\tau_1\to\rho_1} \otimes \alpha_{\tau_2\to\rho_2}$$

where $\mu, \tau_1, \rho_1, \tau_2, \rho_2 \in \mathcal{P}_{\text{fin}}(\omega)$ are such that $\tau_1 \cup \tau_2 = \tau$, $\rho_1 \cup \rho_2 = \rho$ and $\tau_1 \cap \tau_2 = \rho_1 \cap \rho_2 = \emptyset$.

We write $\mathcal{T}\rceil\nabla\Updownarrow$ for the set $\bigcup_{\tau\in\mathcal{P}_{\text{fin}}(\omega)} \mathcal{T}\rceil\nabla\Updownarrow_\tau$ of all terms, $\mathcal{P}$ for the set $\bigcup_{\tau\in\mathcal{P}_{\text{fin}}(\omega)} \mathcal{P}_\tau$ of all predicate symbols, and $\mathcal{A}$ for the set $\bigcup_{\tau,\rho\in\mathcal{P}_{\text{fin}}(\omega)} \mathcal{A}_{\tau\to\rho}$ of all function symbols. When $\tau = \rho$, we simply write $P_\rho?$ for the function symbol $?^\tau P_\rho$.

Definition 3.2. We now define by induction the set $\mathcal{L}$ of *formulas* of our logic:

$$\varphi ::= P_\tau^{\geq r}(t_\tau) \mid \neg\varphi \mid \varphi \wedge \varphi \mid \forall x_\tau \varphi \mid \forall p_\tau \varphi \mid \forall a_{\rho\to\tau}\varphi$$

where $\tau \in \mathcal{P}_{\text{fin}}(\omega)$, $t_\tau \in Term_\tau$, $P_\tau \in \mathcal{P}_\tau$ and $r \in [0, 1]$ is a definable real number (described below before Definition 3.3).

The intended meaning of our basic formula $P_\tau^{\geq r}(t_\tau)$ is that a *quantum system in state t_τ will yield the answer 'yes'* (i.e. it will collapse to a state satisfying property P_τ) *with a probability at least r whenever a binary measurement of property P_τ is performed*. The rest of our logical formulas are built from such basic formulas using standard Boolean connectives, as well as three types of quantifiers: first-order quantifiers $\forall x_\tau$ ranging over quantum states, second-order quantifiers $\forall p_\tau$ over quantum (testable) predicates, and second-order quantifiers $\forall a_{\tau\to\rho}$ ranging over quantum actions.

The notions of free variables, bound variables, etc. are defined in the standard way. As usual, a formula $\varphi \in \mathcal{L}$ is called *closed* if it has no free (state, predicate or action) variables. A *pure* formula is a closed formula containing no (state, predicate or action) constants.

Semantics of QPDSOL Following standard practice, we introduce the notion of *frame* (also known as *structure* in the semantics of first-order logic), by which we mean a structure that fixes the (state, predicate and action) constants. Then, given a frame, we define a *model* on it (also known as an *interpretation* in the semantics of first-order logic), which can determine the denotation of each remaining term, predicate symbol and function symbol. Finally, we define the *satisfaction relation*.

Recall that we say that a real number r is *definable* if there is a formula $\varphi(x)$ in the first-order language of $(\mathbb{R}, +, \cdot, 0, 1)$ such that $(\mathbb{R}, +, \cdot, 0, 1) \models \varphi[r] \wedge \forall x(\varphi(x) \to x = r)$. We also say that a complex number z is *simple* if $z = a + bi$ for definable real numbers a and b. Extending the terminology, we say that a state of the τ-system, a testable property of the τ-system and an action from the τ-system to ρ-system are *definable* if they can be represented under the fixed basis respectively by a $2^{|\tau|}$-tuple (with the state identified with the representative of it), a $2^{|\tau|} \times 2^{|\tau|}$-matrix (with the closed linear subspace identified with the corresponding projector), and a $2^{|\rho|} \times 2^{|\tau|}$-matrix (with the action identified with a linear map that induces it) of simple complex numbers.

Definition 3.3. An $\mathcal{H}$*-valuation* is a function V defined on a subset of $\mathcal{P} \cup \mathcal{A} \cup \mathcal{T}\rceil\nabla\Updownarrow$ and satisfying the following conditions:

- $V(t_\tau) \in \Sigma_\tau$ if $t_\tau \in \mathcal{T}\rceil\nabla\Updownarrow_\tau$;
- $V(P_\tau)$ is a testable property of τ-system, if $P_\tau \in \mathcal{P}_\tau$;
- $V(\alpha_{\tau\to\rho})$ is an action from Σ_τ to Σ_ρ if $\alpha_{\tau\to\rho} \in \mathcal{A}_{\tau\to\rho}$.

Definition 3.4. A *frame* $\mathfrak{F}$ is an $\mathcal{H}$-valuation whose domain is the set of all (state, predicate and action) constants and whose values are all definable.

Actually, for the decidability result to hold, a frame must be a computable function in some sense. We neglect this technicality here.

Definition 3.5. A *model* $\mathfrak{M}$ on a frame $\mathfrak{F}$ is an $\mathcal{H}$-valuation whose domain is $\mathcal{P} \cup \mathcal{A} \cup \mathcal{T}\rceil\nabla\Updownarrow$, that extends $\mathfrak{F}$ and that satisfies the following, for any terms $t_\tau, t_{\tau_1}, t_{\tau_2}$, function symbols $\alpha_{\tau\to\rho}, \beta_{\rho\to\mu}, \alpha_{\tau_1\to\rho_1}, \alpha_{\tau_2\to\rho_2}$, and predicate symbols $P_\tau, Q_\tau, P_\rho, P_{\tau_1}, Q_{\tau_2}$ such that $\tau_1 \cap \tau_2 = \emptyset$ and $\rho_1 \cap \rho_2 = \emptyset$:

$\mathfrak{M}(t_{\tau_1} \otimes t_{\tau_2})$	$=$	$\mathfrak{M}(t_{\tau_1}) \otimes \mathfrak{M}(t_{\tau_2})$
$\mathfrak{M}(\alpha_{\tau\to\rho}(t_\tau))$	$=$	$\mathfrak{M}(\alpha_{\tau\to\rho})(\mathfrak{M}(t_\tau))$
$\mathfrak{M}(\alpha_{\tau\to\rho}; \beta_{\rho\to\mu})$	$=$	$\mathfrak{M}(\beta_{\rho\to\mu}) \circ \mathfrak{M}(\alpha_{\tau\to\rho})$
$\mathfrak{M}(\alpha^\dagger_{\tau\to\rho})$	$=$	$(\mathfrak{M}(\alpha_{\tau\to\rho}))^\dagger$
$\mathfrak{M}(\alpha_{\tau_1\to\rho_1} \otimes \alpha_{\tau_2\to\rho_2})$	$=$	$\mathfrak{M}(\alpha_{\tau_1\to\rho_1}) \otimes \mathfrak{M}(\alpha_{\tau_2\to\rho_2})$
$\mathfrak{M}(?^\tau P_\rho)$	$=$	$?^\tau(\mathfrak{M}(P_\rho))$
$\mathfrak{M}(\sim P_\tau)$	$=$	$\sim\mathfrak{M}(P_\tau)$
$\mathfrak{M}(P_\tau \cap Q_\tau)$	$=$	$\mathfrak{M}(P_\tau) \cap \mathfrak{M}(Q_\tau)$
$\mathfrak{M}(P_{\tau_1} \otimes Q_{\tau_2})$	$=$	$\mathfrak{M}(P_{\tau_1}) \otimes \mathfrak{M}(Q_{\tau_2})$
$\mathfrak{M}(\alpha_{\tau\to\rho}[P_\tau])$	$=$	$\mathfrak{M}(\alpha_{\tau\to\rho})[\mathfrak{M}(P_\tau)]$
$\mathfrak{M}([\alpha_{\tau\to\rho}]P_\rho)$	$=$	$[\mathfrak{M}(\alpha_{\tau\to\rho})]\mathfrak{M}(P_\rho)$

To interpret quantifiers, for each (state, predicate, or action) variable v we introduce an equivalence relation $\sim_v$ among models on the same frame such that $\mathfrak{M} \sim_v \mathfrak{M}'$ iff $\mathfrak{M}(v') = \mathfrak{M}'(v')$ for all variables v' except possibly v.

Definition 3.6. The *satisfaction relation* between a model $\mathfrak{M}$ and a formula is defined recursively, where v is any (state, predicate, or action) variable,

$$
\begin{array}{lcl}
\mathfrak{M} \models P^{\geq r}_\tau(t_\tau) & \iff & |\langle\psi|\, ?^\tau(\mathfrak{M}(P_\tau))\,|\psi\rangle\,|^2 \geq r\||\psi\rangle\|^2\|?^\tau(\mathfrak{M}(P_\tau))\,|\psi\rangle\|^2, \\
 & & \text{for any vector } |\psi\rangle \in \mathfrak{M}(t_\tau) \\
\mathfrak{M} \models \neg\varphi & \iff & \mathfrak{M} \not\models \varphi, \\
\mathfrak{M} \models \varphi \wedge \psi & \iff & \mathfrak{M} \models \varphi \text{ and } \mathfrak{M} \models \psi, \\
\mathfrak{M} \models \forall v\varphi & \iff & \mathfrak{M}' \models \varphi, \text{ for all } \mathfrak{M}' \sim_v \mathfrak{M}.
\end{array}
$$

Obviously, other Boolean connectives such as $\vee$, $\rightarrow$ and $\leftrightarrow$ can be defined in the usual manner. Existential quantifiers over states, predicates and actions can also be defined in the usual manner. Moreover, this logic is at least as expressive as the first-order language of the lattice $L(\mathbb{C}^{2^n})$, which is discussed in (Dunn et al. 2005).

Now we introduce some useful abbreviations:

$$P_\tau^{\leq r}(t_\tau) \stackrel{\text{def}}{=} (\sim P)_\tau^{\geq(1-r)}(t_\tau) \qquad P_\tau^{=r}(t_\tau) \stackrel{\text{def}}{=} P_\tau^{\geq r}(t_\tau) \wedge P_\tau^{\leq r}(t_\tau)$$

$$P_\tau^{<r}(t_\tau) \stackrel{\text{def}}{=} \neg P_\tau^{\geq r}(t_\tau) \qquad P_\tau^{>r}(t_\tau) \stackrel{\text{def}}{=} \neg P_\tau^{\leq r}(t_\tau)$$

$$s_\tau \perp t_\tau \stackrel{\text{def}}{=} s_\tau^{\leq 0}(t_\tau)$$

$$s_\tau \doteq t_\tau \stackrel{\text{def}}{=} \left[s_\tau^{=1}(t_\tau) \wedge \neg(s_\tau \perp t_\tau)\right] \vee \left[(s_\tau \perp s_\tau) \wedge (t_\tau \perp t_\tau)\right]$$

Essentially, the meaning of $P_\tau^{\leq r}(t_\tau)$ (or respectively $P_\tau^{=r}(t_\tau)$, $P_\tau^{<r}(t_\tau)$, $P_\tau^{>r}(t_\tau)$) is that a quantum system in state t_τ will yield the answer 'yes' (i.e. it will collapse to a state satisfying property P_τ) with a probability $\leq r$ (or respectively $= r$, $< r$, $> r$) whenever a binary measurement of property P_τ is performed. Moreover, $\mathfrak{M} \models s_\tau \perp t_\tau$ iff s_τ and t_τ denote two orthogonal states. (Note that the impossible state $\widehat{\mathbf{0}_\tau}$ is the only state that is orthogonal to itself.) Finally, we have $\mathfrak{M} \models s_\tau \doteq t_\tau$ iff s_τ and t_τ refer to *the same state*: the first disjunct ensures that s_τ and t_τ are equal but neither denotes $\widehat{\mathbf{0}_\tau}$ (note that $s_\tau^{=1}(t_\tau)$ and $s_\tau \perp t_\tau$ are together satisfiable where either s_τ or t_τ is interpreted by $\widehat{\mathbf{0}_\tau}$), while the second disjunct ensures that both s_τ and t_τ denote $\widehat{\mathbf{0}_\tau}$.

We now define the notion of validity.

Definition 3.7. A formula φ of $\mathcal{L}$ is said to be *valid in a frame* $\mathfrak{F}$, written $\mathfrak{F} \models \varphi$, if $\mathfrak{M} \models \varphi$ for every model $\mathfrak{M}$ on $\mathfrak{F}$. A formula φ of $\mathcal{L}$ is said to be *valid*, written $\models \varphi$, if $\mathfrak{F} \models \varphi$ for every frame $\mathfrak{F}$.

As in classical predicate logic, we have

Lemma 1. *For every* closed *formula φ in $\mathcal{L}$ and every frame $\mathfrak{F}$, $\mathfrak{F} \models \varphi$ iff there is a model $\mathfrak{M}$ on $\mathfrak{F}$ such that $\mathfrak{M} \models \varphi$. For every* pure *formula φ in $\mathcal{L}$, $\models \varphi$ iff there is a frame $\mathfrak{F}$ such that $\mathfrak{F} \models \varphi$.*

4 Examples

Here we show how our language can be used to express many properties of quantum algorithms. We start with introducing some notation that will be commonly used in the following examples.

First, for each qubit i, we introduce state constants 0_i and 1_i to denote the state generated by $|0\rangle_i$ and $|1\rangle_i$, respectively.

We furthermore have the following action constants for a single qubit i, and for some, we provide the matrix representation (in the fixed bases) of linear maps which are usually used to induce the actions interpreting these constants:

- I_i interpreted as the identity action,
- H_i the action induced by the Hadamard gate with matrix $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$,
- X_i the action induced by the Pauli X gate $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$,
- Z_i the action induced by the Pauli Z gate $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

We furthermore have an action symbol $\mathcal{CNOT}_{ij}$ $(i \neq j)$ for the *controlled-NOT* action with control qubit i and target qubit j usually induced by a linear map with the matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

For any distinct i and j, we also define an abbreviation for an action that interchanges the states of qubits i and j:

$$\mathcal{FP}_{ij} \stackrel{\text{def}}{=} \mathcal{CNOT}_{ij}; \mathcal{CNOT}_{ji}; \mathcal{CNOT}_{ij}.$$

We introduce an abbreviation $\mathcal{CS}_\tau(t_\tau)$ for the formula saying that a state t_τ is a classical state:

$$\mathcal{CS}_\tau(t_\tau) \stackrel{\text{def}}{=} \exists\{x_i \,|\, i \in \tau\}\Big(t_\tau \doteq \otimes_{i\in\tau} x_i \wedge \bigwedge_{i\in\tau}(x_i \doteq 0_i \vee x_i \doteq 1_i)\Big),$$

where $\exists\{x_i \,|\, i \in \tau\}$ means a sequence of existential quantifiers on state variables of type $i \in \tau$. Similarly, we introduce an abbreviation $\mathcal{U}\backslash\rangle\sqcup(\alpha_{\tau\to\tau})$ for the formula saying that the variable $\alpha_{\tau\to\tau}$ denotes (an action induced by) a unitary operator on a τ-system:

$$\mathcal{U}\backslash\rangle\sqcup(\alpha_{\tau\to\tau}) \stackrel{\text{def}}{=} \forall x_\tau(\alpha_{\tau\to\tau}; \alpha^\dagger_{\tau\to\tau}(x_\tau) \doteq x_\tau).$$

Next, we write $H^{\otimes\tau}$ for $\otimes_{i\in\tau}H_i$ and $I^{\otimes\tau}$ for $\otimes_{i\in\tau}I_i$. Finally, we recursively introduce an abbreviation $\alpha^n_{\tau\to\tau}$ for the action obtained by iterating the action $\alpha_{\tau\to\tau}$ for n times:

$$\begin{aligned}\alpha^0_{\tau\to\tau} &= I^{\otimes\tau} \text{ (the identity map on } \tau\text{-system)};\\ \alpha^{n+1}_{\tau\to\tau} &= \alpha^n_{\tau\to\tau};\alpha_{\tau\to\tau} \text{ (for } n\geq 1).\end{aligned}$$

4.1 Quantum teleportation

In quantum teleportation, Alice and Bob, who are separated by a long distance, share a pair of qubits in Bell state $\frac{1}{\sqrt{2}}(|0\rangle_2|0\rangle_3+|1\rangle_2|1\rangle_3)$ (qubit 2 being with Alice, and 3 being with Bob). Alice would like to let Bob have a qubit whose state is the same as the state q of her qubit 1 (which we represent as a state variable of type $\{1\}$). She first interacts the qubit with her end of the Bell state. Define

$$\mathcal{PRE}(q)\stackrel{\text{def}}{=}(\mathcal{CNOT}_{12};(H_1\otimes I_2))\otimes I_3\Big(q\otimes(\mathcal{CNOT}_{23};(H_2\otimes I_3)(0_2\otimes 0_3))\Big).$$

She then measures her qubits 1 and 2, and depending on the result sends Bob instructions as to any further operation that must be performed on his qubit 3.

The *standard frame for Teleportation* is the frame that interprets as intended all the constants occurring in the Teleportation protocol: the constants 0_i and 1_i for each $i\in\{1,2,3\}$ as well as I_2, I_3, H_1, H_2 $\mathcal{CNOT}_{12}$, $\mathcal{CNOT}_{23}$ and $\mathcal{FP}_{13}$.

The correctness of the Teleportation protocol is equivalent to the validity in its standard frame of the formula

$$\begin{aligned}&\forall q[(q\otimes 0_2\otimes 0_3)\doteq(0_1?\otimes 0_2?\otimes I_3);(\mathcal{FP}_{13}\otimes I_2)(\mathcal{PRE}(q))\\ &\quad\wedge(q\otimes 1_2\otimes 0_3)\doteq(0_1?\otimes 1_2?\otimes I_3);(I_1\otimes I_2\otimes X_3);(\mathcal{FP}_{13}\otimes I_2)(\mathcal{PRE}(q))\\ &\quad\wedge(q\otimes 0_2\otimes 1_3)\doteq(1_1?\otimes 0_2?\otimes I_3);(I_1\otimes I_2\otimes Z_3);(\mathcal{FP}_{13}\otimes I_2)(\mathcal{PRE}(q))\\ &\quad\wedge(q\otimes 1_2\otimes 1_3)\doteq(1_1?\otimes 1_2?\otimes I_3);(I_1\otimes I_2\otimes(X_3;Z_3));(\mathcal{FP}_{13}\otimes I_2)(\mathcal{PRE}(q))].\end{aligned}$$

4.2 Quantum search algorithm

In the search problem, we are given a unitary operator O, which is usually called an *oracle*, acting on $N+1$ qubits (we assume them to be indexed by elements in $\mathbf{N}+\mathbf{1}$), such that there is a classical state $|f_0\rangle$ with the property that, for each classical state $|f\rangle$ and $b\in\mathbf{2}$,

$$O(|f\rangle\otimes|b\rangle_N)=\begin{cases}|f\rangle\otimes|1-b\rangle_N, & \text{if } f=f_0,\\ |f\rangle\otimes|b\rangle_N, & \text{if } f\in{}^{\mathbf{N}}\mathbf{2}\setminus\{f_0\}\end{cases}\tag{1}$$

The aim of the algorithm is to find out the classical state $|f_0\rangle$.

To formalize the correctness of this algorithm, we use an action variable O of type $\mathbf{N}+\mathbf{1} \to \mathbf{N}+\mathbf{1}$ to denote the oracle. Moreover, we assume that we have an action constant $PS_{\mathbf{N}}$ of type $\mathbf{N} \to \mathbf{N}$ for the action induced by the conditional phase shift gate on the first N qubits, whose matrix under the fixed basis is the following:

$$\begin{bmatrix} \mathbf{Z} & \mathbf{O}_{2\times(N-2)} \\ \mathbf{O}_{(N-2)\times 2} & -\mathbf{I}_{(N-2)\times(N-2)} \end{bmatrix}.$$

Here $\mathbf{O}_{2\times(N-2)}$ is the 2 by $N-2$ matrix of only 0 entries, and similarly for $\mathbf{O}_{(N-2)\times 2}$, and $\mathbf{I}_{(N-2)\times(N-2)}$ is the $N-2$ by $N-2$ identity matrix.

As before, the *standard frame* for the $(N+1)$-qubit quantum search algorithm is the one that interprets as intended all the above constants, as well as all the constants 0_i and 1_i. For convenience, we make the following abbreviation

$$\begin{aligned} &Oracle_{\mathbf{N+1}}(O) \stackrel{\text{def}}{=} \mathcal{U}\backslash\rangle\sqcup(O) \wedge \exists x_{\mathbf{N}}\Big[\mathcal{CS}_{\mathbf{N}}(x_{\mathbf{N}}) \wedge \forall y_{\mathbf{N}}\Big(\mathcal{CS}_{\mathbf{N}}(y_{\mathbf{N}}) \\ &\to (x_{\mathbf{N}} \doteq y_{\mathbf{N}} \to O(y_{\mathbf{N}}\otimes 0_{N+1}) \doteq y_{\mathbf{N}}\otimes 1_{N+1} \wedge O(y_{\mathbf{N}}\otimes 1_{N+1}) \doteq y_{\mathbf{N}}\otimes 0_{N+1}) \\ &\wedge (x_{\mathbf{N}} \perp y_{\mathbf{N}} \to O(y_{\mathbf{N}}\otimes 0_{N+1}) \doteq y_{\mathbf{N}}\otimes 0_{N+1} \wedge O(y_{\mathbf{N}}\otimes 1_{N+1}) \doteq y_{\mathbf{N}}\otimes 1_{N+1})\Big)\Big] \end{aligned}$$

for the formula saying that O is an action induced by an oracle acting on the $(\mathbf{N}+\mathbf{1})$-system satisfying Eq.(1).

The correctness of $(N+1)$-qubit Quantum Search Algorithm (with $N>2$) is equivalent to the validity in its standard frame of the following formula:

$$\begin{aligned} &\forall O \forall x_{\mathbf{N}}\Big\{Oracle_{\mathbf{N+1}}(O) \wedge \mathcal{CS}_{\mathbf{N}}(x_{\mathbf{N}}) \wedge O(x_{\mathbf{N}}\otimes 0_N) \doteq x_{\mathbf{N}}\otimes 1_N \wedge O(x_{\mathbf{N}}\otimes 1_N) \doteq \\ &\quad x_{\mathbf{N}}\otimes 0_N \to (x_{\mathbf{N}}\otimes H_N(1_N))^{>0.5}\Big(H^{\otimes(\mathbf{N+1})}; (O; ((H^{\otimes\mathbf{N}}; PS_{\mathbf{N}}; H^{\otimes\mathbf{N}})\otimes I_N))^K(0_{\mathbf{N}}\otimes 1_N)\Big)\Big\}, \end{aligned}$$

where K is the largest natural number less than $\frac{\pi}{4}\sqrt{2^N}$.

4.3 Deutsch-Josza algorithm

In the Deutsch-Josza problem, we are given a unitary operator O (usually called an oracle) acting on $N+1$ qubits (we assume them to be indexed by elements in $\mathbf{N}+\mathbf{1}$), which is known to satisfy one of the following properties:

(i) The oracle is *constant* (having the same value for all inputs): there is $i \in \{0,1\}$ s.t. $O(|f\rangle\otimes|b\rangle_N) = |f\rangle\otimes|b\oplus i\rangle_N$ for all $b \in \mathbf{2}$ and classical state $|f\rangle$, with $f \in {}^{\mathbf{N}}\mathbf{2}$;

(ii) The oracle is *balanced* (equal to 1 for exactly half of all the possible inputs, and 0 for the other half): there is $X \subseteq {}^{\mathbf{N}}\mathbf{2}$ s.t. $|X| = 2^{N-1}$ and $O(|f\rangle\otimes|b\rangle_N)$ is $|f\rangle\otimes|1-b\rangle_N$ if $f \in X$, and is $|f\rangle\otimes|b\rangle_N$, otherwise, for all $b \in \mathbf{2}$.

The goal of the algorithm is to determine which of the two properties holds for O.

To formalize the correctness of this algorithm, we use an action variable O of type $\mathbf{N+1} \to \mathbf{N+1}$ to denote the oracle. For convenience, we introduce some abbreviations: first, let us denote by $ConOra(O)$ the formula

$$\mathcal{U}\backslash\rangle\sqcup(O) \wedge \Big[\forall x_{\mathbf{N}}\Big(\mathcal{CS}_{\mathbf{N}}(x_{\mathbf{N}}) \to O(x_{\mathbf{N}}\otimes 0_{N+1}) \doteq x_{\mathbf{N}}\otimes 0_{N+1} \wedge O(x_{\mathbf{N}}\otimes 1_{N+1}) \doteq x_{\mathbf{N}}\otimes 1_{N+1}\Big) \\ \vee \forall x_{\mathbf{N}}\Big(\mathcal{CS}_{\mathbf{N}}(x_{\mathbf{N}}) \to O(x_{\mathbf{N}}\otimes 0_{N+1}) \doteq x_{\mathbf{N}}\otimes 1_{N+1} \wedge O(x_{\mathbf{N}}\otimes 1_{N+1}) \doteq x_{\mathbf{N}}\otimes 0_{N+1}\Big)\Big]$$

saying that O is an action induced by a constant oracle; second, we denote by $BalOra(O)$ the formula (where $k = 2^{N-1}$)

$$\mathcal{U}\backslash\rangle\sqcup(O) \wedge \exists x^1_{\mathbf{N}} \ldots \exists x^k_{\mathbf{N}}\Big[\Big(\bigwedge_{i=1}^{k} \mathcal{CS}_{\mathbf{N}}(x^i_{\mathbf{N}})\Big) \wedge \Big(\bigwedge_{1\le i<j\le k} x^i_{\mathbf{N}} \perp x^j_{\mathbf{N}}\Big) \wedge \forall y_{\mathbf{N}}\Big(\mathcal{CS}_{\mathbf{N}}(y_{\mathbf{N}}) \to \\ \Big(\bigvee_{i=1}^{k} y_{\mathbf{N}} \doteq x^i_{\mathbf{N}} \to O(y_{\mathbf{N}}\otimes 0_{N+1}) \doteq y_{\mathbf{N}}\otimes 1_{N+1} \wedge O(y_{\mathbf{N}}\otimes 1_{N+1}) \doteq y_{\mathbf{N}}\otimes 0_{N+1}\Big) \\ \wedge \Big(\bigwedge_{i=1}^{k} y_{\mathbf{N}} \perp x^i_{\mathbf{N}} \to O(y_{\mathbf{N}}\otimes 0_{N+1}) \doteq y_{\mathbf{N}}\otimes 0_{N+1} \wedge O(y_{\mathbf{N}}\otimes 1_{N+1}) \doteq y_{\mathbf{N}}\otimes 1_{N+1})\Big)\Big)\Big]$$

saying that O is an action induced by a balanced oracle.

Finally, the *correctness of the $(N+1)$-qubit Deutsch-Jozsa algorithm* (for any natural number N) is equivalent to the assertion that the following formula is valid in its standard frame:

$$\forall O\Big\{ConOra(O) \vee BalOra(O) \to \\ \Big[\ \Big(ConOra(O) \leftrightarrow H^{\otimes(\mathbf{N+1})}; O; H^{\otimes(\mathbf{N+1})}(0_{\mathbf{N}} \otimes 1_N) \doteq 0_{\mathbf{N}} \otimes 1_N\Big) \\ \wedge \Big(BalOra(O) \leftrightarrow H^{\otimes(\mathbf{N+1})}; O; H^{\otimes(\mathbf{N+1})}(0_{\mathbf{N}} \otimes 1_N) \perp 0_{\mathbf{N}} \otimes 1_N\Big)\ \Big]\Big\}.$$

5 Decidability

The set of validities of **QPDSOL** on any *given frame* is decidable. Using the same proof strategy, the validity problem for *pure* formulas over (the class of) *all frames* is also decidable. In this section, we sketch the proofs of these results.

The basic technique for proving these decidability results is a generalization and extension of the method used in (Dunn et al. 2005): We express validity of formulas of $\mathcal{L}$ without free variables in a given frame $\mathfrak{F}$ via truth of first-order sentences of

$(\mathbb{R}, +, \cdot, 0, 1)$; then the decidability of our logic follows from Tarski's theorem in (Tarski 1948) which states that the first-order theory of $(\mathbb{R}, +, \cdot, 0, 1)$ is decidable. This idea is unfolded into several technical steps.

In the first step, we need to deal with intersection of testable properties. For a function symbol of the form $(P_\tau \cap Q_\tau)?$, it is well known that calculating the matrix of the corresponding projector typically involves a process of taking limits and hence can not be expressed in the first-order theory of $(\mathbb{R}, +, \cdot, 0, 1)$. The key to solving this is the observation that complex predicate symbols, i.e. those built with $\cap, \otimes, \sim$ and other operations, can be recursively eliminated from our language with the help of quantifiers (over states). Let $\mathcal{L}^*$ be the result of this translation. Its formulas consist of those built as follows (where constraints on the types are those given in Definition 3.1 and 3.2, but with the additional requirement that for each singleton $\tau = \{i\}$, there exists a constant 0_τ that denotes $\widehat{|0\rangle}_i$, so as to facilitate the translation of generalized projectors):

$$\begin{aligned} t_\tau &::= x_\tau \mid c_\tau \mid x_{\tau_1} \otimes x_{\tau_2} \mid \alpha_{\rho\to\tau}(x_\rho) \\ P_\tau &::= p_\tau \mid T_\tau \\ \alpha_{\tau\to\rho} &::= a_{\tau\to\rho} \mid C_{\tau\to\rho} \mid a^\dagger_{\rho\to\tau} \mid a_{\tau_1\to\rho_1} \otimes a'_{\tau_2\to\rho_2} \mid P_\tau? \\ \varphi &::= x_\tau^{\prec r}(t_\tau) \mid x_\tau^{=r}(t_\tau) \mid \neg\varphi \mid \varphi \wedge \varphi \mid \forall x_\tau \varphi \mid \forall p_\tau \varphi \mid \forall a_{\rho\to\tau}\varphi. \end{aligned}$$

With the possible exception of the constants 0_τ, we have that $\mathcal{L}^* \subseteq \mathcal{L}$, and the semantics of $\mathcal{L}^*$ is the same as for $\mathcal{L}$. One can define a function $\nabla : \mathcal{L} \to \mathcal{L}^*$ by recursion (and hence it is computable) s.t. $\mathfrak{M} \models \varphi \Leftrightarrow \mathfrak{M} \models \nabla(\varphi)$ for every model $\mathfrak{M}$. To illustrate why this is the case and how it helps to solve the problem, we exhibit one case in its definition:

$$\begin{aligned} \nabla(x_\tau^{=r}((P_\tau \cap Q_\tau)?(t_\tau))) = \exists y_\tau \exists z_\tau [\ & \nabla(t_\tau \doteq y_\tau \oplus z_\tau) \wedge \nabla(P_\tau^{=1}(y_\tau)) \wedge \nabla(Q_\tau^{=1}(y_\tau)) \\ & \wedge x_\tau^{=r}(y_\tau) \wedge \forall u_\tau(\nabla(P_\tau^{=1}(u_\tau)) \wedge \nabla(Q_\tau^{=1}(u_\tau)) \to z_\tau^{=0}(u_\tau))\ ] \end{aligned}$$

where x_τ is a state variable, t_τ is a term and $t_\tau \doteq y_\tau \oplus z_\tau$ is defined to be $\forall v_\tau(v_\tau^{=0}(y_\tau) \wedge v_\tau^{=0}(z_\tau) \to v_\tau^{=0}(t_\tau))$, which means that t_τ "lies on the plane generated by" y_τ and z_τ.

In the second step, we define for each frame $\mathfrak{F}$, a function $\mathcal{TR}_\mathfrak{F} : \mathcal{L}^* \to \mathcal{L}_\mathbb{C}$, where $\mathcal{L}_\mathbb{C}$ is the first-order language of $(\mathbb{C}, +, \cdot, \bar{\cdot}, \prec, \mathbb{C})$, where $\bar{\cdot}$ is the conjugate operator, $\prec$ is a binary relation between complex numbers such that $a + bi \prec c + di$ iff $a < c$, and the last component $\mathbb{C}$ is the set of numbers named by a constant. Towards this aim, we first formalize in $\mathcal{L}_\mathbb{C}$ the matrix representation of the interpretation in $\mathfrak{F}$ of terms, predicate symbols and function symbols. This is possible because every term, predicate symbol and function symbol involves only finitely many qubits indicated by its type. In fact, one can define by recursion a computable function $\mathfrak{F}^\natural$ from the set of terms, predicate

symbols and function symbols that can occur in formulas in $\mathcal{L}^*$ to the set of finite sets of terms in $\mathcal{L}_{\mathbb{C}}$. For the base case, we define $\mathfrak{F}^\sharp(x_\tau)$, $\mathfrak{F}^\sharp(p_\tau)$ and $\mathfrak{F}^\sharp(a_{\tau\to\rho})$ to be the sets of variables indexed by ${}^\tau\mathbf{2}$, ${}^\tau\mathbf{2}\times^\tau\mathbf{2}$ and ${}^\rho\mathbf{2}\times^\tau\mathbf{2}$ in such a way that different state, predicate or action variables are mapped to disjoint sets of variables. Moreover, $\mathfrak{F}^\sharp(c_\tau)$, $\mathfrak{F}^\sharp(T_\tau)$ and $\mathfrak{F}^\sharp(C_{\tau\to\rho})$ are indexed in a similar way but they are sets of constants. Care must be taken to ensure that the constants are defined according to the interpretation in $\mathfrak{F}$. For complex symbols built with operations, we can mimic the manipulation of vectors and matrices. For example, assume that we have defined $\mathfrak{F}^\natural(x_\tau)$ to be the set of variables $\{\mathsf{x}[f] \mid f \in^\tau \mathbf{2}\}$ and $\mathfrak{F}^\natural(y_\rho)$ to be $\{\mathsf{y}[g] \mid g \in^\rho \mathbf{2}\}$ respectively, then we can mimic the Kronecker product of matrices and define $\mathfrak{F}^\natural(x_\tau \otimes y_\rho)$ to be the set of terms $\{\mathsf{x}\otimes\mathsf{y}[h] \mid h \in^{\tau\cup\rho}\mathbf{2}\}$ s.t. $\mathsf{x}\otimes\mathsf{y}[h] = \mathsf{x}[h\restriction\tau]\cdot_{\mathbb{C}}\mathsf{y}[h\restriction\rho]$, where $\cdot_{\mathbb{C}}$ is the symbol for multiplication in $\mathcal{L}_{\mathbb{C}}$. Using the function $\mathfrak{F}^\natural$, we proceed to define $\mathcal{TR}_{\mathfrak{F}}$ in such a way that given a model $\mathfrak{M}$ on the frame $\mathfrak{F}$, $\mathfrak{M}\models\varphi$ iff $(\mathbb{C},+,\cdot,\bar{\ },\prec,\mathbb{C})\models_{\mathfrak{M}}\mathcal{TR}_{\mathfrak{F}}(\varphi)$, for every $\varphi\in\mathcal{L}^*$. Here the subscript in "$\models_{\mathfrak{M}}$" is an interpretation (added to the structure $(\mathbb{C},+,\cdot,\bar{\ },\prec,\mathbb{C})$) of the free variables in $\mathcal{TR}_{\mathfrak{F}}(\varphi)$ according to the model $\mathfrak{M}$. In defining $\mathcal{TR}_{\mathfrak{F}}$ as such, care is taken in order to verify that quantification over (finitely many) variables in $\mathfrak{F}^\sharp(x_\tau)$, $\mathfrak{F}^\sharp(p_\tau)$ or $\mathfrak{F}^\sharp(a_{\tau\to\rho})$ in the input formula really corresponds to quantification of x_τ, p_τ or $a_{\tau\to\rho}$ in the translated formula.

In the third step, we focus on the behaviour of $\mathcal{TR}_{\mathfrak{F}}$ on the set of closed formulas. Since the definition of frames ensures that the matrix representation of the interpretation of constant symbols only has simple complex numbers as entries, the translation $\mathcal{TR}_{\mathfrak{F}}(\varphi)$ of a closed formula φ of $\mathcal{L}^*$ in a given frame $\mathfrak{F}$ is actually a first-order sentence of $(\mathbb{C},+,\cdot,\bar{\ },\prec,\mathcal{S})$, where $\mathcal{S}$ is the set of simple complex numbers (see page 46). A consequence of this is that pure formulas of $\mathcal{L}$ are translated via $\mathcal{TR}_{\mathfrak{F}}$ into first-order sentences of $(\mathbb{C},+,\cdot,\bar{\ },\prec)$, because there are no constants in a pure formula. Therefore, by Lemma 1 and the property of $\mathcal{TR}_{\mathfrak{F}}$ by definition, we know that on a given frame $\mathfrak{F}$, $\mathfrak{F}\models\varphi$ iff $(\mathbb{C},+,\cdot,\bar{\ },\prec,\mathcal{S})\models\mathcal{TR}_{\mathfrak{F}}\circ\nabla(\varphi)$, for every closed formula $\varphi\in\mathcal{L}$.

The final step is to reduce the first-order theory of $(\mathbb{C},+,\cdot,\bar{\ },\prec,\mathcal{S})$ to the first-order theory of the reals. This is a simple translation, where each simple complex number is mapped to a pair of definable real numbers, and addition and multiplication are mapped according to complex arithmetic. Thus the decidability of our logic follows from these reductions and Tarksi's theorem. To summarize, we have the following decidability result.

Theorem 1. *The set $\{\varphi\in\mathcal{L} \mid \varphi$ is closed and $\mathfrak{F}\models\varphi\}$ is decidable, for any given frame $\mathfrak{F}$. Moreover, the set $\{\varphi\in\mathcal{L} \mid \varphi$ is pure and $\models\varphi\}$ is decidable.*

6 Conclusions

This paper extends decidability results from (Dunn et al. 2005) and (Baltag et al. 2012) to a language that is much more versatile in its ability to express quantum algorithms and their correctness. Our techniques can be applied to a wider range of quantum logics, giving a general recipe for showing decidability as long as definability of the sentences and operators can be done along the lines presented in this paper. In addition we have described how to express the correctness of Quantum Teleportation, the Quantum Search algorithm and the Deutsch-Josza algorithm; however this is not an exhaustive list of algorithms whose correctness can be expressed in our language. The Fourier transform can easily be expressed in our language and this may lead to a wealth of further examples, notably those involving the hidden subgroup problem, such as order-finding and factoring; however we leave these for future work. Other future tasks involve finding a complete axiomatization and determining the complexity of the decision procedure.

Acknowledgements The research of J. Bergfeld, K. Kishida and J. Sack has been funded by VIDI grant 639.072.904 of the NWO. The research of S. Smets is funded by the VIDI grant 639.072.904 of the NWO and by the FP7/2007-2013/ERC Grant agreement no. 283963. The research of S. Zhong has been funded by China Scholarship Council.

References

S. Abramsky and B. Coecke. A categorical semantics of quantum protocols. In *Proceedings of the 19th IEEE conference on Logic in Computer Science (LiCS'04)*, pages 415–425. IEEE Press, 2004.

D. Aerts. Description of compound physical systems and logical interaction of physical systems. In E. Beltrametti and B. van Fraassen, editors, *Current Issues on Quantum Logic*, pages 381–405. Kluwer Academic, 1981.

A. Baltag and S. Smets. Complete Axiomatizations for Quantum Actions. *International Journal of Theoretical Physics*, 44(12):2267–2282, 2005.

A. Baltag and S. Smets. LQP: The Dynamic Logic of Quantum Information. *Mathematical Structures in Computer Science*, 16(3):491–525, 2006.

A. Baltag, J. Bergfeld, K. Kishida, S. Smets, and S. Zhong. A Decidable Dynamic Logic for Quantum Reasoning. *EPTCS*, in print, 2012.

A. Baltag, J. Bergfeld, K. Kishida, J. Sack, S. Smets, and S. Zhong. PLQP & company: Decidable logics for quantum algorithms. Submitted to the International Journal of Theoretical Physics, 2013.

G. Birkhoff and J. von Neumann. The Logic of Quantum Mechanics. *The Annals of Mathematics*, 37:823–843, 1936.

R. Chadha, P. Mateus, A. Sernadas, and C. Sernadas. Extending classical logic for reasoning about quantum systems. In K. Engesser, D. M. Gabbay, and D. Lehmann, editors, *Handbook of Quantum Logic and Quantum Structures: Quantum Logic*, pages 325–371. Elsevier, 2009.

M. L. Dalla Chiara, R. Giuntini, and R. Greechie. *Reasoning in quantum theory: sharp and unsharp quantum logics*, volume 22 of *Trends in logic*. Kluwer Acadamic Press, Dordrecht, 2004.

J. M. Dunn, T. J. Hagge, L. S. Moss, and Z. Wang. Quantum Logic as Motivated by Quantum Computing. *The Journal of Symbolic Logic*, 70(2):353–359, 2005.

L. Henkin. Completeness in the Theory of Types. *The Journal of Symbolic Logic*, 15: 81–91, 1950.

M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2011.

C. Piron. *Foundations of Quantum Physics*. W.A. Benjamin Inc., 1976.

M. Rabin. Decidability of second order theories and automata on infinite trees. *Transactions of the American Mathematical Society*, pages 1–35, 1969.

C. Randall and D. Foulis. Tensor products of quantum logics do not exist. *Notices Amer. Math. Soc.*, 26(6), 1979.

P. Selinger. Towards a quantum programming language. *Mathematical Structures in Computer Science*, 14:527–586, 7 2004.

A. Tarski. *A Decision Method for Elementary Algebra and Geometry*. RAND Corporation, Santa Monica, California, 1948.

The Logic of Evidence-Based Knowledge

Chenwei Shi

Department of Philosophy, Tsinghua University
shichenwei88@gmail.com

Abstract

This paper presents a logic for reasoning about evidence and knowledge. Following van Benthem and Pacuit's semantic approach to evidence, we make use of Nozick's tracking theory to define the key concept *sound enough evidence*, based on which we then provide an account of evidence-based knowledge. In our framework the newly defined knowledge does not imply the evidence-based belief but respects the epistemic closure. The related philosophical issues are also discussed.

1 Introduction

Example 1 (The Witness of The Murder)**.** One day a man was murdered in a room. The detective Sherlock took charge of this case. Later a man came to him and said that he had witnessed the whole murder and recognized the murderer (Killer) because the window of that room was opposite to that of his own room and at that time he was looking out of the window

Scenario 1 without any instruments;

Scenario 2 with a telescope;

and happened to see the murder. With the help of polygraph, Sherlock was certain about the honesty of the witness. But he still could not make the judgement that the murderer was definitely the man identified by the witness. (To be continued.)

The explanation for Sherlock's caution is simple: He can only be sure that the witness has the evidence to make him *believe* that Killer is the murderer but can not be sure whether the evidence the witness has is *sound enough* to make him *know* that Killer is the murderer.

In the above explanation, three notions, "evidence", "belief" and "knowledge", play a crucial role. To understand it better, a further analysis of the relation between these concepts is needed. Furthermore, we would like to answer the following question: How does Sherlock make sure that the witness's evidence is sound enough?

The studies concerning those three concepts can be found in many logical and philosophical literature (e.g. Dubois and Prade 1992, Baltag et al. 2012; 2013, Williamson 1997, Moser 1991). Different from all of them, in this paper we will follow the analysis of the relationship between evidence and belief in (van Benthem and Pacuit 2011), in which evidence structure was introduced to standard doxastic models and a dynamic logic of evidence-based belief was proposed. We attempt to analyse the relationship between evidence and knowledge and propose a new way of defining knowledge based on evidence.

Then the main issue comes down to the problem how to find the right evidences which underlie an agent's knowledge. The observation is $E_K \subseteq E_B$, where E_B and E_K represent the set of propositions directly supported by the agent's evidences from which she can derive her belief and knowledge respectively. In other words, while the propositions in E_B are only the propositions we have evidence for, the propositions in E_K are the propositions we have sound enough evidence for. The definition of knowledge in (Cornelisse 2011) suggests that the relation between E_B and E_K should be $E_K = E_T \subseteq E_B$, where E_T denotes the set of true propositions in E_B. Different from (Cornelisse 2011), however, we claim that the relation should be $E_K \subseteq E_T \subseteq E_B$, for which we will argue philosophically and technically in this paper. Therefore it is pivotal to find a mechanism which can pick the propositions in E_K from the propositions in E_B to define knowledge.

This paper presents a way in which E_K can be picked out from E_B, inspired by Holliday (2012)'s formalization of Nozick (1981)'s tracking theory, and then defines knowledge based on evidences. After the introduction of the logic of evidence-based belief (van Benthem and Pacuit 2011) in Section 2 and the couternfactual belief model, (Holliday 2012) in Section 3, Section 4 proposes a new logic, the logic of sound enough evidence, in which we can express the concept "sound enough evidence for" and discuss the relation between it and another concept "evidence for". Section 5 extends the logic of sound enough evidence with a new operator K^E (knowledge based on evidences) to form the logic of evidence-based knowledge, and analyses the corresponding logical and philosophical issues. Finally, some further research directions are listed.

2 Logic of evidence-based belief

This section mainly introduces the logic of evidence-based belief (EBL) proposed by van Benthem and Pacuit (2011) and discusses the intuition behind the semantic truth definitions. The reader is referred to (van Benthem et al. 2012) for the axiomatization of this logic.

Definition 2.1 (Evidence and belief language)**.** Let **At** be a set of atomic propositions. The evidence and belief language $\mathcal{L}$ is defined by

$$p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B\varphi \mid \Box\varphi \mid A\varphi$$

where $p \in \mathbf{At}$. The definition of $\vee, \rightarrow$, and $\leftrightarrow$ is as usual in terms of $\neg$ and $\wedge$.

The intended reading of $\Box\varphi$ is "the agent has evidence that implies φ(the agent has "evidence for" φ). And $B\varphi$ says that "the agent believes that φ is true". The universal modality ($A\varphi$: "φ is true in all states") is included for technical convenience.

The evidences accepted by the agent are not necessarily jointly consistent, although every evidence itself should be consistent. In Example 1, Sherlock takes the witness's testimony into account and at the same time he can also take other evidences of Killer's absence into consideration, but he will never accept the evidence which is itself contradictory. Therefore the "evidence for" operator is not a normal modal operator and neighbourhood models are the natural choice for the semantics of EBL.

Definition 2.2 (Evidence models)**.** An **evidence model** is a tuple $\mathcal{M} = \langle W, E, V\rangle$ with W a non-empty set of worlds, $E \subseteq W \times \wp(W)$ an evidence relation, and $V : \mathbf{At} \rightarrow \wp(W)$ a valuation function. $E(w)$ is written for the set $\{X \mid wEX\}$, capturing the evidences possessed by the agent in w. Two constraints are imposed on the evidence sets:

- For each $w \in W, \emptyset \notin E(w)$ (evidence per se is never contradictory);
- For each $w \in W, W \in E(w)$ (agents know their space).

Note that $E(w)$ is not assumed to be closed under supersets and disjoint evidence sets are allowed for, whose combination may lead to trouble. Thus to derive the belief operator from the evidence model, the following definition is needed:

Definition 2.3. A w-**scenario** is a maximal collection $C \subseteq E(w)$ that has the fip (i.e., the finite intersection property: for each finite subfamily $\{X_1, \ldots, X_n\} \subseteq C, \bigcap_{1\leq i\leq n} X_i \neq \emptyset$). A collection is called a **scenario** if it is a w-scenario for some state w.

Truth of formulas in $\mathcal{L}$ is defined as follows:

Definition 2.4 (Truth conditions)**.** Given a model $\mathcal{M} = \langle W, E, V\rangle$ with W be an evidence model. Truth of a formula $\varphi \in \mathcal{L}$ is defined inductively as follows:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$ $\quad (p \in \mathbf{At})$
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models \Box\varphi$ iff there is an X with wEX and for all $v \in X$, $\mathcal{M}, v \models \varphi$
- $\mathcal{M}, w \models B\varphi$ iff for each w-scenario $C \subseteq E(w)$ and for all $v \in \bigcap C$, $\mathcal{M}, v \models \varphi$
- $\mathcal{M}, w \models A\varphi$ iff for all $v \in W$, $\mathcal{M}, w \models \varphi$.

The "$B\varphi$" item reflects the mechanism of the agent's believing a proposition φ, that is considering all the maximally consistent theories based on the evidences collected at w to see whether the proposition φ is true in all these theories, where the maximally consistent theory is represented by the w-scenario.

Now that the epistemic operator belief (B) can be derived from the evidence structure, then how about knowledge (K)? Because the propositions the agent has evidence for at certain possible state can not only be true but also be false, the "evidence for"($\Box$) operator represents only the evidences based on which the agent can get belief but not knowledge. For this reason it is necessary to find the set E_K (the propositions we have sound enough evidence for) based on which the agent get her knowledge. Although it is obvious that $E_K \subseteq E_B$ as mentioned in the Introduction, it is not obvious how to pick them from E_B. Before the way of picking is presented in Section 4, the origin of this idea should be first introduced.

3 The counterfactual belief (CB) model

This section introduces Holliday (2012)'s formalization[1] of Nozick (1981)'s tracking theory which argued that sensitivity and adherence – the conjunction of which is tracking – are necessary and sufficient for one's belief to constitute knowledge, where the sensitivity and adherence are as follows:

- if φ were false, the agent would not believe φ (sensitivity);
- if φ were true, the agent would believe φ (adherence)

[1] cf. the Chapter 3 of (Nozick 1981) for the philosophical discussion.

Definition 3.1 (CB model)**.** A **counterfactual belief model** is a tuple $\mathcal{M} = \langle W, D, \leqslant, V\rangle$, where

1. w is a non-empty set;
2. D is a serial binary relation on w;
3. $\leqslant$ assigns to each $w \in W$ a binary relation $\leqslant_w$ on some $W_w \subseteq W$;
   (a) $\leqslant_w$ is reflexive and transitive;
   (b) for all $v \in W_w$, $w \leqslant_w v$;
4. V assigns to each $p \in \mathbf{At}$ a set $V(p) \subseteq W$.

D can be thought of as a *doxastic accessibility*, so that a belief operator B can be defined in this structure as follows:

- $\mathcal{M}, w \models B\varphi$ iff for all $v \in W$ such that wDv, $\mathcal{M}, v \models \varphi$.

$u \leqslant_w v$ can be seen as a relation of *comparative similarity* with regard to w, with which the counterfactuals can be defined as in (Lewis 1973) And Condition 3(b) means that the actual world is always a relevant alternative. In (Holliday 2012), Holliday analyses different effects of other restrictions on the validities and assumes that $\leqslant_w$ is well-founded which does not affect the results of (Holliday 2012):

- $\leqslant_w$ is **well-founded** iff for every non-empty $S \subseteq W_w$, $Min_{\leqslant_w}(S) \neq \emptyset$

where $Min_{\leqslant_w}(S) = \{v \in S \cap W_w \mid \text{there is no } u \in S \text{ such that } u <_w v\}$.

In such a setting, the truth of counterfactuals ($\boxright$) can be defined:

- $\varphi \boxright \psi$ is true at a world w iff the closest φ-worlds to w according to $\leqslant_w$ are ψ-worlds

Thus, all the conditions sufficient and necessary for knowledge can be expressed in the CB model:

Definition 3.2 (Truth conditions)**.** Given a well-founded CB model $\mathcal{M} = \langle W, D, \leqslant, V\rangle$ with $w \in W$ and φ in the epistemic-doxastic language, define $\mathcal{M}, w \models \varphi$ as follows:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$ $\qquad (p \in \mathbf{At})$
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$

- $\mathcal{M}, w \models B\varphi$ iff for all $v \in W$ such that wDv, $\mathcal{M}, v \models \varphi$
- $\mathcal{M}, w \models K\varphi$ iff $\mathcal{M}, w \models B\varphi$ and (sensitivity)$\forall v \in Min_{\leqslant_w}(\overline{[\![\varphi]\!]}) : \mathcal{M}, v \models \neg B\varphi$ and (adherence)$\forall v \in Min_{\leqslant_w}([\![\varphi]\!]) : \mathcal{M}, v \models B\varphi$

where $[\![\varphi]\!] = \{v \in W \mid \mathcal{M}, v \models \varphi\}$ and $\overline{[\![\varphi]\!]} = \{v \in W \mid v \notin [\![\varphi]\!]\}$.

The definition of knowledge here can be seen as a selection from one's belief: what kind of belief can be qualified as knowledge? And this selection is a counterfactual test of the agent's epistemic state to check whether the agent's epistemic state meets the requirement of knowledge. Recall our question at the beginning: what kind of evidences for the agent's belief can be qualified as evidences for the agent's knowledge? Then it is natural to think that this counterfactual test can be used again, not of the agent's epistemic state, but of the agent's evidential state.

4 A logic of sound enough evidence (SEL)

We first continue the story in Example 1.
To make clear whether the witness knows that the Killer is the murderer (we assume that in fact the Killer is the murderer), Sherlock took another suspect Keller who was also the acquaintance of the witness to the room where the murder happened and stimulated the crime scene including the time, position of the murderer, light and so on. And he asked Keller to stand at the position of murderer and the witness to stand in front of his window to identify who the guy at the position of murderer was

Scenario 1 without any instruments;

Scenario 2 with a telescope.

At last, the witness

Scenario 1 did not recognize who he was;

Scenario 2 recognized who he was.

Therefore Sherlock

Scenario 1 still could not conclude that Killer is the murderer;

Scenario 2 could conclude that the Killer is the murderer.

This section develops a framework of evidence logic, which can be used to explain the difference between the two scenarios in the examples.

4.1 Language of SEL

Definition 4.1 (Language of SEL)**.** Let **At** be a set of atomic sentence symbols, the language $\mathcal{L}_E$ is defined by

$\varphi_0 := p \mid \neg\varphi_0 \mid \varphi_0 \wedge \varphi_0 \mid \Box\varphi_0$

$\varphi := \varphi_0 \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \mathrel{\Box\!\!\rightarrow} \varphi$

where $p \in \mathbf{At}$. Additional connectives are defined as usual and the duals of $\varphi \mathrel{\Box\!\!\rightarrow} \psi, \Box\varphi$ are $\varphi \mathrel{\Diamond\!\!\rightarrow} \psi, \Diamond\varphi$ respectively.

$\Box\!\!\rightarrow$ is the operator for counterfactual conditionals and $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ can be read as "If it were the case that φ, then it would be the case that ψ".

In this logic, "sound evidence for" is a derived operator, defined as an abbreviation by putting $\Box_K\varphi := \varphi \wedge (\neg\varphi \mathrel{\Box\!\!\rightarrow} \Box\neg\varphi) \wedge (\varphi \mathrel{\Box\!\!\rightarrow} (\Box\varphi \wedge \neg\Box\neg\varphi))$. The interpretation of $\Box_K\varphi$ is that the agent has sound enough evidence for φ, while $\Box\varphi$ says that the agent has evidence for φ.

For simplicity in this version of SEL and some philosophical reason[2], evidence for counterfactual conditionals is not taken into consideration. It is for this reason that only φ_0 formulas can appear inside "evidence for" operator. And we write the language without operator $\Box\!\!\rightarrow$ as $\mathcal{L}_0$.

4.2 Counterfactual evidence model

Recall the last paragraph of Section 3. We want to execute the counterfactual test for the agent's evidences. To capture such test in the evidence model, it is necessary to introduce the relation $\leqslant$, by which the relevance between different possible worlds can be compared in the new introduced model:

Definition 4.2 (Truth conditions)**.** Given a model $\mathcal{M} = \langle W, E, \leqslant\rangle$ and state $w \in W$, truth of a formula $\varphi \in \mathcal{L}_E$ is defined inductively as follows:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$ $\quad (p \in \mathbf{At})$
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models \varphi \mathrel{\Box\!\!\rightarrow} \psi$ iff $\forall v \in Min_{\leqslant_w}([\![\varphi]\!]) : \mathcal{M}, w \models \psi$

[2] The counterfactual conditionals, relating to the relevance ordering of the possible worlds, seems totally different from the fact which can be known by us through empirical sense (evidence) and inference.

- $\mathcal{M}, w \models \Box\varphi$ iff there is an $X \in E(w)$ and for all $v \in X$, $\mathcal{M}, v \models \varphi$ and $\varphi \in \mathcal{L}_0$

- $\mathcal{M}, w \models \Box_K\varphi$ iff $\exists X \subseteq [\![\varphi]\!] : X \in E(w)$ & $\forall Y \subseteq \overline{[\![\varphi]\!]} : Y \notin E(w)$,
  $\forall v \in Min_{\leqslant_w}(\overline{[\![\varphi]\!]})(\exists X \subseteq \overline{[\![\varphi]\!]} : X \in E(v))$,
  $\forall v \in Min_{\leqslant_w}([\![\varphi]\!])(\exists X \subseteq [\![\varphi]\!] : X \in E(v))$ & $\forall Y \subseteq \overline{[\![\varphi]\!]} : Y \notin E(v))$

To avoid some obvious counterexamples, we revise Nozick's tracking theory[3]: the agent has sound enough evidence for φ if and if only in the actual state where φ is true, she has evidence for φ and has no evidence for $\neg\varphi$, at the same time, she would have evidence for $\neg\varphi$ in the most similar counterfactual states, and would have evidence for φ and accept no evidence for $\neg\varphi$ in the most similar possible states where φ was the case. We can also call such evidential state truth-tracking evidential state.

Now it can be explained why Sherlock made different judgements in different scenarios. The difference between the two scenarios can be shown in the CE models $\mathcal{M}_1$ and $\mathcal{M}_2$ (see Figure 1 and Figure 2) where Killer is the murderer (i) in world w, Keller is the murderer (e) in world v, Killer and Keller are both the murderer ($i \wedge e$) in world u and Killer and Keller are neither the murderer in world t:

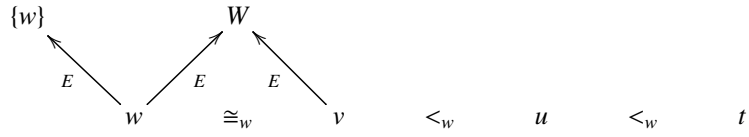


Figure 1: CE model for Scenario 1: $\mathcal{M}_1$

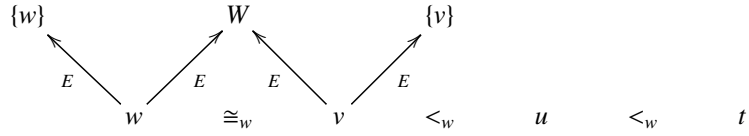


Figure 2: CE model for Scenario 2: $\mathcal{M}_2$

[3]We can also define a weak version of $\Box_K\varphi$ with only the second condition different: $\forall v \in Min_{\leqslant_w}(\overline{[\![\varphi]\!]})(\forall X \subseteq [\![\varphi]\!] : X \notin E(v))$. However, according to this weak version, the evidential state of witness in Scenario 1 also meet the requirements of knowledge, which is counterintuitive.

We do not mean here that our definition of $\Box_K\varphi$ is the unique possibility without any counterexample. Nonetheless, concerning the example in this paper and other similar situations, such a definition is acceptable. Maybe different definitions of $\Box_K\varphi$ can suit different situations so that the comparison between different definitions is necessary. But we do not touch this issue for now.

It is easy to verify that $\mathcal{M}_2, w \models \Box_K i$ but $\mathcal{M}_1, w \models \neg\Box_K i$, which means that from the perspective of Sherlock, the witness in Scenario 2 is in the truth-tracking evidential state but the witness in Scenario 1 is not. And the following fact tells us why Sherlock made different judgements in the different scenarios:

Fact 4.1. *In the class of CE models, we have the following validity:*[4]

- $\models \Box_K\varphi \to \varphi$

Proof. Let $\mathcal{M}$ be any counterfactual evidence model and w any state in $\mathcal{M}$ such that $\mathcal{M}, w \models \Box_K\varphi$. We need to show that $(\mathcal{M}, w) \models \varphi$. Suppose $\mathcal{M}, w \models \neg\varphi$ for contradiction. As $w \in Min_{\leqslant_w}(\overline{[\![\varphi]\!]})$ (according to the assumption and Definition 3.1 3(b)), it follows that $(\mathcal{M}, w) \models \Box\neg\varphi$ by the definition of $\Box_K\varphi$. At the same time, we can get $\mathcal{M}, w \models \neg\Box\neg\varphi$ by the definition of $\Box_K\varphi$. Contradiction. □

With this fact we can now prove the claim made in the Introduction:

Remark 1. *Given any CE model $\mathcal{M} = \langle W, E, \leqslant, V\rangle$ and any $w \in W$, we can represent the following three expressions: "the agent has evidence for φ", "the agent has true evidence for φ" and "the agent has sound enough evidence for φ" as $\mathcal{M}, w \models \Box\varphi, \mathcal{M}, w \models \Box\varphi \wedge \varphi$ and $\mathcal{M}, w \models \Box_K\varphi$ respectively. Thus we can denote E_B, E_T, E_K stated in Section 1 as $E_B(w) = \{\varphi \mid \mathcal{M}, w \models \Box\varphi\}, E_T(w) = \{\varphi \mid \mathcal{M}, w \models \Box\varphi \wedge \varphi\}$ and $E_K(w) = \{\varphi \mid \mathcal{M}, w \models \Box_K\varphi\}$. By the fact $\models \Box_K\varphi \to \Box\varphi \wedge \varphi$ and $\models \Box\varphi \wedge \varphi \to \Box\varphi$, we have $E_K \subseteq E_T \subseteq E_B$.*

4.3 Relationship between $\Box\varphi$ and $\Box_K\varphi$

According to the truth condition of $\Box$, $\Box\!\!\rightarrow$ and $\Box_K$, the following fact follows immediately:

Fact 4.2.

$$\models \varphi \wedge (\neg\varphi \,\Box\!\!\rightarrow\, \Box\neg\varphi) \wedge (\varphi \,\Box\!\!\rightarrow\, (\Box\varphi \wedge \neg\Box\neg\varphi)) \leftrightarrow \Box_K\varphi$$

Proof. From right to left, it is trivial, noticing the fact $\models \Box_K\varphi \to \varphi$. From left to right, noticing the fact $w \in Min_{\leqslant_w}([\![\varphi]\!])$, it is also obvious. □

The fact suggests that the two operators $\Box$ and $\Box_K$ are closely related. It seems that $\Box_K(\varphi \wedge \psi) \to \Box_K\varphi$ is a validity like the validity for $\Box$: $\Box(\varphi \wedge \psi) \to \Box\varphi$. But it is not valid in fact. Moreover, there is no any evidential closure.[5] Because the operator

[4] Without specification, in what follows all the validities we present are validities in the class of CE models.

[5] This concept is similar to the concept "epistemic closure".

$\Box_K$ describes only our evidential state, in which there is no room for inference and reasoning.

However, we can still ask: what condition on earth do we need to ensure some evidential closure for $\Box_K$? The following theorem gives the answer.[6]

Theorem 1 (Closure theorem)**.** *Let*

$$\chi_n := \varphi_0 \wedge \Box_K\varphi_1 \wedge \ldots \wedge \Box_K\varphi_n \rightarrow \Box_K\psi_1 \vee \ldots \vee \Box_K\psi_m$$

be a formula where $\varphi_1, \ldots, \varphi_n$ and $\psi_1, \ldots, \psi_m$ are all propositional formulae and φ_0 is a propositional conjunction. If

$$\models \bigwedge_{\varphi\in\Theta} \varphi \leftrightarrow \psi \Rightarrow \models (\bigwedge_{\varphi\in\Theta} \Box\varphi \leftrightarrow \Box\psi) \wedge (\bigvee_{\varphi\in\Theta} \Box\neg\varphi \leftrightarrow \Box\neg\psi) \tag{1}$$

where Θ can be any set of formulae, then χ_n is valid iff

(a) $\varphi_0 \wedge \ldots \wedge \varphi_n \rightarrow \bot$ is valid or

(b) for some $\Phi \subseteq \{\varphi_1, \ldots, \varphi_n\}$ and $\psi \in \{\psi_1, \ldots, \psi_m\}$, $\bigwedge_{\varphi\in\Phi} \varphi \leftrightarrow \psi$ is valid

Proof. The proof is similar to the proof of Theorem 5.2 in (Holiday 2013). The details can be found in Appendix A. □

4.4 Axiomatization of the evidence logic

Combining the axioms for $\Box\!\!\rightarrow$ (see Board 2004, p. 54) and the axioms for $\Box$ (see van Benthem et al. 2012), the axiom system CES for EL is the following:

taut: all propositional tautologies

$\Box\!\!\rightarrow$-1 $\varphi \Box\!\!\rightarrow \varphi$

$\Box\!\!\rightarrow$-2 $((\varphi \Box\!\!\rightarrow \psi) \wedge (\varphi \Box\!\!\rightarrow (\psi \rightarrow \chi))) \rightarrow (\varphi \Box\!\!\rightarrow \chi)$

$\Box\!\!\rightarrow$-3 $(\varphi \Box\!\!\rightarrow \psi) \rightarrow (((\varphi \wedge \psi) \Box\!\!\rightarrow \chi) \leftrightarrow (\varphi \Box\!\!\rightarrow \chi))$

$\Box\!\!\rightarrow$-4 $\neg(\varphi \Box\!\!\rightarrow \neg\psi) \rightarrow (((\varphi \wedge \psi \Box\!\!\rightarrow \chi) \leftrightarrow (\varphi \Box\!\!\rightarrow (\psi \rightarrow \chi)))$

$\Box\!\!\rightarrow$-5 $\varphi \wedge (\varphi \Box\!\!\rightarrow \psi) \rightarrow \psi$

$\top$- and non $\bot$-evidence: $\Box\top \wedge \neg\Box\bot$

[6] This theorem is inspired by Theorem 2.1 in (Holliday 2012).

$\Box$-monotonicity: $\dfrac{\varphi \to \psi}{\Box\varphi \to \Box\psi}$

MP: Modus Ponens

LE: From $\varphi \leftrightarrow \psi$ infer $(\varphi \Box\!\!\rightarrow \chi) \leftrightarrow (\psi \Box\!\!\rightarrow \chi)$

$N_{\bigcirc}$**:** Necessitation for $\bigcirc = \Box, \varphi \Box\!\!\rightarrow$

Theorem 2. *The evidence logic is sound and weakly complete for the class of counterfactual evidence models.*

Proof. The proof of weak completeness is somewhat similar to the proof system in (Board 2004, Proof of Theorem 2, p.77), but with several key differences. The first difference is that the subformulas of φ may now include formulas of the form $\Box\psi$, from which we construct maximally consistent sets (MSCs). Besides, we stipulate that $\Box\top$ and $\neg\Box\bot$ are both in the MSC. The second difference is the definition of "$\leqslant$" in the canonical model. The third difference is that we must construct a evidence relation in the canonical model for EL, which can be referred to (van Benthem et al. 2012). The detailed can be found in Appendix B. Also the proof of soundness can be found there. $\Box$

5 A logic of evidence-based knowledge

Up to now, we only talk about the agent's evidential state, and say nothing about one's epistemic state. In this section we try to define the agent's knowledge based on her evidence, or more exactly on E_K and discuss the relation between these two kind of states. We simply add operator K^E to the language $\mathcal{L}_E$ to form the new language of EKL $\mathcal{L}_{EK}$.

Definition 5.1 (Language of EKL)**.** Let **At** be a set of atomic sentence symbols, the language $\mathcal{L}_E K$ is defined by

$\varphi_0 := p \mid \neg\varphi_0 \mid \varphi_0 \wedge \varphi_0 \mid \Box\varphi_0$

$\varphi := \varphi_0 \mid \neg\varphi \mid \varphi \wedge \varphi \mid K^E\varphi_0 \mid \varphi \Box\!\!\rightarrow \varphi$

where $p \in \mathbf{At}$.

The new operator K^E is for "evidence-based knowledge" and $K^E\varphi$ can be read "the agent knows φ based on her evidence".

We can define knowledge as in Definition 3.2 (we call it CB-knowledge). But besides this definition, in the counterfactual evidence model there can be another evidence-based definition of knowledge K^E (we call it CE-knowledge).

Definition 5.2 (Truth condition for $\mathcal{L}_{EK}$ in a CE model)**.** Given a model $\mathcal{M} = \langle W, E, \leqslant, V\rangle$ and state $w \in W$, truth of a formula $\varphi \in \mathcal{L}$ is defined inductively, all the cases are the same as in Definition 4.2 except the case for operator K^E:

- $\mathcal{M}, w \models K^E\varphi$ iff $\{\psi \mid \mathcal{M}, w \models \Box_K\psi\} \models \varphi$.

Collecting all the propositions supported directly in the truth-tracking evidential state, we form our knowledge based on them (E_K) by inference and reasoning.

Fact 5.1. *In the class of CE models, $\Box_K\varphi \to K^E\varphi$ is valid but $K^E\varphi \to \Box_K\varphi$ is not valid.*

In addition, our knowledge satisfies the following properties:

Fact 5.2. *In the class of CE models, we have the following validity:*

- $\models K^E\varphi \to \varphi$
- $\models K^E\varphi \to (K^E(\varphi \to \psi) \to K^E\psi)$

Proof. We only prove the first one, as the second is trivial.

Let $\mathcal{M}$ be any counterfactual evidence model and w any state in $\mathcal{M}$ such that $\mathcal{M}, w \models K^E\varphi$. We need to show that φ also holds on w. By the definition of $K^E\varphi$, $\{\psi \mid \mathcal{M}, w \models \Box_K\psi\} \models \varphi$. By Fact 4.1, for any $\psi \in \{\psi \mid \mathcal{M}, w \models \Box_K\psi\}$, $(\mathcal{M}, w) \models \psi$. Thus $\mathcal{M}, w \models \varphi$. □

The first item of Fact 5.2 is of no doubt a required property of knowledge. But the second one seems controversial. We will discuss this issue in the next subsection.

As for the axiomatization and completeness of EKL for the class of counterfactual evidence models, we leave it as an open question for the time being and focus on the philosophical issues brought about by the evidence-based knowledge in this paper.

5.1 Relationship between $B, \Box_K$ and K^E

Because the CE model is in fact a simple extension of the evidence model, the operators belief (B) can be defined in it as in the evidence model. And because of the existence of $\leqslant$ in CE models, and the definability of the belief operator, the CB-knowledge can also be defined in the CE models as in the CB models. Therefore this section discusses some philosophical issue relating to relationship between these operators ($B, \Box_K, K^E$) in the CE models.

Knowledge without belief

In both epistemology and epistemic logic, the dominant opinion on the relationship between belief and knowledge is that belief is a necessary condition for knowledge. However, the definition of CE-knowledge in Section 4.1 does not follow such a tradition:

Fact 5.3. $\nvDash K^E\varphi \to B\varphi$

Proof. We construct a counterfactual evidence model $\mathcal{M} = \langle W, E, \leqslant\rangle$ where $W = \{w, v, u, t\}$, $V(p) = \{w, v\}$, $V(q) = \{u, t\}$ (see Figure 3).
It is easy to check that $(\mathcal{M}, w) \models \Box_K p$. It follows that $(\mathcal{M}, w) \models K^E p$ from the fact that

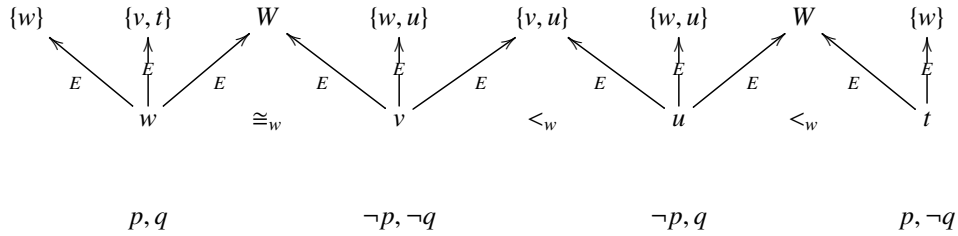


Figure 3: countermodel for $K^E\varphi \to B\varphi$

$(\mathcal{M}, w) \models \Box_K p \to K^E p$. However, we can also check that $(\mathcal{M}, w) \models \neg Bp$ according to the Definition 2.4. □

To elucidate this fact better, we make use of an example from (Lewis 1996):

> I even allow knowledge without belief, as in the case of the timid student who knows the answer but has no confidence that he has it right, and so does not believe what he knows. (p.556)

Let p be the answer to the question mentioned in the example (we denote it as A) and q be the proposition "The student has the ability to solve the problem like A". Then the model constructed in the above model characterizes the example substantially: The student in Lewis's example has an evidence for the answer ($\{w\} \in E(w)$), and has sound enough evidence for the answer ($(\mathcal{M}, w) \models \Box_K p$). Thus, the student knows the answer ($(\mathcal{M}, w) \models K^E p$). However, at the same time, he also has evidence for his inability to solve the question like A ($\{v, t\} \in E(w)$), which contradicts his another evidence ($\{w\} \in E(w)$). Therefore, he does not believe what he knows ($(\mathcal{M}, w) \models \neg Bp$).

In some sense, our definition of knowledge can be seen as one explanation of the so-called "knowledge without belief".

The epistemic closure and the gap between $\Box_K$ and K^E

As mentioned in the end of the last subsection, the second item of Fact 5.2 is controversial. Many epistemologists deny that this principle holds (Dreske 1970; 1981, Nozick 1981). Because the closure under known implication can be made use of by the sceptic to reach his conclusion about the unreliability of our knowledge. To save the epistemic closure and avoid slipping into scepticism at the same time, the logics proposed in this paper suggests to distinguish two "sources of knowledge"–*externally situated evidence* (sound enough evidence) and *internally situated reasons* (inference and reasoning) so that the reliability of knowledge is ensured by the sound enough evidence and the epistemic closure is kept because of the *internally situated reasons*.

Such defence for epistemic closure is similar to (Klein.P. 1995)'s theory, which distinguishes two "sources of justification". And our above defence can be seen as an analogous defence of closure for knowledge except that we require *externally situated evidence* to be sound enough. However, Klein's theory is blamed for the problem of knowledge inflation by Holliday in Section 4.3 of (Holliday 2012).[7] Holliday points out that there is a gap between the knowledge gotten through *externally situated evidence* and the knowledge gotten through *internally situated reasons*, because it can not be explained why an uneliminated possibility v with respect to all propositions, *including the P that gives the internally situated reasons*, after the reasons, however, will become eliminated with respect to any Q entailed by P, including P itself. Then does our defence also suffer from the similar problem? For K^E, once we know P, we get it from sound enough evidence or by inference. We know it without any uncertainty and there is not any uneliminated possibility.[8]

However, we do not deny the existence of the gap here. The logics proposed in this paper suggest that the gap does not locate between the knowledge gotten through *externally situated evidence* and the knowledge gotten through *internally situated reasons*, but between the evidences and the knowledge: how we can get knowledge from evidence and how we can step from evidential state to epistemic state? In a sense, this question is similar to the problem Immanuel Kant dedicated to solve in his *The Critique of Pure Reason*: How are synthetic a priori judgements possible? The further analysis of this philosophical problem is beyond the scope of this paper. It needs to be stressed here only that in such perspective the epistemic closure can be saved, because the theory of truth tracking (the source of closure failure) is used to decide the external resource of knowledge (E_K), but not the knowledge itself. The closure failure occurs in the evidential state but not in the epistemic state.

[7] We refer the reader to Section 4.3 of (Holliday 2012) for the details, and only draw the outline here.

[8] All the knowledge we discuss here is implicit knowledge and it suffers from the omniscience problem. But that is another story.

6 Related work, conclusion and future work

In the logic of evidence-based knowledge, the definition of knowledge (belief) involves a philosophical assumption that we can derive knowledge (belief) from evidence and evidence is priori to knowledge (belief). Baltag et al. (2012) and Baltag et al. (2013) share such an assumption.[9] Nevertheless, they adopt a totally different approach to analyse those concepts. Combining an innovative modification of the Fitting semantics (Fitting 2005) for Artemov's Justification Logic (Artemov 2008) and ideas of belief revision awareness logics (van Benthem and Velazquez-Quesada 2010) and (Baltag and Smets 2008), Baltag et al. (2012) introduces many evidence-related notions like "acceptance", "admissibility" and "availability", and characterizes knowledge and belief in terms of the evidential reasoning that justifies those attitudes. Furthermore, Baltag et al. (2013) adopts the notion of "conclusive evidence" to study evidence-based notions. Though very different from our ideas, those works are inspiring. We think that the tracking theory can also be applied to Artemov's justification logic, and the logic of evidence-based belief (van Benthem and Pacuit 2011) could provide a direct, semantic approach to evidence that may have natural connections with the work of Baltag et al. (2012) and Baltag et al. (2013).

To sum up, the main aim of this paper is to define knowledge based on evidence. For this purpose, we introduced a logic of sound enough evidence and extended it to the logic of evidence-based knowledge. The newly defined knowledge operator has some interesting properties. Firstly, it does not imply belief. Secondly, although its definition depends on the use of the theory of truth tracking (which is the source of closure failure of CB-knowledge in Section 3), CE-knowledge ensures epistemic closure. The distinction between the evidential state and the epistemic state can well explain the difference between CB-knowledge and CE-knowledge. And such a distinction also brings about some further philosophical problems, like the gap mentioned above.

For future work, we would like to study the other resource of knowledge – "inference", which was mentioned a lot in this paper and moreover, it is closely related to the problem of omniscience. And we also want to explore the multi-agent situation and the dynamics of the evidence-based knowledge.

References

S. Artemov. The logic of justification. *The Review of Symbolic Logic*, 1:477–513,

[9] Different from this assumption, Williamson (1997) defends the principle that knowledge, and only knowledge, constitutes evidence ($E = K$). The claim of Williamson (1997) naturally implies that belief is derived from knowledge, because for Williamson belief can be derived from evidence.

2008.

A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonano, W. van der Hoek, and M. Wooldridge, editors, *Logic and the foundations of game and decision theory (LOFT 7)*, pages 9–58. Armsterdam University Press, 2008.

A. Baltag, B. Renne, and S. Smets. The logic of justified belief change, soft evidence and defeasible knowledge. In L. Ong and R. de Queiroz, editors, *Proceedings of the 19th Workshop of Logic, Language, Information and Computation (WoLLIC 2012)*, volume 7456 of *Lecture Notes in Computer Science*, pp. 168–190, Buenos Aires, Argentina. Springer-Verlag 2012.

A. Baltag, B. R, and S. S. The logic of justified belief, explicit knowledge and conclusive evidence. Manuscript, March 2013.

J. van Benthem and E. Pacuit. Dynamic logics of evidence-based beliefs. *Studia Logica*, 99:61–92, 2011.

J. van Benthem and F. Velazquez-Quesada. The dynamics of awareness. *Synthese*, 177:5–27, 2010.

J. van Benthem, D. Fernandez, and E. Pacuit. Evidence logic: a new look at neighborhood structure. *Advances in Modal Logic*, 9:97–118, 2012.

O. Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49: 49–80, 2004.

I. Cornelisse. Context dependence of epistemic operators in dynamic evidence logic. Master's thesis, University of Amsterdeam, ILLC, 2011.

F. Dreske. Epistemic operators. *The Journal of Philosophy*, 67:1007–1023, 1970.

F. Dreske. The pragmatic dimention of knowledge. *Philosophical Studies*, 40:363–378, 1981.

D. Dubois and H. Prade. Evidence, knowledge, and belief functions. *International Journal of Approximate Reasoning*, 6:295–319, 1992.

M. Fitting. The logic of proofs, semantically. *Annals of Pure and Apllied Logic*, 132: 1–25, 2005.

W. Holiday. Epistemic closure and epistemic logic i: Relevant alternatives and subjunctivism. Final version to appear in *Journal of Philosophical Logic*, 2013.

W. Holliday. *Knowing what follows: epistemic closure and epistemic logic*. PhD thesis, Stanford University, 2012.

P. Klein. Skepticism and closure: why the evil genius argument fails. *Philosophical Topics*, 23:213–236, 1995.

D. Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973.

D. Lewis. Elusive knowledge. *Australasian Journal of Philosophy*, 74:549–567, 1996.

P. K. Moser. *Knowledge and evidence*. Cambridge University Press, 1991.

R. Nozick. *Philosophical Explanations*. Harvard University Press, 1981.

T. Williamson. Knowledge as evidence. *Mind*, 106:717–742, 1997.

Appendix

A Proof of Theorem 1

From right to left

Lemma 1. *If the condition (b) of Theorem 1 holds and* $\bigcap_{\varphi\in\Phi} Min_{\leqslant_w}([\![\varphi]\!]) \neq \emptyset$*, then for any pointed CE model* $\mathcal{M}, w$

$$Min_{\leqslant_w}(\overline{[\![\psi]\!]}) \subseteq \bigcup_{\varphi\in\Phi} Min_{\leqslant_w}(\overline{[\![\varphi]\!]})$$

Proof. See Lemma 5.10 in (Holiday 2013). □

With this lemma, we can prove the right-to-left directions of Theorem 1.

Proof. If (a) holds, then it is immediate that $\chi_{n,m}$ is valid, since its antecedent is always false. For (b), we assume for a pointed CE model $\mathcal{M}, w$ that

$$\mathcal{M}, w \models \bigwedge_{\varphi\in\Phi} \Box_K \varphi \tag{2}$$

then by the truth definition (Def. 3.2),

$$\mathcal{M}, w \models \bigwedge_{\varphi\in\Phi} (\Box\varphi \wedge \neg\Box\neg\varphi), \tag{3}$$

$$\bigcup_{\varphi\in\Phi} Min_{\leqslant_w}(\overline{[\![\varphi]\!]}) \subseteq \bigcup_{\varphi\in\Phi} [\![\Box\neg\varphi]\!] \bigcap_{\varphi\in\Phi} Min_{\leqslant_w}([\![\varphi]\!]) \subseteq \bigcap_{\varphi\in\Phi}([\![\Box\varphi]\!] \cap \overline{[\![\Box\neg\varphi]\!]}) \tag{4}$$

On the other hand it follows from (b) and (1) in Theorem 1,

$$\bigcap_{\varphi\in\Phi} [\![\Box\varphi]\!] = [\![\Box\psi]\!], \bigcup_{\varphi\in\Phi} \overline{[\![\Box\varphi]\!]} = \overline{[\![\Box\psi]\!]}, \tag{5}$$

$$\bigcup_{\varphi\in\Phi} [\![\Box\neg\varphi]\!] = [\![\Box\neg\psi]\!], \bigcap_{\varphi\in\Phi} \overline{[\![\Box\neg\varphi]\!]} = \overline{[\![\Box\neg\psi]\!]} \tag{6}$$

By (8) and (9), (5) implies $\mathcal{M}, w \models \Box\psi \wedge \neg\Box\neg\psi$. By (b), Lemma 1(a) and (9), (7) implies $Min_{\leqslant_w}(\overline{[\![\psi]\!]}) \subseteq [\![\Box\neg\psi]\!]$. Then $\mathcal{M}, w \models \psi$ (suppose not, $w \in Min_{\leqslant_w}(\overline{[\![\psi]\!]}) \subseteq [\![\Box\neg\psi]\!]$, then we have $\mathcal{M}, w \models \neg\Box\psi$, which contradicts $\mathcal{M}, w \models \Box\psi \wedge \neg\Box\neg\psi$).

It follows that $Min_{\leqslant_w}([\![\psi]\!]) = \bigcap_{\varphi\in\Phi} Min_{\leqslant_w}([\![\varphi]\!])$. Then by (b), (8) and (9), (7) implies $Min_{\leqslant_w}([\![\psi]\!]) \subseteq [\![\Box\psi \wedge \neg\Box\neg\psi]\!]$.

Therefore, $\mathcal{M}, w \models \Box_K\psi$, given $\mathcal{M}, w \models \bigwedge_{\varphi\in\Phi} \Box_K\varphi$ □

From left to right

Proof. Suppose that neither (a) nor (b) holds for $\chi_{n,m}$, we will prove there is a pointed CE model $\mathcal{M}, w$ such that $\mathcal{M}, w \not\models \chi_{n,m}$.

For each $k \leqslant m$, let $S_k = \{i \mid 0 \leqslant i \leqslant n \text{ and } \models \psi_k \rightarrow \varphi_i\}$. Since (b) does not hold for $\chi_{n,m}$, there must be

$$\not\models \bigwedge_{i\in S_k} \varphi_i \rightarrow \psi_k \tag{7}$$

Construct $\mathcal{M} = \langle W, E, \leqslant\rangle$ as follows (see Figure 4):

$W = \{w\} \cup \{v_k \mid k \leqslant m\} \cup \{u\}$;

$E = \langle w, \{W, \{w\}\}\rangle, \langle v_k, \{W, \{u\}\}\rangle, \langle u, \{W\}\rangle$;

$\leqslant_w = \{\langle w, w\rangle\} \cup \bigcup_{k\leqslant m}\{\langle w, v_k\rangle\} \cup \bigcup_{k,j\leqslant m}\{\langle v_k, v_j\rangle\}$;

For $x \in W \setminus \{w\}$, $\leqslant_x$ is any total preorder on any $W_x \subseteq W$ with $x \in Min_{\leqslant_x}(W)$;

V is any valuation function on W such that $\mathcal{M}, w \models \varphi_0 \wedge \ldots \wedge \varphi_n$, and for all $k \leqslant m$, $\mathcal{M}, v_k \models \bigwedge_{i\in S_k} \varphi_i \wedge \neg\psi_k$ *and* $\mathcal{M}, u \models \bigwedge_{\varphi\in\Phi} \neg\varphi \wedge \bigwedge_{\psi\in\Psi} \psi$.

Then it is easy to check that $\mathcal{M}, w \not\models \Box_K\psi_k$ for all $k \leqslant m$, while $\mathcal{M}, w \models \bigwedge_{\varphi\in\Phi} \Box_K\varphi$.
□

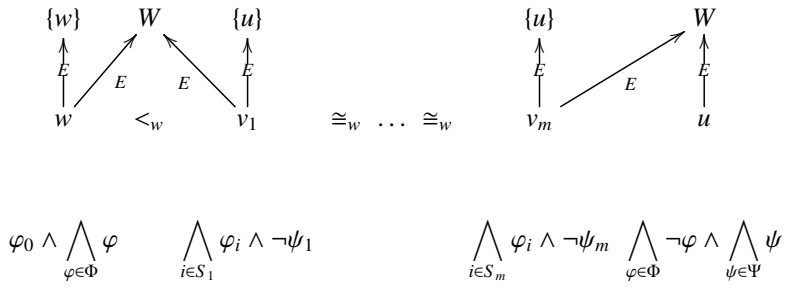


Figure 4: countermodel for $\chi_{n,m}$

B Proof of Theorem 2

Soundness

The soundness of most axioms and rules can be referred to (van Benthem et al. 2012) and (Board 2004). Here we only prove the case of $\Box\!\!\rightarrow$-4 and the case of $\Box\!\!\rightarrow$-5.

$\Box\!\!\rightarrow$-4: We must show that for any counterfactual evidence model $\mathcal{M}$ and any state $w \in \mathcal{M}$, $(\mathcal{M}, w) \models \neg(\varphi \ \Box\!\!\rightarrow \ \neg\psi) \rightarrow (((\varphi \wedge \psi \ \Box\!\!\rightarrow \chi) \leftrightarrow (\varphi \ \Box\!\!\rightarrow \ (\psi \rightarrow \chi)))$. Assume $\mathcal{M}, w \models \neg(\varphi \ \Box\!\!\rightarrow \ \neg\psi)$, then $Min_{\leqslant_w}([\![\varphi]\!] \cap W_w) \cap [\![\psi]\!] \neq \emptyset$. By the transitivity and totality of $\leqslant_w$, we have $Min_{\leqslant_w}([\![\varphi]\!] \cap [\![\psi]\!] \cap W_w) = Min_{\leqslant_w}([\![\varphi]\!] \cap W_w) \cap [\![\psi]\!]$. Therefore $\forall y \in Min_{\leqslant_w}([\![\varphi]\!] \cap [\![\psi]\!] \cap W_w), (\mathcal{M}, w) \models \chi$ iff $\forall y \in Min_{\leqslant_w}([\![\varphi]\!] \cap W_w) \cap [\![\psi]\!], (\mathcal{M}, w) \models \chi$. It follows immediately that $((\varphi \wedge \psi \ \Box\!\!\rightarrow \chi) \leftrightarrow (\varphi \ \Box\!\!\rightarrow \ (\psi \rightarrow \chi))$.

$\Box\!\!\rightarrow$-5: We must show that $(\mathcal{M}, w) \models \varphi \wedge (\varphi \ \Box\!\!\rightarrow \ \psi) \rightarrow \psi$. Assume that $(\mathcal{M}, w) \models \varphi \wedge (\varphi \ \Box\!\!\rightarrow \psi)$. By the Definition 3.1 3(b) and $w \in [\![\varphi]\!], w \in Min_{\leqslant_w}([\![\varphi]\!])$. By $(\mathcal{M}, w) \models \varphi \ \Box\!\!\rightarrow \psi, w \in [\![\psi]\!]$.

Completeness

Following the standard strategy, we show that every $CEKS$-consistent formula in $\mathcal{L}_{EK}$ is satisfiable with respect to a counterfactual evidence model. To produce the satisfying model we firstly construct maximally consistent sets (MCSs) from the subformulas of φ.

Let $Sub(\varphi)$ consist of all subformulas of φ: $\psi \in Sub(\varphi)$ if either (a) $\psi = \varphi$, or (b) φ is of the form $\neg\varphi', \varphi' \wedge \varphi'', \varphi' \ \Box\!\!\rightarrow \ \varphi''$, or $\Box\varphi'$, and $\psi \in Sub(\varphi')$ or $\psi \in sub(\varphi'')$; let $Sub^+(\varphi)$ be the smallest set such that (a) if $\psi \in Sub(\varphi)$ then $\psi \in Sub^+(\varphi)$; (b) if $\psi, \chi \in Sub^+(\varphi)$, then $\neg\psi, \psi \wedge \chi \in Sub^+(\varphi)$.

Let $Sub^{++}(\varphi)$ be the set of all formulas of $Sub^{+}(\varphi)$ and all formulas of the form $\chi \Box\!\!\rightarrow \psi$, where $\psi, \chi \in Sub^{+}(\varphi)$; and let $Sub^{++}_{neg}(\varphi)$ consist of (a) all the formulas in $Sub^{++}(\varphi)$ and their negations; (b) $\Box\top$ and $\neg\Box\bot$. Let $Con(\varphi)$ be the set of maximal CEKS-consistent subsets of $Sub^{++}_{neg}(\varphi)$.

Before the construction of the canonical model, we need some notations and definitions:

Notation. $\Gamma^{\bigcirc}_{\varphi} = \{\psi \mid \bigcirc\psi \in \Gamma_{\varphi} : \bigcirc = \chi \Box\!\!\rightarrow, \Box, K^{E}\}$

For the evidence relation E in the canonical model for EKL (see van Benthem et al. 2012, Definition 4.3 and 4.4 (iii), p.10.), we need to define evidence sets on the canonical model:

Definition 1. *Given a MCS* Γ *with* $\Box\alpha \in \Gamma$*, we define the* α*-neighborhood of* Γ *as*

$$\mathcal{N}^{\alpha}(\Gamma) = \{\Delta \in Con(\varphi) \mid \alpha \in \Delta\}$$

For $\leqslant_{\Gamma}$ in the canonical model, because of the difference of our definition of $Min_{\leqslant_{\Gamma}}$ from Board's, we use Board's definition as a pre-relation R_{Γ}:

Definition 2. $\Delta R_{\Gamma}\Theta$ *if there is some* $\psi \in Sub^{+}(\varphi) \cap \Delta \cap \Theta$ *such that* $\Gamma^{\psi\Box\!\!\rightarrow} \subseteq \Delta$

Then let $W^{\Gamma} = \{\Delta \mid \Delta R_{\Gamma}\Theta\}$.

Now we are ready to construct the canonical model M_{φ}

Definition 3. *Let* $M_{\varphi} = \langle W, E, \leqslant, V\rangle$ *where*

1. $W = Con(\varphi)$
2. $E(\Gamma) = \{\mathcal{N}^{\alpha}(\Gamma) \mid \Box\alpha \in \Gamma\}$
3. $\Delta \leqslant_{\Gamma} \Theta$ *iff* $\Delta R_{\Gamma}\Theta$ *and* $\Theta \in W^{\Gamma}$
4. *For each* $p \in \mathbf{At} \cap \Gamma, \Gamma \in V(p)$ *iff* $p \in \Gamma$.

Next we prove the truth lemma:

Lemma 2. *Truth Lemma For every* $\psi \in Sub^{++}_{neg}(\varphi)$ *and every* $\Gamma \in Con(\varphi)$*,*

$$\psi \in \Gamma \Leftrightarrow (\mathcal{M}_{\varphi}, \Gamma) \models \psi$$

Proof. The proof is by induction on the structure of ψ. The only interesting cases are the modality $\Box$ and $\Box\!\!\rightarrow$.

For the case of $\Box$, we only prove that right-to-left direction, and the converse is totally the same as that in (van Benthem et al. 2012) (Proof of Proposition 4.11, p.11). Assume that $(\mathcal{M}_\varphi, \Gamma) \models \Box\psi$. Then there exist an evidence $X = \mathcal{N}^\alpha(\Gamma) \in E(\Gamma)$ such that for all $\Delta \in X, (\mathcal{M}_\varphi, \Delta) \models \psi$. By the inductive hypothesis, for all $\Delta \in X, \psi \in \Delta$. By the definition of $\mathcal{N}^\alpha(\Gamma)$, for all $\Delta \in W$, if $\alpha \in \Delta, \psi \in \Delta$. It follows that for all $\Delta \in W, \alpha \rightarrow \psi \in \Delta$. Thus, $\alpha \rightarrow \psi \in \Gamma$. By $\Box$-monotonicity and $\Box\alpha \in \Gamma$, $\Box\psi \in \Gamma$, as desired.

For the case of $\Box\!\!\rightarrow$, although there is no essential difference between the proof of (Board 2004) for the case of $B^\chi\zeta$ and our proof, there are some modifications made by us. For clearness, we write down the details of the modified part.

Assume that $\psi \in \Gamma$ where ψ is of the form $\chi \ \Box\!\!\rightarrow \zeta$. It follows immediately that $\zeta \in \Gamma^{\chi\Box\rightarrow}$. Consider the set $Min_{\leqslant_\Gamma}([\![\chi]\!] \cap W_\Gamma)$. (Notice the difference between W^Γ and W_Γ). If this set is empty, then we have $(\mathcal{M}_\varphi, \Gamma) \models \chi \ \Box\!\!\rightarrow \zeta$ from the truth condition of $\Box\!\!\rightarrow$.

Suppose then there is some $\Delta \in Min_{\leqslant_\Gamma}([\![\chi]\!] \cap W_\Gamma)$. It is easy to prove that $W_\Gamma \subseteq W^\Gamma$. Therefore there is some $\xi \in Sub^+(\varphi) \cap \Delta$ such that $\Gamma^{\xi\Box\rightarrow} \subseteq \Delta$. We will prove that $\zeta \in \Delta$. Since $\Gamma^{\xi\Box\rightarrow} \subseteq \Delta, \Gamma^{\xi\Box\rightarrow}$ must be a CES-consistent set, $\Gamma^{\chi\Box\rightarrow}$ is a CES-consistent set too. Suppose not: then we have some finite set of formulas $F = \{\varphi_1, \ldots, \varphi_K\} \subseteq \Gamma^{\chi\Box\rightarrow}$ satisfying $CES \vdash \neg(\varphi_1 \wedge \ldots \wedge \varphi_k)$. Letting η denote $(\varphi_1, \ldots, \varphi_k)$, we have:

1.	$CES \vdash \neg\eta$	Assumption
2.	$CES \vdash \eta \rightarrow \xi$	1,Taut,MP
3.	$CES \vdash (\chi \ \Box\!\!\rightarrow \eta) \rightarrow (\chi \ \Box\!\!\rightarrow \xi)$	2,RE,$\Box\!\!\rightarrow$-2,Taut,MP
4.	$CES \vdash (\chi \ \Box\!\!\rightarrow \xi) \rightarrow (((\chi \wedge \xi) \ \Box\!\!\rightarrow \eta) \leftrightarrow (\chi \ \Box\!\!\rightarrow \eta))$	$\Box\!\!\rightarrow$-3
5.	$CES \vdash (\chi \ \Box\!\!\rightarrow \eta) \rightarrow ((\chi \wedge \xi) \ \Box\!\!\rightarrow \eta)$	3,4,Taut,MP
6.	$CES \vdash \neg(\xi \ \Box\!\!\rightarrow \neg\chi) \rightarrow (((\xi \wedge \chi) \ \Box\!\!\rightarrow \eta) \leftrightarrow (\xi \ \Box\!\!\rightarrow (\chi \rightarrow \eta)))$	$\Box\!\!\rightarrow$-4
7.	$CES \vdash \neg(\xi \ \Box\!\!\rightarrow \neg\chi) \rightarrow (((\chi \wedge \xi) \ \Box\!\!\rightarrow \eta) \leftrightarrow (\xi \ \Box\!\!\rightarrow (\chi \rightarrow \eta)))$	6,LE,Taut,MP
8.	$CES \vdash (\chi \rightarrow \eta) \rightarrow \neg\chi$	1,Taut,MP
9.	$CES \vdash (\xi \ \Box\!\!\rightarrow \chi) \rightarrow ((\chi \rightarrow \eta) \rightarrow (\xi \ \Box\!\!\rightarrow \neg\chi))$	8,N,$\Box\!\!\rightarrow$-2,Taut,MP
10.	$CES \vdash \neg(\xi \ \Box\!\!\rightarrow \neg\chi) \wedge (\chi \ \Box\!\!\rightarrow \eta) \rightarrow (\xi \ \Box\!\!\rightarrow \neg\chi)$	5,7,9,Taut,MP

It follows that $\chi \in \Delta$ from the hypothesis of induction and $\Delta \in [\![\chi]\!]$. It implies that $\neg(\xi \ \Box\!\!\rightarrow \neg\chi) \in \Gamma$ since $\xi \ \Box\!\!\rightarrow \neg\chi \in Sub^{++}(\varphi)$. For $\eta \in \Gamma^{\chi\Box\rightarrow}, \chi \ \Box\!\!\rightarrow \eta \in \Gamma$. By line 10, $\xi \ \Box\!\!\rightarrow \neg\chi \in \Gamma$. It follows that $\neg\chi \in \Gamma^{\xi\Box\rightarrow} \subseteq \Delta$, contradicting the fact $\chi \in \Delta$. So $\Gamma^{\chi\Box\rightarrow}$ is a CES-consistent set and it has a maximal CES-consistent extension Λ.

For the left part of the proof, we refer the reader to (Board 2004)'s proof. □

We have shown that the truth lemma holds for all formulas $\psi \in Sub(\varphi)$. To complete the proof of completeness, we need to show that M_φ really is a counterfactual evidence structure. Here we only prove that for all $\Gamma \in W, \leqslant_\Gamma$ satisfies the condition "for

all $\Delta \in W_\Gamma, \Gamma \leqslant_\Gamma \Delta$". The proof of well-foundness, transitivity and totality of $\leqslant_\Gamma$ for all Γ can be found in (Board 2004). And It is easy to see that "for all $\Gamma \in W, W \in E(\Gamma)$ and $\emptyset \notin E(\Gamma)$" follows from our stipulation that $\Box\top, \neg\Box\bot$ are both in all the MCSs and the definition of $E(\Gamma)$.

Lemma 3. *For all* $\Delta \in W_\Gamma, \Gamma \leqslant_\Gamma \Delta$

Proof. For any $\Delta \in W$, let $\psi \in Sub^+(\varphi) \cap \Gamma \cap \Delta$. For any $\chi \in \Gamma^{\psi\Box\rightarrow}$, $\chi \in \Gamma$, since $\psi \wedge (\psi \Box\rightarrow \chi) \rightarrow \chi \in \Gamma$ by $\Box\rightarrow$-5. Therefore, $\Gamma^{\psi\Box\rightarrow} \subseteq \Gamma$, which means that $\Gamma \leqslant_\Gamma \Delta$ by the definition of $\leqslant_\Gamma$. □

With all above work, the evidence logic's weak-completeness are proved for the class of counterfactual evidence models.

Hybrid-Logical Reasoning in False-Belief Tasks

Torben Braüner

Programming, Logic and Intelligent Systems Research Group, Roskilde University
torben@ruc.dk

Abstract

The main aim of the present paper is to use a proof system for hybrid modal logic to formalize what are called false-belief tasks in cognitive psychology, thereby investigating the interplay between cognition and logical reasoning about belief. We consider two different versions of the Smarties task, involving respectively a shift of perspective to another person and to another time. Our formalizations disclose that despite this difference, the two versions of the Smarties task have exactly the same underlying logical structure. We also consider the Sally-Anne task, having a somewhat more complicated logical structure, presupposing a "principle of inertia" saying that a belief is preserved over time, unless there is belief to the contrary.[1]

1 Introduction

In the area of cognitive psychology there is a reasoning task called the *Smarties task*. The following is one version of the Smarties task.

> A child is shown a Smarties tube where unbeknownst to the child the Smarties have been replaced by pencils. The child is asked: "What do you think is inside the tube?" The child answers "Smarties!" The tube is then shown to contain pencils only. The child is then asked: "If your mother

[1] The present paper is a reformatted version of the paper (Braüner 2013) originally published in *Proceedings of Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2013)*. Beside reformatting, a couple of minor errors have been corrected.

> comes into the room and we show this tube to her, what will she think is inside?"

It is well-known from experiments that most childred above the age of four correctly say "Smarties" (thereby attributing a false belief to the mother) whereas younger children say "Pencils" (what they know is inside the tube). For autistic[2] children the cutoff age is higher than four years, which is one reason to the interest in the Smarties task.

The Smarties task is one out of a family of reasoning tasks called *false-belief tasks* showing the same pattern, that most children above four answer correctly, but autistic children have to be older. This was first observed in the paper (Baron-Cohen et al. 1985) in connection with another false-belief task called the *Sally-Anne task*. Starting with the authors of that paper, many researchers in cognitive psychology have argued that there is a link between autism and a lack of what is called *theory of mind*, which is a capacity to imagine other people's mental states, for example their beliefs. For a very general formulation of the theory of mind deficit hypothesis of autism, see the book (Baron-Cohen 1995).

Giving a correct answer to the Smarties task involves a shift of perspective to another person, namely the mother. You have to put yourself in another person's shoes, so to speak. Since the capacity to take another perspective is a precondition for figuring out the correct answer to the Smarties task and other false-belief tasks, the fact that autistic children have a higher cutoff age is taken to support the claim that autists have a limited or delayed theory of mind. For a critical overview of these arguments, see the book (Stenning and van Lambalgen 2008) by Keith Stenning and Michiel van Lambalgen. The books (Stenning and van Lambalgen 2008) and (Baron-Cohen 1995) not only consider theory of mind at a cognitive level, such as in connection with false-belief tasks, but they also discuss it from a biological point of view.

In a range of works van Lambalgen and co-authors have given a detailed logical analysis (but not a full formalization) of the reasoning taking place in the Smarties task and other false-belief tasks in terms of closed-world reasoning as used in non-monotonic logics, see in particular (Stenning and van Lambalgen 2008). The analysis of the Smarties task of (Stenning and van Lambalgen 2008) (in Subsection 9.4.4) makes use of a modality B for belief satisfying two standard modal principles.[3] The first principle is $B(\varphi \rightarrow \psi) \rightarrow (B\varphi \rightarrow B\psi)$ (principle (9.5) at page 251 in Stenning and van

[2] Autism is a psychiatric disorder with the following three diagnostic criteria: 1. Impairment in social interaction. 2. Impairment in communication. 3. Restricted repetitive and stereotyped patterns of behavior, interests, and activities. For details, see *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV)*, published by the American Psychiatric Association.

[3] Strictly speaking, the modality B in (Stenning and van Lambalgen 2008) is not formalized in terms of modal logic, but in terms of what is called event calculus, where B is a predicate that can take formulas as arguments.

Lambalgen 2008). The second principle is the rule called necessitation, that is, from φ derive $B\varphi$ (this principle is not mentioned explicitly in (Stenning and van Lambalgen 2008), but is implicit in the analysis given at the bottom of page 256). These two principles together imply that belief is closed under logical consequence, that is, $B\psi$ can be derived from $\varphi \to \psi$ together with $B\varphi$, which at least for human agents is implausible (when the modal operator stands for knowledge, this is called logical omniscience).

In the present paper we give a logical analysis of the perspective shift required to give a correct answer to the Smarties and Sally-Anne tasks, and we demonstrate that these tasks can be fully formalized in a hybrid-logical proof system not assuming principles implying logical omniscience, namely the natural deduction system described in Chapter 4 of the book (Braüner 2011), and the paper (Braüner 2004) as well. Beside not suffering from logical omniscience, why is a *natural deduction* system for *hybrid modal logic* appropriate to this end?

- The subject of proof-theory is the notion of proof and formal, that is, symbolic, systems for representing proofs. Formal proofs built according to the rules of proof systems can be used to represent—describe the structure of—mathematical arguments as well as arguments in everyday human practice. Beside giving a way to distinguish logically correct arguments from incorrect ones, proof systems also give a number of ways to characterize the structure of arguments. Natural deduction style proofs are meant to formalize the way human beings actually reason, so natural deduction is an obvious candidate when looking for a proof system to formalize the Smarties task in.

- In the standard Kripke semantics for modal logic, the truth-value of a formula is relative to points in a set, that is, a formula is evaluated "locally" at a point, where points usually are taken to represent possible worlds, times, locations, epistemic states, persons, states in a computer, or something else. Hybrid logics are extended modal logics where it is possible to directly refer to such points in the logical object language, whereby locality can be handled explicitly, for example, when reasoning about time one can formulate a series of statements about what happens at specific times, which is not possible in ordinary modal logic. Thus, when points in the Kripke semantics represent local perspectives, hybrid-logical machinery can handle explicitly the different perspectives in the Smarties task.

For the above reasons, we have been able to turn our informal logical analysis of the Smarties and Sally-Anne tasks into formal hybrid-logical natural deduction proofs closely reflecting the shift between different perspectives.

The natural deduction system we use for our formalizations is a modified version of a natural deduction system for a logic of situations similar to hybrid logic, originally introduced by Jerry Seligman (1997). The modified system was introduced in the paper (Braüner 2004), and later on considered in Chapter 4 of the book (Braüner 2011), both by the present author. In what follows we shall simply refer to the modified system as Seligman's system.

Now, Seligman's system allows any formula to occur in it, which is different from the most common proof systems for hybrid logic that only allow formulas of a certain form called satisfaction statements. This is related to a different way of reasoning in Seligman's system, which captures particularly well the reasoning in the Smarties and Sally-Anne tasks. We prove a completeness result which also says that Seligman's system is analytic, that is, we prove that any valid formula has a derivation satisfying the subformula property. Analyticity guarentees that any valid argument can be formalized using only subformulas of the premises and the conclusion. The notion of analyticity goes back to G.W. Leibniz (1646–1716) who called a proof analytic if and only if the proof is based on concepts contained in the proven statement, the main aim being to be able to construct a proof by an analysis of the result (cf. Baaz and Leitsch 2011).

The present paper is structured as follows. In the second section we recapitulate the basics of hybrid logic, readers well-versed in hybrid logic can safely skip this section. In the third section we introduce Seligman's natural deduction system for hybrid logic and in the fourth section we give a first example of reasoning in this system. In the fifth and sixth sections we formalize two versions of the Smarties task using this system, and in the seventh section we formalize the Sally-Anne task. A discussion can be found in the eightth section, in the nineth section there are some brief remarks on other work, and in the final section some remarks on further work. In the appendix we prove the above mentioned completeness result, which also demonstrates analyticity.

2 Hybrid logic

The term "hybrid logic" covers a number of logics obtained by adding further expressive power to ordinary modal logic. The history of what now is known as hybrid logic goes back to the philosopher Arthur Prior's work in the 1960s. See the handbook chapter (Areces and ten Cate 2007) for a detailed overview of hybrid logic. See the book (Braüner 2011) on hybrid logic and its proof-theory.

The most basic hybrid logic is obtained by extending ordinary modal logic with *nominals*, which are propositional symbols of a new sort. In the Kripke semantics a nominal is interpreted in a restricted way such that it is true at exactly one point. If the

points are given a temporal reading, this enables the formalization of natural language statements that are true at exactly one time, for example

it is five o'clock May 10th 2007

which is true at the time five o'clock May 10th 2007, but false at all other times. Such statements cannot be formalized in ordinary modal logic, the reason being that there is only one sort of propositional symbol available, namely ordinary propositional symbols, which are not restricted to being true at exactly one point.

Most hybrid logics involve further additional machinery than nominals. There is a number of options for adding further machinery; here we shall consider a kind of operator called *satisfaction operators*. The motivation for adding satisfaction operators is to be able to formalize a statement being true at a particular time, possible world, or something else. For example, we want to be able to formalize that the statement "it is raining" is true at the time five o'clock May 10th 2007, that is, that

at five o'clock May 10th 2007 it is raining.

This is formalized by the formula $@_a r$ where the nominal a stands for "it is five o'clock May 10th 2007" as above and where r is an ordinary propositional symbol that stands for "it is raining". It is the part $@_a$ of the formula $@_a r$ that is called a satisfaction operator. In general, if a is a nominal and φ is an arbitrary formula, then a new formula $@_a \varphi$ can be built (in some literature the notation $a : \varphi$ is used instead of $@_a \varphi$). A formula of this form is called a *satisfaction statement*. The formula $@_a \varphi$ expresses that the formula φ is true at one particular point, namely the point to which the nominal a refers. Nominals and satisfaction operators are the most common pieces of hybrid-logical machinery, and are what we need for the purpose of the present paper.

In what follows we give the formal syntax and semantics of hybrid logic. It is assumed that a set of ordinary propositional symbols and a countably infinite set of nominals are given. The sets are assumed to be disjoint. The metavariables p, q, r, ... range over ordinary propositional symbols and a, b, c, ... range over nominals. Formulas are defined by the following grammar.

$$S \ ::= \ p \mid a \mid S \wedge S \mid S \rightarrow S \mid \bot \mid \Box S \mid @_a S$$

The metavariables φ, ψ, θ, ... range over formulas. Negation is defined by the convention that $\neg\varphi$ is an abbreviation for $\varphi \rightarrow \bot$. Similarly, $\Diamond\varphi$ is an abbreviation for $\neg\Box\neg\varphi$.

Definition 2.1. A *model* for hybrid logic is a tuple $(W, R, \{V_w\}_{w \in W})$ where

1. W is a non-empty set;

2. R is a binary relation on W; and
3. for each w, V_w is a function that to each ordinary propositional symbol assigns an element of $\{0, 1\}$.

The pair (W, R) is called a *frame*. Note that a model for hybrid logic is the same as a model for ordinary modal logic. Given a model $\mathfrak{M} = (W, R, \{V_w\}_{w \in W})$, an *assignment* is a function g that to each nominal assigns an element of W. The relation $\mathfrak{M}, g, w \models \varphi$ is defined by induction, where g is an assignment, w is an element of W, and φ is a formula.

$$
\begin{array}{rll}
\mathfrak{M}, g, w \models p & \text{iff} & V_w(p) = 1 \\
\mathfrak{M}, g, w \models a & \text{iff} & w = g(a) \\
\mathfrak{M}, g, w \models \varphi \wedge \psi & \text{iff} & \mathfrak{M}, g, w \models \varphi \text{ and } \mathfrak{M}, g, w \models \psi \\
\mathfrak{M}, g, w \models \varphi \rightarrow \psi & \text{iff} & \mathfrak{M}, g, w \models \varphi \text{ implies } \mathfrak{M}, g, w \models \psi \\
\mathfrak{M}, g, w \models \bot & \text{iff} & \text{falsum} \\
\mathfrak{M}, g, w \models \Box\varphi & \text{iff} & \text{for any } v \in W \text{ such that } wRv, \mathfrak{M}, g, v \models \varphi \\
\mathfrak{M}, g, w \models @_a\varphi & \text{iff} & \mathfrak{M}, g, g(a) \models \varphi
\end{array}
$$

By convention $\mathfrak{M}, g \models \varphi$ means $\mathfrak{M}, g, w \models \varphi$ for every element w of W and $\mathfrak{M} \models \varphi$ means $\mathfrak{M}, g \models \varphi$ for every assignment g. A formula φ is *valid* if and only if $\mathfrak{M} \models \varphi$ for any model $\mathfrak{M}$.

3 Seligman's system

In this section we introduce Seligman's natural deduction systems for hybrid logic. Before defining the system, we shall sketch the basics of natural deduction. Natural deduction style derivation rules for ordinary classical first-order logic were originally introduced by Gerhard Gentzen (1969) and later on developed much further by Dag Prawitz (1965, 1971). See (Troelstra and Schwichtenberg 1996) for a general introduction to natural deduction systems. With reference to Gentzen's work, Prawitz made the following remarks on the significance of natural deduction.

> ... the essential logical content of intuitive logical operations that can be formulated in the languages considered can be understood as composed of the atomic inferences isolated by Gentzen. It is in this sense that we may understand the terminology *natural* deduction.
>
> Nevertheless, Gentzen's systems are also natural in the more superficial sense of corresponding rather well to informal practices; in other words, the structure of informal proofs are often preserved rather well when formalized within the systems of natural deduction. (Prawitz 1971, p. 245)

Similar views on natural deduction are expressed many places, for example in a textbook by Warren Goldfarb.

> What we shall present is a system for *deductions*, sometimes called a system of *natural deduction*, because to a certain extent it mimics certain natural ways we reason informally. In particular, at any stage in a deduction we may introduce a new premise (that is, a new supposition); we may then infer things from this premise and eventually eliminate the premise (*discharge* it). (Goldfarb 2003, p. 181)

Basically, what is said by the second part of the quotation by Prawitz, and the quotation by Goldfarb as well, is that the structure of informal human arguments can be described by natural deduction derivations.

Of course, the observation that natural deduction derivations often can formalize, or mimic, informal reasoning does not itself prove that natural deduction is the mechanism underlying human deductive reasoning, that is, that formal rules in natural deduction style are somehow built into the human cognitive architecture. However, this view is held by a number of psychologists, for example Lance Rips in the book (Rips 1994), where he provides experimental support for the claim.

> ... a person faced with a task involving deduction attempts to carry it out through a series of steps that takes him or her from an initial description of the problem to its solution. These intermediate steps are licensed by mental inference rules, such as modus ponens, whose output people find intuitively obvious. (Rips 1994, p. x)

This is the main claim of the "mental logic" school in the psychology of reasoning (whose major competitor is the "mental models" school, claiming that the mechanism underlying human reasoning is the construction of models, rather than the application of topic-neutral formal rules).

We have now given a brief motivation for natural deduction and proceed to a formal definition. A *derivation* in a natural deduction system has the form of a finite tree where the nodes are labelled with formulas such that for any formula occurrence φ in the derivation, either φ is a leaf of the derivation or the immediate successors of φ in the derivation are the premises of a rule-instance which has φ as the conclusion. In what follows, the metavariables π, τ, ... range over derivations. A formula occurrence that is a leaf is called an *assumption* of the derivation. The root of a derivation is called the *end-formula* of the derivation. All assumptions are annotated with numbers. An assumption is either *undischarged* or *discharged*. If an assumption is discharged, then it is discharged at one particular rule-instance and this is indicated by annotating the

assumption and the rule-instance with identical numbers. We shall often omit this information when no confusion can occur. A rule-instance annotated with some number discharges all undischarged assumptions that are above it and are annotated with the number in question, and moreover, are occurrences of a formula determined by the rule-instance.

Two assumptions in a derivation belong to the same *parcel* if they are annotated with the same number and are occurrences of the same formula, and moreover, either are both undischarged or have both been discharged at the same rule-instance. Thus, in this terminology rules discharge parcels. We shall make use of the standard notation

$$\begin{array}{c} [\varphi^r] \\ \vdots\,\pi \\ \psi \end{array}$$

which means a derivation π where ψ is the end-formula and $[\varphi^r]$ is the parcel consisting of all undischarged assumptions that have the form φ^r.

We shall make use of the following conventions. The metavariables $\Gamma, \Delta, \ldots$ range over sets of formulas. A derivation π is called a *derivation of* φ if the end-formula of π is an occurrence of φ, and moreover, π is called a *derivation from* Γ if each undischarged assumption in π is an occurrence of a formula in Γ (note that numbers annotating undischarged assumptions are ignored). If there exists a derivation of φ from $\emptyset$, then we shall simply say that φ is *derivable*.

A typical feature of natural deduction is that there are two different kinds of rules for each connective; there are rules called introduction rules which introduce a connective (that is, the connective occurs in the conclusion of the rule, but not in the premises) and there are rules called elimination rules which eliminate a connective (the connective occurs in a premiss of the rule, but not in the conclusion). Introduction rules have names in the form $(\ldots I \ldots)$, and similarly, elimination rules have names in the form $(\ldots E \ldots)$.

Now, Seligman's natural deduction system is obtained from the rules given in Figure 1 and Figure 2. We let $\mathbf{N}'_{\mathcal{H}}$ denote the system thus obtained. The system $\mathbf{N}'_{\mathcal{H}}$ is taken from (Braüner 2004) and Chapter 4 of (Braüner 2011) where it is shown to be sound and complete wrt. the formal semantics given in the previous section. As mentioned earlier, this system is a modified version of a system originally introduced in (Seligman 1997). The system of (Seligman 1997) was modified in (Braüner 2004) and (Braüner 2011) with the aim of obtaining a desirable property called closure under substitution, see Subsection 4.1.1 of (Braüner 2011) for further explanation.

Figure 1: Rules for connectives

$$\frac{\varphi \quad \psi}{\varphi \wedge \psi}\,(\wedge I) \qquad \frac{\varphi \wedge \psi}{\varphi}\,(\wedge E1) \qquad \frac{\varphi \wedge \psi}{\psi}\,(\wedge E2)$$

$$\frac{\begin{array}{c}[\varphi]\\ \vdots\\ \psi\end{array}}{\varphi \to \psi}\,(\to I) \qquad \frac{\varphi \to \psi \quad \varphi}{\psi}\,(\to E)$$

$$\frac{\begin{array}{c}[\neg\varphi]\\ \vdots\\ \bot\end{array}}{\varphi}\,(\bot)^{*}$$

$$\frac{a \quad \varphi}{@_a\varphi}\,(@I) \qquad \frac{a \quad @_a\varphi}{\varphi}\,(@E)$$

$$\frac{\begin{array}{c}[\Diamond c]\\ \vdots\\ @_c\varphi\end{array}}{\Box\varphi}\,(\Box I)^{\dagger} \qquad \frac{\Box\varphi \quad \Diamond e}{@_e\varphi}\,(\Box E)$$

$*$ φ is a propositional letter.
† c does not occur free in $\Box\varphi$ or in any undischarged assumptions other than the specified occurrences of $\Diamond c$.

Figure 2: Rules for nominals

$$\frac{\varphi_1 \quad \ldots \quad \varphi_n \qquad \begin{array}{c}[\varphi_1]\ldots[\varphi_n][a]\\ \vdots\\ \psi\end{array}}{\psi}\,(\mathcal{T}erm)^{*} \qquad \frac{\begin{array}{c}[a]\\ \vdots\\ \psi\end{array}}{\psi}\,(\mathcal{N}ame)^{\dagger}$$

$*$ $\varphi_1, \ldots, \varphi_n$, and ψ are all satisfaction statements and there are no undischarged assumptions in the derivation of ψ besides the specified occurrences of $\varphi_1, \ldots, \varphi_n$, and a.
† a does not occur in ψ or in any undischarged assumptions other than the specified occurrences of a.

4 A first example

The way of reasoning in Seligman's system is different from the way of reasoning in most other proof systems for hybrid logic[4]. In this section we give the first example of reasoning using the ($\mathcal{T}erm$) rule (displayed in Figure 2).

Beside the ($\mathcal{T}erm$) rule, the key rules in the example are the rules ($@I$) and ($@E$) (displayed in Figure 1), which are the introduction and elimination rules for the satisfaction operator. The rule ($@I$) formalizes the following informal argument.

> It is Christmas Eve 2011; it is snowing, so at Christmas Eve 2011 it is snowing.

And the rule ($@E$) formalizes the following.

> It is Christmas Eve 2011; at Christmas Eve 2011 it is snowing, so it is snowing.

The ($\mathcal{T}erm$) rule enables hypothetical reasoning where reasoning is about what is the case at a specific time, possibly different from the actual time. Consider the following informal argument.

> At May 10th 2007 it is raining; if it is raining it is wet, so at May 10th 2007 it is wet.

The reasoning in this example argument is about what is the case at May 10th 2007. If this argument is made at a specific actual time, the time of evaluation is first shifted from the actual time to a hypothetical time, namely May 10th 2007, then some reasoning is performed involving the premise "if it is raining it is wet", and finally the time of evaluation is shifted back to the actual time. The reader is invited to verify this shift of time by checking that the argument is correct, and note that the reader himself (or herself) imagines being at the time May 10th 2007. Note that the premise "if it is raining it is wet" represents a causal relation holding at all times.

Now, in a temporal setting, the side-condition on the rule ($\mathcal{T}erm$) requiring that all the formulas $\varphi_1, \ldots, \varphi_n, \psi$ are satisfaction statements (see Figure 2) ensures that these formulas are temporally definite, that is, they have the same truth-value at all times, so the truth-value of these formulas are not affected by a shift of temporal perspective. The rule would not be sound if the formulas were not temporally definite.

[4]We here have in mind natural deduction, Gentzen, and tableau systems for hybrid logic, not axiom systems. Proof systems of the first three types are suitable for actual reasoning, carried out by a human, a computer, or in some other medium. Axiom systems are usually not meant for actual reasoning, but are of a more foundational interest.

We now proceed to the formalization of the above argument about what is the case at May 10th 2007. We make use of the following symbolizations

p It is raining
q It is wet
a May 10th 2007

and we take the formula $p \to q$ as an axiom since it represents a causal relation between p and q holding at all times (note that we use an axiom since the relation $p \to q$ holds between the particular propositions p and q, we do not use an axiom schema since the relation obviously does not hold between any pair of propositions).[5] Then the argument can be formalized as

$$(1)\qquad \dfrac{\dfrac{@_a p \qquad \dfrac{\dfrac{[a] \qquad [@_a p]}{p}(@E) \qquad \dfrac{}{p \to q}(\mathcal{A}xiom)}{q}(\to E)}{@_a q}(@I)}{@_a q}(\mathcal{T}erm)$$

Note that the derivation (1) above is obtained by applying the $(\mathcal{T}erm)$ rule to the subderivation (2) below.

$$(2)\qquad \dfrac{a \qquad \dfrac{\dfrac{a \qquad @_a p}{p}(@E) \qquad \dfrac{}{p \to q}(\mathcal{A}xiom)}{q}(\to E)}{@_a q}(@I)$$

Thus, the $(\mathcal{T}erm)$ rule delimits a piece of reasoning taking place at a certain hypothetical time, which above is the subderivation (2).

The above example argument is similar to an example given in the paper (Seligman 1997). The following is a slightly reformulated version.

[5] One of the anonymous reviewers asked why the premise "if it is raining it is wet" is formalized as $p \to q$ using classical implication, rather than a form of non-monotonic implication. Like in many cases when classical logic is used to formalize natural language statements, there is an idealization in our choice of classical implication. We think this idealization is justified since our main goal is to formalize the perspective shift involved in the example argument, which we presume is orthogonal to the issue of non-monotonicity. We note in passing that our premise "if it is raining it is wet" corresponds to the premise "if alcohol is forbidden Sake is forbidden" in Seligman's example argument briefly described below, and Seligman also uses classical implication, or to be precise, machinery equivalent to classical implication, (Seligman 1997). See also Footnote 7.

> In Abu Dabi alcohol is forbidden; if alcohol is forbidden Sake is forbidden, so in Abu Dabi Sake is forbidden.

Thus, the example of (Seligman 1997) involves spatial locations rather than times, and the shift is to a hypothetical place, namely the city of Abu Dabi.

Formally, the shift to a hypothetical point of evaluation effected by the rule ($\mathcal{T}erm$) can be seen by inspecting the proof that the rule ($\mathcal{T}erm$) is sound: The world of evaluation is shifted from the actual world to the hypothetical world where the nominal a is true (see Figure 2), then some reasoning is performed involving the delimited subderivation which by induction is assumed to be sound, and finally the world of evaluation is shifted back to the actual world. Soundness of the system $\mathbf{N}'_{\mathcal{H}}$, including soundness of the rule ($\mathcal{T}erm$), is proved in Theorem 4.1 in Section 4.3 of (Braüner 2011).

The rule ($\mathcal{T}erm$) is very different from other rules in proof systems for hybrid logic, roughly, this rule replaces rules for equational reasoning in other systems, see for example the rules in the natural deduction system given in Section 2.2 of the book (Braüner 2011).

In passing we mention that the way in which the ($\mathcal{T}erm$) rule delimits a subderivation is similar to the way subderivations are delimited by so-called boxes in linear logic, and more specifically, the way a subderivation is delimited by the introduction rule for the modal operator $\Box$ in the natural deduction system for S4 given in (Bierman and de Paiva 2000), making use of explicit substitutions in derivations.

5 The Smarties task (temporal shift version)

In this section we will give a formalization which has exactly the same structure as the formalization in the previous section, but which in other respects is quite different. It turns out that a temporal shift like the one just described in the previous section also takes place in the following version of the Smarties task, where instead of a shift of perspective to another person, there is a shift of perspective to another time.[6]

> A child is shown a Smarties tube where unbeknownst to the child the Smarties have been replaced by pencils. The child is asked: "What do you think is inside the tube?" The child answers "Smarties!" The tube is then shown to contain pencils only. The child is then asked: "Before this tube was opened, what did you think was inside?"

[6] The author thanks Michiel van Lambalgen for mentioning the Smarties task in an email exchange where the author suggested that the shift of perspective in the hybrid-logical rule ($\mathcal{T}erm$) could be of relevance in connection with the theory of mind view of autism.

See (Gopnik and Astington 1988) for more on the temporal version of the Smarties task. Below we shall formalize each step in the logical reasoning taking place when giving a correct answer to the task, but before that, we give an informal analysis. Let us call the child Peter. Let a be the time where Peter answers the first question, and let t be the time where he answers the second one. To answer the second question, Peter imagines himself being at the earlier time a where he was asked the first question. At that time he deduced that there were Smarties inside the tube from the fact that it is a Smarties tube. Imagining being at the time a, Peter reasons that since he at that time deduced that there were Smarties inside, he must also have come to believe that there were Smarties inside. Therefore, at the time t he concludes that at the earlier time a he believed that there were Smarties inside.

We now proceed to the full formalization. We first extend the language of hybrid logic with two modal operators, D and B. We make use of the following symbolizations

D	Peter deduces that ...
B	Peter believes that ...
p	There are Smarties inside the tube
a	The time where the first question is asked

and we take the principle $D\varphi \rightarrow B\varphi$ as an axiom schema (it holds whatever proposition is substituted for the metavariable φ, hence an axiom schema). This is principle (9.4) in (Stenning and van Lambalgen 2008).[7] Then the shift of temporal perspective in the Smarties task can be formalized very directly in Seligman's system as

$$\dfrac{@_a Dp \qquad \dfrac{\dfrac{[a] \qquad \dfrac{[@_a Dp]}{Dp}(@E) \qquad \dfrac{}{Dp \rightarrow Bp}}{Bp}(\rightarrow E)}{@_a Bp}(@I)}{@_a Bp}(\mathcal{T}erm)$$

Recall that the derivation is meant to formalize each step in Peters's reasoning at the time t where the second question is answered. The premise $@_a Dp$ in the derivation says that Peter at the earlier time a deduced that there were Smarties inside the tube, which he remembers at t.

[7] Analogous to the question in Footnote 5, it can be asked why we use classical implication in $D\varphi \rightarrow B\varphi$, rather than a form of non-monotonic implication. Again, the answer is that this is an idealization, but we presume that the perspective shift involved in the Smarties task is orthogonal to the issue of non-monotonicity, at least from a logical point of view. In this connection we remark that principle (9.4) in (Stenning and van Lambalgen 2008) also uses classical implication (the non-monotonicity in the logical analysis of the Smarties task of (Stenning and van Lambalgen 2008) does not concern principle (9.4), but other principles).

Note that the formalization does not involve the $\Box$ operator, so this operator could have been omitted together with the associated rules ($\Box I$) and ($\Box E$) in Figure 1. Since this proof system is complete, the $\Box$ operator satisfies logical omniscience. The operators D and B are only taken to satisfy the principle $D\varphi \rightarrow B\varphi$, as mentioned above.

Compare the derivation above to the derivation (1) in the previous section and note that the structure is exactly the same. Note that what we have done is that we have formalized the logical reasoning taking place when giving the correct answer "Smarties". Note also that the actual content of the tube, namely pencils, is not even mentioned in the formalization, so it is clear from the formalization that the actual content of the tube is not relevant to figure out the correct answer. Accordingly, our formalization does not tell what goes wrong when a child incorrectly answers "Pencils".

6 The Smarties task (person shift version)

As a stepping stone between the temporal version of the Smarties task we considered in the previous section, and the Sally-Anne task we shall consider in the next section, we in the present section take a look again at the version of the Smarties task described in the introduction. The only difference between the version in the introduction and the version in the previous section is the second question where

> "Before this tube was opened, what did you think was inside?"

obviously gives rise to a temporal shift of perspective, whereas

> "If your mother comes into the room and we show this tube to her, what will she think is inside?"

gives rise to a shift of perspective to another person, namely the imagined mother.

To give a correct answer to the latter of these two questions, the child Peter imagines being the mother coming into the room. Imagining being the mother, Peter reasons that the mother must deduce that there are Smarties inside the tube from the fact that it is a Smarties tube, and from that, she must also come to believe that there are Smarties inside. Therefore, Peter concludes that the mother would believe that there are Smarties inside.

The derivation formalizing this argument is exactly the same as in the temporal case dealt with in previous section but the symbols are interpreted differently, namely as

D	Deduces that ...
B	Believes that ...
p	There are Smarties inside the tube
a	The imagined mother

So now nominals refer to persons rather than times. Accordingly, the modal operator B now symbolize the belief of the person represented by the point of evaluation, rather than Peter's belief at the time of evaluation, etc. Thus, the premise $@_a Dp$ in the derivation in the previous section now says that the imagined mother deduces that there are Smarties inside the tube, which the child doing the reasoning takes to be the case since the mother is imagined to be present in the room.

Incidentally, letting points in the Kripke model represent persons is exactly what is done in Arthur Prior's *egocentric logic*, see Section 1.3 in the book (Braüner 2011), in particular pp. 15–16. In egocentric logic the accessibility relation represents the taller-than relation, but this relation is obviously not relevant here.

7 The Sally-Anne task

In this section we will give a formalization of a somewhat more complicated reasoning task called the Sally-Anne task. The following is one version.

> A child is shown a scene with two doll protagonists, Sally and Anne, having respectively a basket and a box. Sally first places a marble into her basket. Then Sally leaves the scene, and in her absence, the marble is transferred by Anne and hidden in her box. Then Sally returns, and the child is asked: "Where will Sally look for her marble?"

Most children above the age of four correctly responds where Sally must falsely believe the marble to be (in the basket) whereas younger children respond where they know the marble to be (in the box). Again, for autists, the cutoff is higher.

Below we shall formalize the correct response to the task, but before that, we give an informal analysis. Let us call the child Peter again. Let t_1 be the time where he answers the question. To answer the question, Peter imagines himself being Sally at an earlier time t_0 before she leaves the scene, but after she places the marble in her basket. Imagining being Sally, he reasons as follows: At the time t_0 Sally believes that the marble is in the box since she can see it. At the time t_1, after she has returned, she deduces that the marble is still in the box as she has no belief to the contrary, and since Sally deduces that the marble is in the box, she must also come to believe it. Therefore, Peter concludes that Sally believes that the marble is in the box.

In our formalization we make use of a tiny fragment of first-order hybrid logic, involving the unary predicate $P(t)$, the binary predicate $t < u$, and the modal operators S, D and B, but no quantifiers. We make use of the following symbolizations.

$p(t)$	The marble is in the basket at the time t
$t < u$	The time t is before the time u
S	Sees that ...
D	Deduces that ...
B	Believes that ...
a	The person Sally

We also make use of the following three principles.

$$S\varphi \rightarrow B\varphi$$
$$D\varphi \rightarrow B\varphi$$
$$B\varphi(t) \wedge t < u \wedge \neg B\neg\varphi(u) \rightarrow D\varphi(u)$$

The first two are versions of principles (9.2) and (9.4) in the book (Stenning and van Lambalgen 2008) and the third is similar to principle (9.11) in that book. In order to make the formalization more compact, and also more in the spirit of natural deduction style, we do not take the principles as axiom schemas, but instead we turn them into the following proof-rules.

$$\frac{S\varphi}{B\varphi}\,(\mathcal{R}1) \qquad \frac{D\varphi}{B\varphi}\,(\mathcal{R}2) \qquad \frac{B\varphi(t) \quad t < u \quad \neg B\neg\varphi(u)}{D\varphi(u)}\,(\mathcal{R}3)$$

The second and third proof-rule together formalizes a "principle of inertia" saying that a belief is preserved over time, unless there is belief to the contrary.

We liberalize the side-condition on the (Term) rule such that the formulas $\varphi_1, \ldots, \varphi_n$, and ψ may include formulas on the form $t < u$, since we assume that the truth-values of such formulas are not changed by the perspective shift effected by the rule. With this machinery in place, the shift of person perspective in the Sally-Anne task can be formalized as

$$\dfrac{@_a S\,p(t_0) \quad t_0 < t_1 \quad @_a\neg B\neg p(t_1) \qquad \dfrac{[a] \qquad \dfrac{\dfrac{\dfrac{[a] \quad [@_a S\,p(t_0)]}{S\,p(t_0)}\,(\mathcal{R}1)}{Bp(t_0)} \qquad [t_0 < t_1] \qquad \dfrac{[a] \quad [@_a\neg B\neg p(t_1)]}{\neg B\neg p(t_1)}}{Dp(t_1)}\,(\mathcal{R}3)}{Bp(t_1)}\,(\mathcal{R}2)}{@_a Bp(t_1)}\,(\mathcal{T}erm)$$

where we have omitted names of introduction and elimination rules for the satisfaction operator. Recall that this derivation is meant to formalize the child's reasoning at the

time t_1 where the question is answered. The first premise $@_a S\, p(t_0)$ in the derivation says that Sally (the reference the nominal a) at the earlier time t_0 saw that the marble was in the basket, which the child remembers. The third premise $@_a \neg B \neg p(t_1)$ says that Sally at the time t_1 does not believe that the marble is not in the basket, which the child realizes as Sally was absent when the marble was transferred to the box.

Note that the actual position of the marble at the time t_1 is irrelevant to figure out the correct response. Note that in the Sally-Anne task there is a shift of person perspective which we deal with in a modal-logical fashion letting points of evaluation stand for persons, like in the person version of the Smarties task in the previous section, but there is also a temporal shift in the Sally-Anne task, from the time t_0 to the time t_1, which we deal with using first-order machinery.

8 Discussion

In the introduction of the present paper we remarked that reasoning in Seligman's system is different from reasoning in the most common proof systems for hybrid logic, and that reasoning in Seligman's system captures well the reasoning in the Smarties and Sally-Anne tasks, in particular the involved shift between different local perspectives.

More can be said about this difference between the proof systems and how local perspectives are (or are not) represented. A truth-bearer is an entity that is either true or false. According to Peter Simons' paper (Simons 2006), there have historically been two fundamentally opposed views of how truth-bearers have their truth-values.

> One view takes truth to be absolute: a truth-bearer's truth-value (whether truth or falsity) is something it has *simpliciter*, without variation according to place, time, by whom and to whom it is said. The other view allows a truth-bearer's truth-value to vary according to circumstances: typically time or place, but also other factors may be relevant. (Simons 2006, p. 443)

Peter Simons calls the first view the *absolute* view and the second the *centred* view. It is well-known that Arthur Prior often expressed sympathy for what is here called the centred view, most outspoken with respect to time, one reason being that he wanted to allow statements to change truth-value from one time to another. What a truth-bearer's truth-value varies according to, is by Simons called a *location*.

> I understand 'location' broadly to include not just spatial location but also temporal location, spatiotemporal location, modal location, and more

> broadly still location in any relational structure. I consider that the concept of an object being located at a position among other positions is a formal concept, applicable topic-neutrally in any field of discourse. This means that logical considerations about location are not limited in extent or parochial in interest. (Simons 2006, p. 444)

The proposition expressed in the quotation above is defended in Simons' paper (Simons 2004). See also the paper (Simons 2003). Obviously, a frame for modal and hybrid logic is a mathematically precise formulation of Simons' concept of a location, see Definition 2.1.

What does all this have to do with proof systems for hybrid logic? The distinction between the absolute view and the centred view is useful for describing proof systems and the formulas that occur in them. The basic building blocks of the most common proof systems for hybrid logic are satisfaction statements, and satisfaction statements have constant truth-values, so the basic building blocks of such systems are absolute, although it is arguable that such systems have both absolute and centred features since arbitrary subformulas of satisfaction statements do have varying truth-values, and therefore have to be evaluated for truth at some location. On the other hand, the basic building blocks of Seligman's system are arbitrary formulas, and arbitrary formulas have varying truth-values, so this system is centred, involving local perspectives in the reasoning.

9 Some remarks on other work

Beside analysing the reasoning taking place when giving a correct answer to a reasoning task, the works by van Lambalgen and co-authors also analyse what goes wrong when an incorrect answer is given. We note that Stenning and van Lambalgen (2008) warn against simply characterizing autism as a lack of theory of mind. Rather than being an explanation of autism, Stenning and van Lambalgen see the theory of mind deficit hypothesis as "an important label for a problem that needs a label" (cf. Stenning and van Lambalgen 2008, p. 243). Based on their logical analysis, they argue that another psychological theory of autism is more fundamental, namely what is called the *executive function deficit theory*. Very briefly, executive function is an ability to plan and control a sequence of actions with the aim of obtaining a goal in different circumstances.

The paper (Pijnacker et al. 2009) reports empirical investigations of closed-world reasoning in adults with autism. Incidentally, according to the opening sentence of that paper, published in 2009, "While autism is one of the most intensively researched psychiatric disorders, little is known about reasoning skills of people with autism."

With motivations from the theory of mind literature, the paper (van Ditmarsch and Labuschagne 2007) models examples of beliefs that agents may have about other agents' beliefs (one example is an autistic agent that always believes that other agents have the same beliefs as the agent's own). This is modelled by different agents preference relations between states, where an agent prefers one state over another if the agent considers it more likely. The beliefs in question turn out to be frame-characterizable by formulas of epistemic logic.

The paper (Flobbe et al. 2008) reports empirical investigations of what is called *second-order* theory of mind, which is a person's capacity to imagine other people's beliefs about the person's own beliefs (where *first-order* theory of mind is what we previously in the present paper just have called theory of mind). The investigations in (Flobbe et al. 2008) make use of a second-order false-belief task, as well as other tasks.

The paper (Gierasimczuk et al. 2012) does not deal with false-belief tasks or theory of mind, but it is nevertheless relevant to mention since it uses formal proofs to compare the cognitive difficulty of deductive tasks. To be more precise, the paper associates the difficulty of a deductive task in a version of the Mastermind game with the minimal size of a corresponding tableau tree, and it uses this measure of difficulty to predict the empirical difficulty of game-plays, for example the number of steps actually needed for solving a task.

The method of reasoning in tableau systems can be seen as attempts to construct a model of a formula: A tableau tree is built step by step using rules, whereby more and more information about models for the formula is obtained, and either at some stage a model can be read off from the tableau tree, or it can be concluded that there cannot be such a model (in fact, in the case of Gierasimczuk et al. 2012, any formula under consideration has exactly one model, so in that case it is a matter of building a tableau tree that generates this model). Hence, if the building of tableau trees is taken to be the underlying mechanism when a human is solving Mastermind tasks, then the investigations in (Gierasimczuk et al. 2012) can be seen to be in line with the mental models school (see the third section of the present paper).

A remark from a more formal point of view: The tableau system described in (Gierasimczuk et al. 2012) does not include the cut-rule[8]. Much has been written on the size of proofs in cut-free proof systems, in particular, the paper (Boolos 1984) gives examples of first-order formulas whose derivations in cut-free systems are much larger than their derivations in natural deduction systems, which implicitly allow unrestricted cuts (in one case more than 10^{38} characters compared to less than 3280 characters). Similarly, the paper (D'Agostino and Mondadori 1994) points out that ordinary cut-

[8]The cut-rule says that the end of any branch in a tableau tree can extended with two branches with φ on the one branch and $\neg\varphi$ on the other (expressing the bivalence of classical logic).

free tableau systems have a number of anomalies, one of them being that for some classes of propositional formulas, decision procedures based on cut-free systems are much slower than the truth-table method (in the technical sense that there is no polynomial time computable function that maps truth-table proofs of such formulas to proofs of the same formulas in cut-free tableau systems). Instead of prohibiting cuts completely, the paper (D'Agostino and Mondadori 1994) advocates allowing a restricted version of the cut-rule, called the analytic cut-rule.

10 Future work

We would like to extend the work of the present paper to further false-belief tasks, perhaps using different hybrid-logical machinery (and moreover, to see if we can also use hybrid-logical proof-theory to analyse what goes wrong when incorrect answers are given). Not only will formalization of further reasoning tasks be of interest on their own, but we also expect that such investigations can be feed back into logical research, either as corroboration of the applicability of existing logical constructs, or in the form of new logical constructs, for example new proof-rules or new ways to add expressive power to a logic.

We are also interested in further investigations in when two seemingly dissimilar reasoning tasks have the same underlying logical structure, like we in the present paper have disclosed that two different versions of the Smarties task have exactly the same underlying logical structure. Such investigations might be assisted by a notion of identity on proofs (exploiting the longstanding effort in proof-theory to give a notion of identity between proofs, that is, a way to determine if two arguments have common logical structure, despite superficial dissimilarity).

More speculatively, we expect that our formalizations can contribute to the ongoing debate between two dominating views on theory of mind, denoted *theory-theory* and *simulation-theory*. According to theory-theory, theory of mind should be viewed as an explicit theory of the mental realm of another person, like the theories of the physical world usually going under the heading "naive physics", whereas according to simulation-theory, theory of mind should be viewed as a capacity to put yourself in another person's shoes, and simulate the person's mental states.

Acknowledgements The author thanks Thomas Bolander for comments on an early version of this paper. Also thanks to Jerry Seligman for a discussion of the paper. The author acknowledges the financial support received from The Danish Natural Science Research Council as funding for the project HYLOCORE (Hybrid Logic, Computation, and Reasoning Methods, 2009–2013).

References

C. Areces and B. ten Cate. Hybrid logics. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, pages 821–868. Elsevier, 2007.

M. Baaz and A. Leitsch. *Methods of Cut-Elimination*, volume 34 of *Trends in Logic Series*. Springer, 2011.

S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, 1995.

S. Baron-Cohen, A. Leslie, and U. Frith. Does the autistic child have a 'theory of mind'? *Cognition*, 21:37–46, 1985.

G. Bierman and V. de Paiva. On an intuitionistic modal logic. *Studia Logica*, 65: 383–416, 2000.

G. Boolos. Don't eliminate cut. *Journal of Philosophical Logic*, 13:373–378, 1984.

T. Braüner. Two natural deduction systems for hybrid logic: A comparison. *Journal of Logic, Language and Information*, 13:1–23, 2004.

T. Braüner. *Hybrid Logic and its Proof-Theory*, volume 37 of *Applied Logic Series*. Springer, 2011.

T. Braüner. Hybrid-logical reasoning in false-belief tasks. In B. Schipper, editor, *Proceedings of Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 186–195, 2013.

M. D'Agostino and M. Mondadori. The taming of the cut. Classical refutations with analytical cut. *Journal of Logic and Computation*, 4:285–319, 1994.

H. van Ditmarsch and W. Labuschagne. My beliefs about your beliefs – a case study in theory of mind and epistemic logic. *Synthese*, 155:191–209, 2007.

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.

G. Gentzen. Investigations into logical deduction. In M. Szabo, editor, *The Collected Papers of Gerhard Gentzen*, pages 68–131. North-Holland Publishing Company, 1969.

N. Gierasimczuk, H. van der Maas, and M. Raijmakers. Logical and psychological analysis of Deductive Mastermind. In *Proceedings of the Logic & Cognition Workshop at ESSLLI 2012, Opole, Poland, 13–17 August, 2012*, volume 883 of *CEUR Workshop Proceedings*, pages 21–39. CEUR-WS.org, 2012.

W. Goldfarb. *Deductive Logic*. Hackett Pub. Co., 2003.

A. Gopnik and J. Astington. 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, 1988.

J. Pijnacker, B. Geurts, M. van Lambalgen, C. Kan, J. Buitelaar, and P. Hagoort. Defeasible reasoning in high-functioning adults with autism: Evidence for impaired exception-handling. *Neuropsychologia*, 47:644–651, 2009.

D. Prawitz. *Natural Deduction. A Proof-Theoretical Study*. Almqvist and Wiksell, Stockholm, 1965.

D. Prawitz. Ideas and results in proof theory. In J. E. Fenstad, editor, *Proceedings of the Second Scandinavian Logic Symposium*, volume 63 of *Studies in Logic and The Foundations of Mathematics*, pages 235–307. North-Holland, 1971.

L. Rips. *The Psychology of Proof: Deductive Reasoning in Human Thinking*. MIT Press, 1994.

J. Seligman. The logic of correct description. In M. de Rijke, editor, *Advances in Intensional Logic*, volume 7 of *Applied Logic Series*, pages 107 – 135. Kluwer, 1997.

P. Simons. Absolute truth in a changing world. In *Philosophy and Logic. In Search of the Polish Tradition. Essays in Honour of Jan Wolenski on the Occasion of his 60th Birthday*, volume 323 of *Synthese Library*, pages 37–54. Kluwer, 2003.

P. Simons. Location. *Dialectica*, 58:341–347, 2004.

P. Simons. The logic of location. *Synthese*, 150:443–458, 2006. Special issue edited by T. Braüner, P. Hasle, and P. Øhrstrøm.

K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science*. MIT Press, 2008.

A. Troelstra and H. Schwichtenberg. *Basic Proof Theory*, volume 43 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1996.

A Proof of analyticity

Usually, when considering a natural deduction system, one wants to equip it with a normalizing set of reduction rules such that normal derivations satisfy the subformula property. Normalization says that any derivation by repeated applications of reduction rules can be rewritten to a derivation which is normal, that is, no reduction rules apply. From this it follows that the system under consideration is analytic.

Now, the works (Braüner 2004) and (Braüner 2011), Section 4.3 by the present author devise a set of reduction rules for $\mathbf{N}'_{\mathcal{H}}$ obtained by translation of a set of reduction rules for a more common natural deduction system for hybrid logic. This more common system, which we denote $\mathbf{N}_{\mathcal{H}}$, can be found in (Braüner 2004) and in (Braüner 2011), Section 2.2. All formulas in the system $\mathbf{N}_{\mathcal{H}}$ are satisfaction statements. Despite other desirable features, it is not known whether the reduction rules for $\mathbf{N}'_{\mathcal{H}}$ are normalizing, and normal derivations do not always satisfy the subformula property. In fact, Chapter 4 of the book (Braüner 2011) ends somewhat pessimistically by exhibiting a normal derivation without the subformula property. It is remarked that a remedy would be to find a more complete set of reduction rules, but the counter-example does not give a clue how such a set of reduction rules should look.

In what follows we shall take another route. We prove a completeness result saying that any valid formula has a derivation in $\mathbf{N}'_{\mathcal{H}}$ satisfying a version of the subformula property. This is a sharpened version of a completeness result for $\mathbf{N}'_{\mathcal{H}}$ originally given in (Braüner 2004) and in Section 4.3 of (Braüner 2011) (Theorem 4.1 in (Braüner 2011)). Thus, we prove that $\mathbf{N}'_{\mathcal{H}}$ is analytic without going via a normalization result. So the proof of the completeness result does not involve reduction rules. The result is mathematically weaker than normalization together with the subformula property for normal derivations, but it nevertheless demonstrates analyticity. Analyticity is a major success criteria in proof-theory, one reason being that analytic provability is a step towards automated theorem proving (which obviously is related to Leibniz' aim mentioned in the intoduction of the present paper).

In the proof below we shall refer to $\mathbf{N}_{\mathcal{H}}$ as well as a translation $(\cdot)^\circ$ from $\mathbf{N}_{\mathcal{H}}$ to $\mathbf{N}'_{\mathcal{H}}$ given in (Braüner 2004) and Section 4.3 of (Braüner 2011). This translates a derivation π in $\mathbf{N}_{\mathcal{H}}$ to a derivation π° in $\mathbf{N}'_{\mathcal{H}}$ having the same end-formula and parcels of undischarged assumptions. The reader wanting to follow the details of our proof is advised to obtain a copy of the paper (Braüner 2004) or the book (Braüner 2011). The translation $(\cdot)^\circ$ satisfies the following.

Lemma 1. *Let π be a derivation in $\mathbf{N}_{\mathcal{H}}$. Any formula θ occuring in π° has at least one of the following properties.*

1. *θ occurs in π.*

2. *$@_a\theta$ occurs in π for some satisfaction operator $@_a$.*

3. *θ is a nominal a such that some formula $@_a\psi$ occurs in π.*

Proof. Induction on the structure of the derivation of π. Each case in the translation $(\cdot)^\circ$ is checked. □

Note that in item 1 of the lemma above, the formula θ must be a satisfaction statement since only satisfaction statements occur in π. In what follows $@_d\Gamma$ denotes the set of formulas $\{@_d\xi \mid \xi \in \Gamma\}$.

Theorem 1. *Let π be a normal derivation of $@_d\varphi$ from $@_d\Gamma$ in $\mathbf{N}_\mathcal{H}$. Any formula θ occuring in π° has at least one of the following properties.*

1. *θ is of the form $@_a\psi$ such that ψ is a subformula of φ, some formula in Γ, or some formula of the form c or $\Diamond c$.*

2. *θ is a subformula of φ, some formula in Γ, or some formula of the form c or $\Diamond c$.*

3. *θ is a nominal.*

4. *θ is of the form $@_a(p \to \bot)$ or $p \to \bot$ where p is a subformula of φ or some formula in Γ.*

5. *θ is of the form $@_a\bot$ or $\bot$.*

Proof. Follows from Lemma 1 above together with Theorem 2.4 (called the quasi-subformula property) in Subsection 2.2.5 of (Braüner 2011). □

We are now ready to give our main result, which is a sharpened version of the completeness result given in Theorem 4.1 in Section 4.3 of (Braüner 2011).

Theorem 2. *Let φ be a formula and Γ a set of formulas. The first statement below implies the second statement.*

1. *For any model $\mathcal{M}$, any world w, and any assignment g, if, for any formula $\xi \in \Gamma$, $\mathcal{M}, g, w \models \xi$, then $\mathcal{M}, g, w \models \varphi$.*

2. *There exists of derivation of φ from Γ in $\mathbf{N}'_\mathcal{H}$ such that any formula θ occuring in the derivation has at least one of the five properties listed in Theorem 1.*

Proof. Let d be a new nominal. It follows that for any model $\mathcal{M}$ and any assignment g, if, for any formula $@_d\xi \in @_d\Gamma$, $\mathcal{M}, g \models @_d\xi$, then $\mathcal{M}, g \models @_d\varphi$. By completeness of the system $\mathbf{N}_{\mathcal{H}}$, Theorem 2.2 in Subsection 2.2.3 of the book (Braüner 2011), there exists a derivation π of $@_d\varphi$ from $@_d\Gamma$ in $\mathbf{N}_{\mathcal{H}}$. By normalization, Theorem 2.3 in Subsection 2.2.5 of the book, we can assume that π is normal. We now apply the rules $(@I)$, $(@E)$, and $(\mathcal{N}ame)$ to π° obtaining a derivation of φ from Γ in $\mathbf{N}'_{\mathcal{H}}$ satisfying at least one of the properties mentioned in Theorem 1. □

Remark: If the formula occurrence θ mentioned in the theorem above is not of one of the forms covered by item 4 in Theorem 1, and does not have one of a finite number of very simple forms not involving propositional symbols, then either θ is a subformula of φ or some formula in Γ, or θ is of the form $@_a\psi$ such that ψ is a subformula of φ or some formula in Γ. This is the version of the subformula property we intended to prove.

Iterating Semantic Automata

Shane Steinert-Threlkeld and Thomas F. Icard, III.

Department of Philosophy, Stanford University
shanest@stanford.edu, icard@stanford.edu

Abstract

The semantic automata framework, developed originally in the 1980s, provides computational interpretations of generalized quantifiers. While recent experimental results have associated structural features of these automata with neuroanatomical demands in processing sentences with quantifiers, the theoretical framework has remained largely unexplored. In this paper, after presenting some classic results on semantic automata in a modern style, we present the first application of semantic automata to polyadic quantification, exhibiting automata for iterated quantifiers. We also discuss the role of semantic automata in linguistic theory and offer new empirical predictions for sentence processing with embedded quantifiers.[1]

1 Introduction

The interpretation of natural language determiner phrases as generalized quantifiers has led to deep and subtle insights into linguistic quantification. While the original goal of interpreting determiner phrases uniformly as higher-order properties is now seen as perhaps too simplistic,[2] the very idea that determiners can be assigned meanings which correctly predict their contribution to the meanings of sentences is of fundamental importance in semantics. Generalized quantifier theory, and arguably model-theoretic

[1] The final version of this paper appeared in *Linguistics and Philosophy*. The final publication is available at link.springer.com. Please use that version for citations.

[2] See Szabolcsi 2009 for an overview of some recent developments in quantifier theory. As she notes (p.5), "these days one reads more about what [generalized quantifiers] cannot do than about what they can."

semantics in general, has largely developed independently of detailed questions about language processing. If one's aim is to understand how language can express truths, abstracting away from language users, then this orientation is arguably justified.[3] However, if the aim is to understand the role quantification plays in human cognition, model-theoretic interpretation by itself is too abstract. Patrick Suppes (1980) aptly summarized the point more than three decades ago:

> "It is a surprising and important fact that so much of language ... can be analyzed at a nearly satisfactory formal level by set-theoretical semantics, but the psychology of users is barely touched by this analysis." (p. 27)

Consider, for instance, the basic question of how a quantified sentence is verified as true or false. Generalized quantifier theory by itself has nothing to say about this question. One may worry that the psychological details would be too complex or unsystematic to admit useful and elegant theorizing of the sort familiar in formal semantics. However, in the particular case of verification, we believe the analysis of quantifier phrases by *semantic automata* provides a promising intermediate level of study between the abstract, ideal level of model theory and the mosaic, low-level details of processing.

Semantic automata, originally pioneered by Johan van Benthem (1986), offer an algorithmic, or procedural, perspective on the traditional meanings of quantifier phrases as studied in generalized quantifier theory. They are thus ideally suited to modeling verification-related tasks. A semantic automaton represents the *control structure* involved in assessing whether a quantified sentence is true or false. While there has been relatively little theoretical work in this area since van Benthem (1986) (though see Mostowski 1991; 1998), a series of recent imaging and behavioral experiments has drawn on semantic automata to make concrete predictions about quantifier comprehension (McMillan et al. 2005; 2006, Szymanik and Zajenkowski 2010a;b). These experiments establish (among other results, to be discussed further below) that working memory is recruited in the processing of sentences involving certain quantifiers, which corresponds to an analogous memory requirement on automata. Such studies provide impetus to revisit the semantic automata framework from a theoretical perspective. In this paper we extend the framework from simple single (monadic) quantifier sentences to sentences involving iterated quantification. This extension in turn raises new questions about processing.

[3] A classic statement of this approach to semantics can be found in (Lewis 1970, p. 170): "I distinguish two topics: first, the description of possible languages or grammars as abstract semantic systems whereby symbols are associated with aspects of the world; and second, the description of the psychological and sociological facts whereby one of these abstract semantic systems is the one used by a person or population. Only confusion comes of mixing these two aspects." Lewis, Montague, and others clearly took the first as their object of study.

In Section 2, we give a quick review of generalized quantifiers, followed by an extended introduction to semantic automata for single quantifier sentences in Section 3. Section 4 includes a more detailed discussion of how semantic automata might fit into semantic theorizing, at a level in between model-theoretic semantics and language processing. Finally, our main technical contribution is in Section 5 where we show how to extend the framework to iterations of quantifiers. A general construction method is given for combining automata for single quantifiers into automata for iterations. We then discuss further open empirical questions and other issues raised by this work.

2 Generalized quantifiers

Definition 2.1 (Mostowski 1957, Lindström 1966)**.** A *generalized quantifier* Q of type $\langle n_1, \ldots, n_k \rangle$ is a class of models $\mathcal{M} = \langle M, R_1, \ldots, R_k \rangle$ closed under isomorphism, where each $R_i \subseteq M^{n_i}$.[4] A generalized quantifier is *monadic* if $n_i = 1$ for all i, and *polyadic* otherwise.

We write $Q_M\ R_1\ \ldots\ R_k$ as shorthand for $\langle M, R_1, \ldots, R_k \rangle \in Q$. Usually the subscripted M is omitted for readability. Thus, e.g., for type $\langle 1, 1 \rangle$ we write $Q\ A\ B$, where A and B are predicates. This connects with the more familiar definition given in linguistic semantics as can be seen by the following examples:

$$\begin{aligned} \mathit{all} &= \{\langle M, A, B \rangle \mid A \subseteq B\} \\ \mathit{some} &= \{\langle M, A, B \rangle \mid A \cap B \neq \emptyset\} \end{aligned}$$

The isomorphism closure condition (partially) captures the intuition that quantifiers are sensitive only to the size of the relevant subsets of M and not the identity of any particular elements or the order in which they are presented.

A useful classification of generalized quantifiers is given by the standard logical hierarchy. We restrict attention to type $\langle 1, 1 \rangle$, and we distinguish only between first-order and higher-order definability.

Definition 2.2. A generalized quantifier Q of type $\langle 1, 1 \rangle$ is *first-order definable* if and only if there is a first-order language $\mathcal{L}$ and an $\mathcal{L}$-sentence φ whose non-logical vocabulary contains only two unary predicate symbols A and B such that for any model $\mathcal{M} = \langle M, A, B \rangle$,

$$Q_M AB \quad \Leftrightarrow \quad \langle M, A, B \rangle \models \varphi.$$

[4] For more complete introductions to the theory of generalized quantifiers see Barwise and Cooper 1981, van Benthem 1986, Westerståhl 1989, Keenan 1996, Keenan and Westerståhl 2011.

The generalization to higher-order (non-first order) definability is obvious. As examples, *all*, *some*, and *at least three* are first-order definable:

$$\begin{aligned} all_M AB &\Leftrightarrow \langle M, A, B\rangle \models \forall x\,(Ax \rightarrow Bx)\,; \\ some_M AB &\Leftrightarrow \langle M, A, B\rangle \models \exists x\,(Ax \wedge Bx)\,; \\ \textit{at least three}_M AB &\Leftrightarrow \langle M, A, B\rangle \models \exists x, y, z\ \varphi(x, y, z), \end{aligned}$$

where $\varphi(x, y, z)$ is the formula

$$x \neq y \wedge y \neq z \wedge x \neq z \wedge Ax \wedge Bx \wedge Ay \wedge By \wedge Az \wedge Bz.$$

Most, *an even number of*, and *an odd number of* are (only) higher-order definable. For *most*, see, e.g., Appendix C of Barwise and Cooper (1981).

Because the space of type $\langle 1, 1\rangle$ quantifiers places few constraints on possible determiner meanings, several properties have been offered as potential semantic universals, narrowing down the class of possible meanings. These properties seem to hold of (at least a majority of) quantifiers found in natural languages. Two of these will play a pivotal role in the development of semantic automata, due to their role in Theorem 1 below:

CONS $Q_M AB$ iff $Q_M A(A \cap B)$.

EXT $Q_M AB$ iff $Q_{M'} AB$ for every $M \subseteq M'$.

Lemma 1. *A quantifier Q satisfies* **CONS** + **EXT** *iff, for all $\mathcal{M} = \langle M, A, B\rangle$:*

$$Q_M AB \Leftrightarrow Q_A A(A \cap B).$$

Theorem 1. *A quantifier Q satisfies* **CONS** *and* **EXT** *if and only if for every M, M' and $A, B \subseteq M$, $A', B' \subseteq M'$, if $|A - B| = |A' - B'|$ and $|A \cap B| = |A' \cap B'|$, then $Q_M AB \Leftrightarrow Q_{M'} A'B'$.*

Proof. Suppose Q satisfies **CONS** and **EXT**. If $|A-B| = |A'-B'|$ and $|A\cap B| = |A'\cap B'|$, then we have bijections between the set differences and intersections which can be combined to give a bijection from A to A'. Thus $Q_A A(A \cap B) \Leftrightarrow Q_{A'} A'(A' \cap B')$ by isomorphism closure. By two applications of Lemma 1, $Q_M AB \Leftrightarrow Q_{M'} A'B'$.

In the other direction, for any given $\langle M, A, B\rangle$, let $M' = A' = A$ and $B' = A \cap B$. The assumption yields $Q_M AB \Leftrightarrow Q_{M'} A'B' \Leftrightarrow Q_A A(A\cap B)$, which by Lemma 1 implies **CONS** + **EXT**. □

In other words, quantifiers that satisfy **CONS** and **EXT** can be summarized succinctly as binary relations on natural numbers. Given Q we define:

$$Q^c_M xy \quad \Leftrightarrow \quad \exists\, A, B \subseteq M \ \text{ s.t. } \ Q_M AB \text{ and } |A - B| = x, |A \cap B| = y.$$

Standard generalized quantifiers can thus be seen as particular simple cases.

$$\begin{aligned} every^c_M\ xy &\Leftrightarrow x = 0 \\ some^c_M\ xy &\Leftrightarrow y > 0 \\ at\ least\ three^c_M\ xy &\Leftrightarrow y \geq 3 \\ most^c_M\ xy &\Leftrightarrow y > x \\ an\ even\ number\ of^c_M\ xy &\Leftrightarrow y = 2n \text{ for some } n \in \mathbb{N} \end{aligned}$$

Theorem 1 guarantees that the relation Q^c is always well defined.

2.1 Iterating monadic quantifiers

To handle sentences such as

(1) (a) One of our neighbors stole all but four of the sunflowers.
(b) Three explorers discovered most of the islands.

in which quantified phrases appear both in object position and subject position, we need to look at so-called polyadic lifts of monadic quantifiers. Intuitively, these sentences express complex properties of the respective transitive verbs. Since these verbs take two arguments, it will be impossible to give truth-conditions using monadic predicates.

In particular, we will need iterations of type $\langle 1, 1\rangle$ quantifiers. For notation, if R is a binary relation, we write

$$R_x = \{y \mid Rxy\}$$

If Q_1 and Q_2 are type $\langle 1, 1\rangle$, then $It(Q_1, Q_2)$ will be of type $\langle 1, 1, 2\rangle$, defined:

$$It(Q_1, Q_2)\ A\ B\ R \quad \Leftrightarrow \quad Q_1\ A\ \{x \mid Q_2\ B\ R_x\}$$

We will sometimes use the alternative notation $Q_1 \cdot Q_2$ for $It(Q_1, Q_2)$.[5]

Sentences with embedded quantifiers can be formalized as iterations. For instance, the sentences in (1) would typically be assigned truth conditions in (2):

(2) (a) $It(some, all\_but\_four)$ *neighbor sunflowers stole*
(b) $It(three, most)$ *explorers islands discovered*

For example, (2) (a) holds iff

$$neighbor \cap \{x \mid all\_but\_four\ sunflowers\ stole_x\} \neq \emptyset$$

In Section 5 we will show how to define automata corresponding to such iterations. But first, in the next section, we introduce automata for single quantifier sentences.

[5]This is a special case of a general definition for iterating quantifiers. For details, see Chapter 10 of Peters and Westerståhl 2006.

3 Semantic automata

Throughout this section all quantifiers are assumed to satisfy **CONS** and **EXT**.[6] The basic idea behind semantic automata is as follows: given a model $\mathcal{M} = \langle M, A, B\rangle$ and an enumeration of A, we define a string in $s \in \{0,1\}^*$ by assigning 0 to elements of $A \setminus B$ and 1 to $A \cap B$. Note that we can use any enumeration of A since quantifiers are closed under isomorphism. To ensure that these strings are finite, we consider only finite models. It then follows by Theorem 1 that

$$\mathcal{M} \in Q \quad \Leftrightarrow \quad \langle \#_0(s), \#_1(s)\rangle \in Q^c,$$

where $\#_0$ and $\#_1$ are recursively defined functions yielding the number of zeros and of ones in a string, respectively. The goal is to define machines that correspond to quantifiers, in the sense that they accept exactly the strings encoding models in the denotation of the quantifier. The *language of* Q is the set

$$\mathcal{L}_Q = \{s \in \{0,1\}^* \mid \langle \#_0(s), \#_1(s)\rangle \in Q^c\}.$$

Definition 3.1. Let $\mathcal{M} = \langle M, A, B\rangle$ be a model, $\vec{a}$ an enumeration of A, and $n = |A|$. We define $\tau(\vec{a}, B) \in \{0,1\}^n$ by

$$(\tau(\vec{a}, B))_i = \begin{cases} 0 & a_i \in A \setminus B \\ 1 & a_i \in A \cap B \end{cases}$$

Thus, τ defines the string corresponding to a particular finite model.

Lemma 2. *If a quantifier Q satisfies* **CONS** *and* **EXT**, *then the language $\mathcal{L}_Q$ is permutation-invariant. In other words, if $\langle M, A, B\rangle \in Q$, then $\tau(\vec{a}, B) \in \mathcal{L}_Q$ for all enumerations $\vec{a}$ of A.*

Proof. Membership in $\mathcal{L}_Q$ depends only on the number of ones and zeros in a string, neither of which is affected by permutations.

□

The simplest class of automata are the (deterministic) finite-state automata:[7]

Definition 3.2. A *deterministic finite-state automaton (DFA)* is a tuple $\langle Q, \Sigma, \delta, q_0, F\rangle$:

[6] These assumptions can be dropped. But with them, we can use a two-letter alphabet in defining our languages below. Without them, we would need a four-letter alphabet. Moreover, little to no coverage of natural language determiners is lost by making these assumptions.

[7] For a canonical reference on automata theory, see Hopcroft et al. 2001.

- Q a finite set of *states*
- $\delta : Q \times \Sigma \to \Sigma$ a *transition function*
- $F \subseteq Q$ the set of *accepting states*
- Σ a finite set of *input symbols*
- q_0 the *start state*

We denote the components of a DFA M by $Q(\mathsf{M})$, $\Sigma(\mathsf{M})$, etc.

A DFA can be given a simple graphical representation. Each state $q \in Q$ corresponds to a node. We will often represent these as circles, omitting the name of the node. Final states (another name for accepting states) in F will be represented as double circles. If $\delta(q, a) = p$, we draw a directed arc from q to p labeled by a. For semantic automata our alphabet Σ will always be $\{0, 1\}$.

For an example, consider *every*. Recall that $every^c_M\ xy \Leftrightarrow x = 0$. Thus, the only words $w \in \{0, 1\}^*$ that should be accepted are those where $\#_0(w) = 0$. So the automaton for *every* will start in an accepting state and stay there so long as it only reads 1s. As soon as it reads a 0, however, the automaton moves to a non-accepting state and remains there. We represent *every* by the DFA in Figure 1.

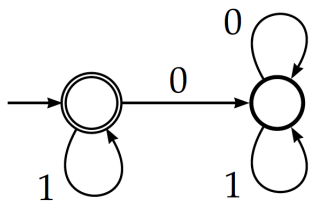


Figure 1: A finite state automaton for *every*

Here is a toy example: $A = \{a, b\}$, $B = \{a, b, c\}$. In this case, A will be represented by the string 11 and the DFA for *every* will start and stay in the accepting state. If, on the other hand, $B = \{a, c\}$, then A will be represented by the string 10. Upon reading 0, the DFA for *every* will move to the non-accepting state and end there.

Recall that $some^c_M\ xy \Leftrightarrow y > 0$. Therefore, a DFA for *some* should accept any word $w \in \{0, 1\}^*$ which contains at least one 1 (i.e. such that $\#_1(w) > 0$). It is easily verified that the DFA depicted in Figure 2 will do just that.

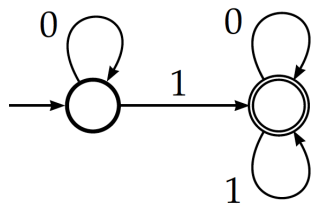


Figure 2: A finite state automaton for *some*

One can prove that all first-order definable quantifiers have languages that can be accepted by DFAs. Motivated by this result, one might hope that only first-order definable quantifiers can be so simulated. It turns out, however, that some higher-order definable quantifiers can also be modeled by finite state automata. Figure 3 shows such an automaton for *an even number of*, which is not first-order definable.

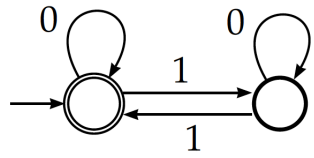


Figure 3: A cyclic finite state automaton for *an even number of*

Switching the end-state from the node on the left to the node on the right renders Figure 3 a (cyclic) finite state automaton for *an odd number of*. In general, the first-order definable quantifiers correspond to a smaller class of DFAs:

Theorem 2 (van Benthem 1986, p.156-157)**.** *A quantifier Q is first-order definable iff $\mathcal{L}_Q$ can be recognized by a permutation-invariant acyclic finite state automaton.*

Moreover, it is not the case that all higher-order quantifiers can be simulated by cyclic finite-state automata. Mostowski (1991) (see also Mostowski 1998) has characterized the class of quantifiers which can. The type $\langle 1 \rangle$ *divisibility quantifier* D_n is defined:

$$\langle M, A \rangle \in D_n \quad \text{iff} \quad |A| \text{ is divisible by } n.$$

Theorem 3 (Mostowski 1991)**.** *Finite state automata accept exactly the class of quantifiers of type* $\langle 1, \ldots, 1\rangle$ *definable in first-order logic augmented with* D_n *for all n.*

To simulate quantifiers such as *less than half* or *most*, neither of which is definable in divisibility logic, we must move to the next level of the machine hierarchy, to pushdown automata.[8] Intuitively, a pushdown automaton augments a DFA with a stack of memory; this stack is a last-in/first-out data structure, onto which we can "push" new content, and from which we can "pop" the topmost element.

Definition 3.3. A (non-deterministic) *pushdown automaton (PDA)* is a tuple $\langle Q, \Sigma, \Gamma, \delta, q_0, Z_0, F\rangle$:

- Q is a finite set of *states*
- Γ is a finite *stack alphabet*
- q_0 is the *start state*
- Z_0 is the *start symbol*
- Σ is a finite set of *input symbols*
- $\delta : Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma \to \mathcal{P}(Q \times \Gamma^*)$ is a *transition function*
- F is the set of *accepting states*

The biggest difference between DFAs and PDAs lies in the transition function. The idea is that δ receives as input the current state, the symbol most recently read, and the symbol at the top of the stack. An output pair (p, γ) indicates that the automaton has moved to state p and replaced the top of the stack with the string γ. Suppose X is the symbol at the top of the stack. If $\gamma = \epsilon$ (here ϵ denotes the empty string), then the top of the stack has been popped. If $\gamma = X$, then no change has been made. If $\gamma = YZ$, then X has been replaced with Z and Y has been pushed onto the stack. While the definition of a PDA allows for any length string to be pushed, we will incidentally work only with strings of length 2.

Graphically, we represent $\delta(q, a, X) = (p, \gamma)$ by a directed arc from q to p labeled by $a, X/\gamma$. Here X/γ is intended to signify that symbol X has been replaced by string γ at the top of the stack. We use x as a variable over characters in an alphabet in order to consolidate multiple labels. In all of the following examples, we assume $\Gamma = \Sigma \cup \{\epsilon\}$.

Figure 4 depicts a PDA for *less than half*. The idea here is that we push 1s and 0s to

[8] In particular, we are moving one level up the Chomsky (1959) hierarchy of formal grammars. The *regular* languages are generated by DFAs, while the *context-free* languages are generated by pushdown automata. That quantifiers such as *most* and *less than half* are not computable by DFAs can be easily proven using the Pumping Lemma for regular languages.

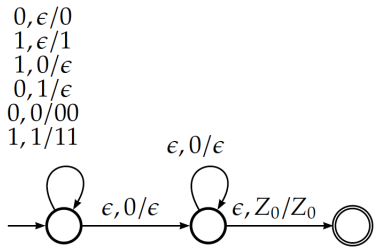


Figure 4: A pushdown automaton for *less than half*

the stack as often as we can, but popping off pairs. If a 1 is read and a 0 is on the stack, we pop the 0 and *vice versa*. This has the affect of "pairing" the members of $A \setminus B$ and $A \cap B$. Because *less than half* holds when the former outnumber the latter, there should be only 0s on the stack at the end of this process. Therefore, the transition to the accepting state occurs only when the string of 0s and 1s has been entirely processed and popping off any remaining 0s exhausts the contents of the stack. Modifying the labels on the edges of the final two states to pop all 1s would render this a PDA for *most*.

The final result reported in this section depends on one final definition.

Definition 3.4. A quantifier Q is *first-order additively definable* if there is a formula φ in the first-order language with equality and an addition symbol $\dot{+}$ such that

$$Q^c_M ab \quad \Leftrightarrow \quad \langle \mathbb{N}, \dot{+}, a, b \rangle \models \varphi(a, b)$$

Theorem 4 (van Benthem 1986, p.163-165)**.** *$\mathcal{L}_Q$ is computable by a pushdown automaton if and only if Q is first-order additively definable.*

4 Automata and processing

How exactly does this work on automata relate to questions about processing? On one hand, the machine representations of quantifiers discussed in the previous section are inspired directly by the standard model-theoretic meanings assumed in classical quantifier theory. On the other hand, the fine structure of these representations promises a potential bridge to lower level questions about language processing. In this section we discuss two important dimensions of this connection:

(1) Semantic automata suggest a relatively clean theoretical separation between semantic competence and performance errors.

(2) Recent imaging and behavioral experiments have revealed that structural differences in automata may be predictive of neuroanatomical demands in quantifier processing.

4.1 Explaining performance errors

The logical theory of generalized quantifiers is occasionally dismissed as being irrelevant to the psychological facts about how people produce and assess quantified sentences, typically by citing data that suggests human reasoning with quantifiers does not match the patterns assumed in standard logical analysis. Indeed, experiments reveal people diverge from standard logical prescriptions, not only in reasoning tasks such as the classical syllogism (see, e.g. Chater and Oaksford 1999), but even in simple verification tasks (see, e.g. McMillan et al. 2005) with a similar pattern in (Szymanik and Zajenkowski 2010a)). One could take this to show, or at least to reinforce the idea, that logical/truth-conditional semantics and the psychology of language are best kept separate, with the former studying abstract normative aspects of meaning and the latter studying the actual mechanisms involved in processing.[9] Yet the normative aspects of meaning, and of quantifier meanings in particular, are clearly relevant to questions about how people use quantifiers, and *vice versa*. While a complete discussion of this issue is beyond the scope of this paper, we would like to point out that semantic automata have the potential to serve as a natural bridge. In principle, they allow for a separation between abstract *control structure* involved in quantifier verification and innumerable other variables that the framework leaves underspecified: order of evaluation, predication judgments, domain restriction, and any contextual or practical factors that might affect these and other variables. This could be viewed as a modest distinction between *competence* and *performance* for quantifier expressions.[10]

We might hypothesize that competence with a particular quantifier involves, at least in part, internalizing the right abstract computational mechanism for verification, in particular that given by the appropriate automaton. How precisely verification is implemented on a given occasion will depend on many factors quite independent from

[9] Recall the quotation from Lewis 1970 in Footnote 3 above.

[10] Our suggestion is compatible with many interpretations of what this distinction comes to. For instance, the relatively non-committal interpretation of Smolensky (1988) says competence of a system or agent is "described by hard constraints" which are violable and hold only in the ideal limit, with unbounded time, resources, and other enabling conditions. The actual implementational details are to be given by "soft constraints" at a lower level, which have their own characteristic effects on performance.

the meanings of quantifiers: prototype effects, saliency effects, time constraints, and so on. Consider, for instance, how one might verify or falsify a sentence like (1):

All U.S. presidents have served at least one year in office. (1)

Supposing one does not already know whether this is true or false, but that an answer must be computed using information stored in memory, one might check famous presidents like George Washington or Abraham Lincoln first and only then move to less salient presidents. Alternatively, if one actually knew James Garfield or William Harrison had served less than a year, such information might be retrieved quickly without first checking more salient presidents. The subtleties of such strategies are fascinating, but arguably go beyond the meanings of quantifiers. Indeed, they arise in tasks and phenomena having nothing to do with quantifiers.

The same can be said for cases where a search is terminated too soon. In example (1) a person might think about a few salient examples and conclude that the sentence is true. For a particular task it may take too long to think about all forty-four presidents, or it may not be worth the effort, and a quick guess is sufficient. However, even in such cases, upon being shown a counterexample, a subject will not insist that the sentence is actually true just because they were unable to identify the counterexample. It is in that way that people are reasonably attuned to the proper, "normative" meanings. Ordinary speakers have a good sense for what needs to be checked to verify a quantified sentence, even if in practice going through the necessary steps is difficult or infeasible. Semantic automata allow separating out control structure from the specific algorithm used to implement the procedure.

In terms of the Marr's famous *levels of explanation* (Marr 1982), the standard model-theoretic semantics of quantification could be seen as a potential computational, or level 1, theory. Semantic automata offer more detail about processing, but, as we have just argued, less than one would expect by a full algorithmic story about processing (level 2), which would include details about order, time, salience, etc. Thus, we might see the semantic automata framework as aimed at level 1.5 explanation,[11] in between levels 1 and 2, providing a potential bridge between abstract model theory and concrete processing details.

[11] Peacocke (1986) coined the term "level 1.5", though in a slightly different context. Soames (1984) independently identified three levels of (psycho)linguistic investigation which appear to (inversely) correlate with Marr's three levels. Pietroski et al. (2009) also articulate the idea that verification procedures for quantifiers (for *most*, specifically) provide a level 1.5 explanation. Independently, the idea of level 1.5 explanation has recently gained currency in Bayesian psychology in relation to *rational process models* (personal communication, Noah Goodman).

4.2 Experimental results

While this separation between competence and performance remains somewhat speculative, recent experimental work has shown that certain structural features of semantic automata are concretely reflected in the neuroanatomical demands on quantifier verification. Recall that certain quantifiers can be computed by memoryless finite state automata whereas others (such as *most*) require a pushdown automaton which has a form of memory in its stack data structure. McMillan et al. (2005) used fMRI to test "the hypothesis that all quantifiers recruit inferior parietal cortex associated with numerosity, while only higher-order quantifiers recruit prefrontal cortex associated with executive resources like working memory." In their study, twelve native English speakers were presented with 120 grammatically simple sentences using a quantifier to ask about a color feature of a visual array. The 120 sentences included 20 instances of 6 different quantifiers: three first-order (*at least 3*, *all*, *some*) and three higher-order (*less than half*, *an odd number of*, *an even number of*). Each subject was first shown the proposition alone on a screen for 10s, then the proposition with a visual scene for 2500ms, then a blank screen for 7500ms during which they were to assess whether the proposition accurately portrayed the visual scene.

While behavioral results showed a statistically significant difference in accuracy between verifying sentences with first-order quantifiers and those with higher-order quantifiers, more interesting for our present purposes was the fact that activation in brain regions (dorsolateral prefrontal and inferior frontal cortices) typically linked with executive functioning such as working memory was found only in processing higher-order quantifiers. They concluded that the formal difference between the machines required to compute quantifier languages seems to reflect a physical difference in neuroanatomical demands during quantifier comprehension.

This first piece of evidence does not tell the whole story, however. The first-order vs. higher-order distinction does not map directly on to the distinction between DFAs and PDAs because of parity quantifiers such as *an even number of*, which are computable by cyclic DFAs. Szymanik (2007) observed that McMillan et al. did not make this distinction and began investigating the demands placed on memory by parity quantifiers. In a subsequent study by Szymanik and Zajenkowski (2010b) reaction times were found to be lowest for Aristotelian quantifiers (*all* , *some*), higher for parity (*an even number of*), yet higher for cardinals of high rank (*at least 8*), and highest for proportionality quantifiers (*most*). This provides some evidence that the complexity of the minimal automaton, and not simply the kind of automaton, may be relevant to questions about processing.[12] A subsequent study by Szymanik and Zajenkowski (2011)

[12]Indeed, a general notion of complexity for automata in the context of language processing would be useful in this context (see also the discussion in Section 5.1).

showed that proportionality quantifiers place stronger demands on working memory than parity quantifiers. This result is consistent with the semantic automata picture, since a PDA computing the relevant parity quantifiers never needs to push more than one symbol on to the stack.

Finally, it is also relevant that McMillan et al. (2005) found no significant difference in activation between judgments involving *at least 3* where the relevant class had cardinality near or distant to three. This suggests subject are invoking a precise number sense in such cases. This contrasts with recent studies by Pietroski et al. (2009) and Lidz et al. (2011), which attempt to distinguish which of several verification procedures are actually used when processing sentences with *most*. Although the present paper shares much in spirit with this project, their protocols show a scene for only 150 or 200ms, which effectively forces the use of the *approximate number system*.[13] The finding is that in such cases people do not seem to be using a pair-matching procedure such as that defined above. We conjecture that with more time the pair-matching algorithm would be used, at least approximately.

The experimental work described in this section has been based solely on automata defined for monadic quantifiers. In order to carry this project further, and to understand its potential and its shortcomings as a high-level processing model, we need to define machines appropriate for more complex quantificational types. In the next section we take an important first step in this direction, defining automata for iterations of quantifiers.[14]

5 Iteration

We now show how to define automata for type $\langle 1, 1, 2\rangle$ iterations of quantifiers by composing the automata already defined for $\langle 1, 1\rangle$ quantifiers. Recall:

$$It(Q_1, Q_2)\ A\ B\ R \quad \Leftrightarrow \quad Q_1\ A\ \{x \mid Q_2\ B\ R_x\}$$

Intuitively, we simply want to run the automaton for Q_1 on the string generated by the sets A and $\{x \mid Q_2\ B\ R_x\}$. The trick, however, comes in "generating" this second set on the fly. Our basic maneuver will be this: we run the Q_2 automaton on B and R_{a_i} for every $a_i \in A$. For each run, we push onto a stack a 1 or a 0 corresponding to whether $a_i \in \{x \mid Q_2\ B\ R_x\}$ or not. Then, we run a transformed version of the Q_1 automaton where every transition has been replaced with one that pops symbols off the

[13] See Dehaene 1997 for the distinction between precise and approximate number systems.

[14] Szymanik (2010) investigated the computational complexity of polyadic lifts of monadic quantifiers. This approach, however, deals only with Turing machines. Our development can be seen as investigating the fine structure of machines for computing quantifier meanings.

stack instead of reading them. In this sense, the stack of the iterated machine will serve as the input tape for the Q_1 machine and we will use the Q_2 machine to generate an appropriate string. We will now make this all precise, first working with quantifiers computable by finite state automata, and then generalizing to those computable only by pushdown automata.

Definition 5.1. Let $\mathcal{M} = \langle M, A, B, R\rangle$ be a model, $\vec{a}$ and $\vec{b}$ enumerations of A and B, with $n = |A|,\ m = |B|$. We overload notation (see Definition 3.1) by allowing τ to take a relation as an extra argument:

$$\tau\left(\vec{a}, \vec{b}, R\right) = \left(\tau\left(\vec{b}, R_{a_i}\right) \boxdot\right)_{i \leq n}$$

where $\tau\left(\vec{b}, R_{a_i}\right)$ is the translation given in Definition 3.1. The operation $(\cdot)_{i\leq n}$ concatenates instances of $(\cdot)$ for $0, \ldots, n$. The $\boxdot$ functions as a separator symbol in a way that will shortly be made precise.

To see this translation in a concrete example, consider a model $\langle M, A, B, R\rangle$ where $M = \{x, y, z\}$, $A = \{x, y\}$, $B = M$, and

$$R = \{\langle x, y\rangle, \langle y, x\rangle, \langle y, y\rangle, \langle y, z\rangle\}$$

Let the enumerations $\vec{a}$ and $\vec{b}$ be given alphabetically. Then $\tau\left(\vec{a}, \vec{b}, R\right)$ will be

$$010 \boxdot 111 \boxdot \tag{2}$$

Definition 5.2. Let Q_1 and Q_2 be quantifiers of type $\langle 1, 1\rangle$. We define *the language of* $Q_1 \cdot Q_2$ by

$$\begin{aligned}\mathcal{L}_{Q_1 \cdot Q_2} = \ & \{w \in (w_i \boxdot)^* \mid i \leq n,\ w_i \in \{0,1\}^* \text{ and} \\ & \langle card\left(\{w_i \mid w_i \notin \mathcal{L}_{Q_2}\}\right), card\left(\{w_i \mid w_i \in \mathcal{L}_{Q_2}\}\right)\rangle \in Q_1^c\}\end{aligned}$$

For $w \in (w_i \boxdot)^*$, we write $numsep\,(w)$ for the number of $\boxdot$ symbols in w.

Note that $w \notin \mathcal{L}_Q$ iff $w \in \mathcal{L}_{\neg Q}$ where $\neg Q$ is the outer negation, given by

$$\langle M, A, B\rangle \in \neg Q \Leftrightarrow \langle M, A, B\rangle \notin Q.$$

We illustrate the definition with a few examples.

Example 1. Here are a few examples using *some*, *every*, and *most*. We omit some of the conditions (e.g., that $i \leq n$) to enhance readability.

$$w \in \mathcal{L}_{every \cdot some} \Leftrightarrow \left\langle card\left(\{w_i \mid w_i \in \mathcal{L}_{some}\}\right), card\left(\{w_i \mid w_i \notin \mathcal{L}_{some}\}\right)\right\rangle \in every^c$$

$$\begin{aligned}
&\Leftrightarrow card\left(\{w_i \mid w_i \notin \mathcal{L}_{some}\}\right) = 0 \\
&\Leftrightarrow card\left(\{w_i \mid \#(1) = 0\}\right) = 0 \\
w \in \mathcal{L}_{some\cdot every} &\Leftrightarrow card\left(\{w_i \mid \#(0) = 0\}\right) > 0 \\
w \in \mathcal{L}_{most\cdot some} &\Leftrightarrow card\left(\{w_i \mid \#(1) > 0\}\right) > card\left(\{w_i \mid \#(1) = 0\}\right)
\end{aligned}$$

One can see that the translation given in (2) will be in $\mathcal{L}_{some\cdot every}$.

As before we have the following relationship between translations of models and languages.

Proposition 1. *Let $\mathcal{M} = \langle M, A, B, R\rangle$ be a model and Q_1, Q_2 quantifiers of type $\langle 1, 1\rangle$. Then for any enumerations $\vec{a}$ and $\vec{b}$ of A and B,*

$$\tau\left(\vec{a}, \vec{b}, R\right) \in \mathcal{L}_{Q_1\cdot Q_2} \Leftrightarrow \langle M, A, B, R\rangle \in It(Q_1, Q_2)$$

5.1 Iterating finite state machines

In order to define PDAs that accept these iterated languages, we first need to define the aforementioned transformation on DFAs which allows a stack to be treated as if it were input.

Definition 5.3. Let M be a DFA. The *pushdown reader of* M, M^p is defined by

- $Q(\mathsf{M}^p) = Q(\mathsf{M})$, $q_0(\mathsf{M}^p) = q_0$, $F(\mathsf{M}^p) = F(\mathsf{M})$;
- $\Sigma(\mathsf{M}^p) = \emptyset$;
- $\Gamma(\mathsf{M}^p) = \Sigma(\mathsf{M})$;
- $\delta(\mathsf{M}^p) = \{\langle q_1, \epsilon, r, \epsilon, q_2\rangle \mid \langle q_1, r, q_2\rangle \in \delta(\mathsf{M})\}$.

In other words, the stack alphabet of the pushdown reader is the input alphabet of the original automaton. The state spaces are the same, but each transition $q_1 \xrightarrow{r} q_2$ in M is replaced by $q_1 \xrightarrow{\epsilon, r/\epsilon} q_2$, i.e. by popping an r from the stack. On its own, a pushdown reader is a fairly meaningless machine since its input alphabet is empty. But they will prove to be a critical component in the PDAs which compute iterated quantification.

Before defining the iteration automata, we provide several helper definitions. For an automaton M, let the *sign of* $q \in Q(\mathsf{M})$ be given by

$$sgn(q) = \begin{cases} 1 & q \in F \\ 0 & q \notin F \end{cases}$$

We define the *sign of* M as

$$sgn(\mathsf{M}) = sgn(q_0(\mathsf{M}))$$

In what follows, the complement operator $(\cdot)^c : \{0,1\} \to \{0,1\}$ maps 1 to 0 and 0 to 1. With these definitions in hand, we can proceed to the central definition of this paper.

Definition 5.4 (Iteration Automaton)**.** Let Q_1 and Q_2 be two DFAs accepting $\mathcal{L}_{Q_1}$ and $\mathcal{L}_{Q_2}$, respectively. The PDA $\mathsf{It}(\mathsf{Q}_1, \mathsf{Q}_2)$ is given by:

- $Q = \{q_I\} \cup Q\left(\mathsf{Q}_1^p\right) \cup Q(\mathsf{Q}_2)$
- $\Sigma = \{0, 1, \boxdot\}$
- $\Gamma = \{0, 1\}$
- Transition function:

$$\begin{aligned}
\delta = {} & \delta\left(\mathsf{Q}_1^p\right) \\
& \cup \{\langle q_I, \varepsilon, x, sgn(\mathsf{Q}_2)\, x, q_0(\mathsf{Q}_2)\rangle \mid i \leq n\} \\
& \cup \{\langle q_1, 1, x, x, q_2\rangle \mid \langle q_1, 1, q_2\rangle \in \delta(\mathsf{Q}_2) \text{ and } sgn(q_1) = sgn(q_2)\} \\
& \cup \{\langle q_1, 0, x, x, q_2\rangle \mid \langle q_1, 0, q_2\rangle \in \delta(\mathsf{Q}_2) \text{ and } sgn(q_1) = sgn(q_2)\} \\
& \cup \{\langle q_1, 1, x, x^c, q_2\rangle \mid \langle q_1, 1, q_2\rangle \in \delta(\mathsf{Q}_2) \text{ and } sgn(q_1) \neq sgn(q_2)\} \\
& \cup \{\langle q_1, 0, x, x^c, q_2\rangle \mid \langle q_1, 0, q_2\rangle \in \delta(\mathsf{Q}_2) \text{ and } sgn(q_1) \neq sgn(q_2)\} \\
& \cup \{\langle q, \boxdot, x, x, q_I\rangle \mid q \in Q(\mathsf{Q}_2)\} \\
& \cup \left\{\langle q_I, \varepsilon, x, x, q_0\left(\mathsf{Q}_1^p\right)\rangle\right\}
\end{aligned}$$

- $q_0 = q_I$
- $F = F\left(\mathsf{Q}_1^p\right)$

The basic idea is as follows: q_I is a new start state. From q_I, we have a ε transition to the start state of Q_2. When we take such a transition, a 1 or a 0 is pushed onto the stack according to whether or not the start state of Q_2 is an accepting state. The role of *sgn* and $(\cdot)^c$ is to ensure that we switch the original symbol pushed onto the stack by the i transition whenever we go from an accepting to a non-accepting state or *vice versa*. In this way, we push exactly one symbol on to the stack for each visit to Q_2: a 1 if it ended in an accepting state and a 0 if not. The $\boxdot$ transitions from each state of Q_2 to q_I enable $\boxdot$ to function as a separating symbol. From q_I, we can also take an ε-transition to Q_1^p; this pushdown reader will then process the stack generated by the visits to Q_2.

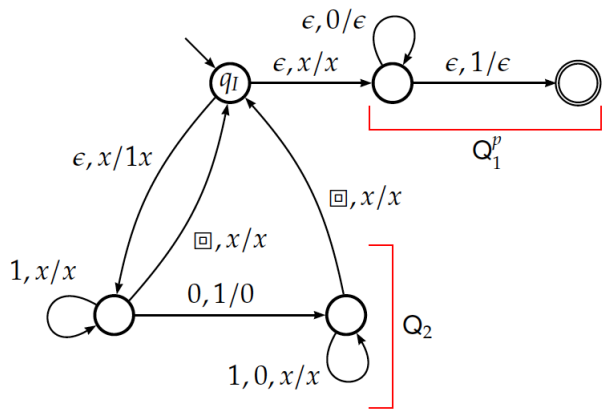


Figure 5: A pushdown automaton for *some A R every B*

Example 2. In Figure 2 is a PDA for computing $some \cdot every$.

Here, Q_1^p is the pushdown reader (see Definition 5.3) of some. Q_2 is the transformed copy of every. Note that we push a 1 onto the stack on the transition from q_I to the start state of every since $q_0(\mathsf{every}) \in F(\mathsf{every})$. Similarly, we pop this 1 and push a 0 on the 0 transition since this goes from an accepting to rejecting state of every.

Consider our earlier string

$$010 \boxdot 111 \boxdot$$

which we know to be in $\mathcal{L}_{some \cdot every}$. When reading 010, this automaton will push a 0 on to the stack, but will push a 1 on to the stack when it reads 111. Thus, some^p will accept the stack input and so the whole string will be accepted.

We record here a basic fact about iterated machines which follows straightforwardly from the definition, and which will be important shortly.

Observation 1. $\mathsf{It}(\mathsf{Q}_1, \mathsf{Q}_2)$ has $1 + |Q(\mathsf{Q}_1)| + |Q(\mathsf{Q}_2)|$ states.

While the informal description of $\mathsf{It}(\mathsf{Q}_1, \mathsf{Q}_2)$ and the example make it seem plausible that this PDA accepts the right iterated language, we now make this equivalence precise. First, we prepare a few preliminary results, for which a basic definition of the notion of computation in a PDA is required.

Definition 5.5. Given a PDA M, a triple $\langle q, w, X\rangle \in Q \times \Sigma^* \times \Gamma^*$ is called an *instantaneous description* of M, specifying the current state, what of the input has not been read, and the current stack contents. The transition function defines a notion of one-step computation: for every $\langle q_1, x, A, q_2, X\rangle \in \delta$, we write

$$\langle q_1, xw, AY\rangle \vdash_{\mathsf{M}} \langle q_2, w, XY\rangle$$

for every $w \in \Sigma^*$ and $Y \in \Gamma^*$, with $\vdash_{\mathsf{M}}^*$ the reflexive, transitive closure of $\vdash_{\mathsf{M}}$.

Intuitively, $\langle q_1, w_1 w_2, AY\rangle \vdash_{\mathsf{M}}^* \langle q_2, w_2, XY\rangle$ means that there is a sequence of transitions starting in q_1 which reads w_1, ends in q_2 and changes the stack from AY to XY.

Lemma 3. *Let Q_1 and Q_2 be quantifiers corresponding to regular languages and $w_i \in \{0,1\}^*$. Abbreviate $\mathsf{It}(\mathsf{Q}_1, \mathsf{Q}_2)$ by M.*

(1) If $w_i \in \mathcal{L}_{Q_2}$, then $(q_I, w_i \boxdot w, X) \vdash_{\mathsf{M}}^* (q_I, w, 1X)$ for any $X \in \Gamma^*,\ w \in \Sigma^*$.

(2) If $w_i \notin \mathcal{L}_{Q_2}$, then $(q_I, w_i \boxdot w, X) \vdash_{\mathsf{M}}^* (q_I, w, 0X)$ for any $X \in \Gamma^*,\ w \in \Sigma^*$.

In other words, for any string w_i, there is a $w_i\boxdot$ path through the iterated PDA such that a 1 or 0 is pushed onto the stack according to whether or not $w_i \in \mathcal{L}_{Q_2}$.

Proof. We prove (1) by induction on the length of w_i. The proof for (2) is wholly analogous. Assume $w_i \in \mathcal{L}_{Q_2}$.

If $|w_i| = 0$ (i.e. $w_i = \epsilon$), then $q_0(\mathsf{Q}_2) \in F(\mathsf{Q}_2)$, i.e. $sgn(\mathsf{Q}_2) = 1$. Thus, we take the $\epsilon, X/1X$ transition from q_I to $q_0(\mathsf{Q}_2)$, immediately followed by the $\boxdot, x/x$ transition back to q_I.

For the inductive step, let $|w_i| = n$ and write $w_i = w_i^- c_i$ where $c_i \in \{0,1\}$. By assumption, Q_2 accepts w_i, so its w_i sequence of transitions ends in some $q_{w_i} \in F(\mathsf{Q}_2)$. We need to check two cases: $w_i^- \in \mathcal{L}_{Q_2}$ or $w_i^- \notin \mathcal{L}_{Q_2}$.

If $w_i^- \in \mathcal{L}_{Q_2}$, then by the inductive hypothesis, there is a $w_i^-\boxdot$ sequence in the iterated PDA sending X to $1X$. Moreover, by assumption, $sgn\left(q_{w_i^-}\right) = sgn(q_{w_i})$. Thus, we replace the $\boxdot$ transition in the $w_i^-\boxdot$ sequence with the $c_i, x/x$ transition from $q_{w_i^-}$ to q_{w_i} that is given by definition of $\delta(\mathsf{It}(\mathsf{Q}_1, \mathsf{Q}_2))$.

If $w_i^- \notin \mathcal{L}_{Q_2}$, then by the inductive hypothesis, there is a $w_i^-\boxdot$ sequence in the iterated PDA sending X to $0X$. This time, by assumption, $sgn\left(q_{w_i^-}\right) \neq sgn(q_{w_i})$. Thus, we replace the $\boxdot$ transition in the $w_i^-\boxdot$ sequence with the $c_i, 0/1$ transition from $q_{w_i^-}$ to q_{w_i} that is given by definition of $\delta(\mathsf{It}(\mathsf{Q}_1, \mathsf{Q}_2))$.

□

Corollary 1. *Let* $w \in (w_i\boxdot)^*$ *where* $w_i \in \{0,1\}^*$. *Then* $(q_I, w, Z_0) \vdash_{\mathsf{M}}^* (q_I, \epsilon, X)$ *for some* X *with* $|X| = numsep(w)$.

Theorem 5. *Let* Q_1 *and* Q_2 *be quantifiers corresponding to regular languages. The language accepted by* $\mathsf{It}(\mathsf{Q}_1, \mathsf{Q}_2)$ *is* $\mathcal{L}_{Q_1 \cdot Q_2}$.

Proof. First, $\mathcal{L}_{Q_1 \cdot Q_2} \subseteq L(\mathsf{It}(\mathsf{Q}_1, \mathsf{Q}_2))$. In particular, we show by induction on $n = numsep(w)$ for $w \in \mathcal{L}_{Q_1 \cdot Q_2}$ that $(q_I, w, Z_0) \vdash_{\mathsf{M}}^* (q_I, \epsilon, X)$ where

$$card(\{w_i \mid w_i \in \mathcal{L}_{Q_2}\}) = \#_1(X)$$
$$card(\{w_i \mid w_i \notin \mathcal{L}_{Q_2}\}) = \#_0(X)$$

The inclusion then follows immediately from the definition of the pushdown reader Q_1^p and the $\epsilon, x/x$ transition from q_I to $q_0\left(\mathsf{Q}_1^p\right)$. If $n = 0$, then $w = \epsilon$. This is in $\mathcal{L}_{Q_1 \cdot Q_2}$ only if $\langle 0, 0\rangle \in Q_1^c$. One can easily check that this entails $q_0(\mathsf{Q}_1) \in F(\mathsf{Q}_1)$, so the ϵ transition from q_I to $q_0\left(\mathsf{Q}_1^p\right)$ takes us to an accepting state of M.

For the inductive step, assume $n > 0$. Write $w = w^- w_i\boxdot$. By inductive hypothesis, we have a w^- path from q_I to q_I generating a stack X such that

$$card(\{w_i \mid w_i \in w^- \text{ and } w_i \in \mathcal{L}_{Q_2}\}) = \#_1(X)$$
$$card(\{w_i \mid w_i \in w^- \text{ and } w_i \notin \mathcal{L}_{Q_2}\}) = \#_0(X)$$

By Lemma 3, there is a w sequence from q_I to q_I which generates a stack $1X$ or $0X$ depending on whether $w_i \in \mathcal{L}_{Q_2}$ or not, exactly as desired.

For the $\mathcal{L}_{Q_1 \cdot Q_2} \supseteq L(\mathsf{It}(\mathsf{Q}_1, \mathsf{Q}_2))$ inclusion, consider $w \in L(\mathsf{M})$. Because $F(\mathsf{M}) = F\left(\mathsf{Q}_1^p\right)$ and Q_1^p only pops from the stack, there must be a w sequence from q_I to q_I generating a stack that contains a word accepted by Q_1. Because the only transitions leaving q_I are ε and the only ones back to q_I are $\boxdot, x/x$, w must be of the form $(w_i\boxdot)^*$. That $w \in \mathcal{L}_{Q_1 \cdot Q_2}^n$ then follows by $numsep(w)$ applications of Lemma 3 and by inspection of Definition 5.3. □

Although we have demonstrated that languages for iterating two quantifiers whose languages can be processed by finite state automata can be processed by pushdown automata, a natural question arises: can these iterated languages also be accepted by finite state automata? In other words, are the regular languages closed under iteration? We show the answer to be positive.

First, note the following fact about regular expressions.[15]

[15] In fact, the substitutions we are defining are known in the compilers literature as *regular definitions*. See, for instance, Aho et al. 2006, Ch. 3.

Lemma 4. *If E is a regular expression in alphabet $\{a_1, \ldots, a_n\}$ and E_i is a regular expression in alphabet Σ_i, then $E[a_1/E_1, \ldots, a_n/E_n]$ is a regular expression in $\bigcup_{i \leq n} \Sigma_i$.*

Definition 5.6. Let E_1 and E_2 be regular expressions in $\{0, 1\}$. We define the iterated regular expressions $It(E_1, E_2)$ by

$$It(E_1, E_2) = E_1\left[0/\left(E_2^c\boxdot\right), 1/\left(E_2\boxdot\right)\right]$$

where E^c denotes a regular expression for the complement of the language generated by E (recall the regular languages are closed under complement).

From Lemma 4 we know that $It(E_1, E_2)$ is a regular expression for every n. Inspection of the above definition makes the following closure result obvious.

Proposition 2. *Let Q_1 and Q_2 be quantifiers with regular languages; write E_{Q_1} and E_{Q_2} for a regular expression generating $\mathcal{L}_{Q_1}$ and $\mathcal{L}_{Q_2}$. Then $It(E_{Q_1}, E_{Q_2})$ generates $\mathcal{L}_{Q_1 \cdot Q_2}$. In other words, $\mathcal{L}_{Q_1 \cdot Q_2}$ is a regular language whenever both $\mathcal{L}_{Q_1}$ and $\mathcal{L}_{Q_2}$ are.*

Note on processing

Already in the single quantifier case, certain quantifiers like *an even number of* have both DFA and PDA representations. It has been suggested, with supporting evidence (Szymanik and Zajenkowski 2010a), that working memory is solicited when processing sentences containing such quantifiers. This provides *prima facie* reason to believe that working memory will be recruited when processing sentences with multiple quantifiers each computable by a DFA. This would show that the PDA representation more closely resembles the actual processing mechanism.

One argument in support of the DFA representation could derive from Fact 1 about the size of the PDAs. Both $\mathsf{It}(\mathsf{some}, \mathsf{every})$ and $\mathsf{It}(\mathsf{every}, \mathsf{some})$ have five states. It can be shown, however, that the minimal DFA accepting $\mathcal{L}_{some \cdot every}$ has four states and that the minimal DFA accepting $\mathcal{L}_{every \cdot some}$, depicted in Figure 6, has three states, strictly fewer than the associated PDAs.

This smaller state space and the apparent superfluity of the stack may favor the DFA representation for such iterations. On the other hand, the PDA construction provides a general method for generating a machine for the iteration of any two quantifiers. There appears to be no such analogously general mechanism for generating the minimal DFAs. Because neither argument on its own can be conclusive, empirical investigation should be done to see how much (if any) working memory is activated in these and similar cases.

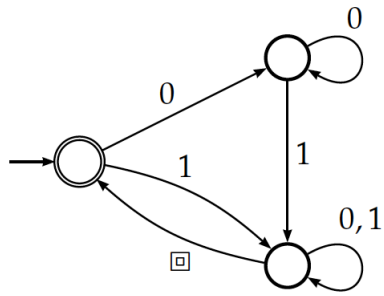


Figure 6: The minimal DFA accepting the language for *every A R some B*

5.2 Iterating with one or more pushdown automata

We now consider the case where one or both of the quantifiers in the iteration defines a non-regular, context-free language. Note that Definition 5.2 and Proposition 1 still apply in this situation. To define machines accepting these iterated languages, we proceed as before by adding a stack to act as an input tape to a pushdown reader. Because one or both of the machines being iterated may in fact be a pushdown automaton, we must generate a two-stack pushdown automaton.

Definition 5.7. A *two-stack pushdown automaton* is exactly like a pushdown automaton except that the transition function is now of the form

$$\delta : Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma \times \Gamma \to \mathcal{P}(Q \times \Gamma^* \times \Gamma^*)$$

We depict a transition $\delta(q, a, X_1, X_2) = (p, \gamma_1, \gamma_2)$ by a directed arc from q to p, labeled by $a, X_1/\gamma_1, X_2/\gamma_2$.

Definition 5.8. Let M be a PDA. The *pushdown reader of* M, M^p, is a two-stack pushdown automaton defined by

- $Q(\mathsf{M}^p) = Q(\mathsf{M})$, $q_0(\mathsf{M}^p) = q_0$, $F(\mathsf{M}^p) = F(\mathsf{M})$
- $\Sigma(\mathsf{M}^p) = \emptyset$
- $\Gamma(\mathsf{M}^p) = \Sigma(\mathsf{M})$
- $\delta(\mathsf{M}^p) = \{\langle q_1, \epsilon, X, r, q_2, \gamma, \epsilon\rangle \mid \langle q_1, r, X, q_2, \gamma\rangle \in \delta(\mathsf{M})\}$

In the pushdown reader, all $r, X/\gamma$ transitions in the original machine become $\epsilon, X/\gamma, r/\epsilon$ transitions. Definition 5.4 of iterated machines easily generalizes when one (or both) of Q_1 and Q_2 is a PDA. The construction is nearly identical, merely keeping track of the extra stack. An example will make this clearer.

Example 3. Figure 7 depicts a two-stack PDA accepting $\mathcal{L}^n_{most \cdot some}$.

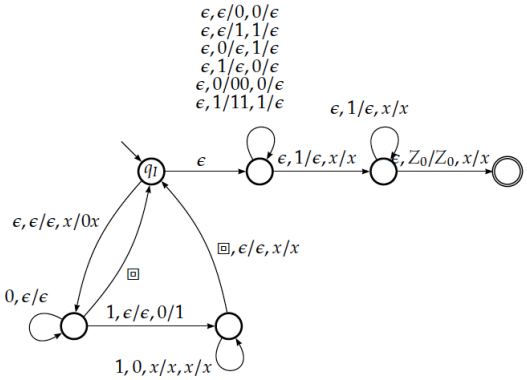


Figure 7: A pushdown automaton for *most A R some B*

As a model of computation, two-stack PDAs have the same power as Turing machines, in that every Turing machine can be simulated by a two-stack PDA and *vice versa*. One might wonder whether our machines really need the power of two stacks or whether, as was the case with the regular languages, the added stack can be eliminated while accepting the same language. In other words, the question arises of whether the context-free languages are closed under iteration (both with regular languages and with other context-free languages).

Again, the answer is positive. The approach directly mirrors that for showing closure of regular languages, replacing 1s and 0s in the language of Q_1 by words in the (appropriately indexed) language of Q_2. First, a simple observation:

Observation 2. For a quantifier Q, if $\mathcal{L}_Q$ is context-free, then so too is $\mathcal{L}_{\neg Q}$.

Proof. Let φ be a formula in first-order additive arithmetic defining Q (as given by Theorem 4). Then $\neg\varphi$ defines $\neg Q$.

□

Theorem 6. *If $\mathcal{L}_{Q_1}$ and $\mathcal{L}_{Q_2}$ are context-free, then so is $\mathcal{L}_{Q_1 \cdot Q_2}$.*

Proof. Let G_1, G_2, and $\overline{G_2}$ be context-free grammars (CFGs) in alphabet $\{0, 1\}$ generating $\mathcal{L}_{Q_1}$, $\mathcal{L}_{Q_2}$, and $\mathcal{L}_{\neg Q_2}$ respectively. Let $n > 0$. We will construct a CFG G in alphabet $\{0, 1, \boxdot\}$ generating $\mathcal{L}_{Q_1 \cdot Q_2}$.

The start symbol of G is the start symbol of G_1. We add a copy of G_2 and call its start symbol S_2. Similarly for $\overline{G_2}$. We then have two new production rules:

$$W \to S_2 \boxdot$$
$$\overline{W} \to \overline{S_2}$$

G simply contains the rules of $G_1\left[0/\overline{W}, 1/W\right]$ and the production rules outlined in the previous paragraph. That G generates $\mathcal{L}_{Q_1 \cdot Q_2}$ follows *via* an analogous argument to that given in the proof of Proposition 2; intuitively, W generates some $w_i \in \mathcal{L}_{Q_2}$ and $\overline{W}$ some $w \in \mathcal{L}^i_{\neg Q_2}$.

□

Example 4. The grammar below generates exactly $\mathcal{L}^n_{most \cdot some}$:

$$\begin{aligned}
M &\to MWM \mid M'WM \mid MWM' \mid M'WM' \\
M' &\to WM'\overline{W} \mid \overline{W}M'W \mid \varepsilon \\
W &\to S_2 \boxdot \\
\overline{W} &\to \overline{S_2} \boxdot \\
S_2 &\to S'_2 1 S'_2 \\
S'_2 &\to S'_2 S'_2 \mid 1 \mid 0 \mid \varepsilon \\
\overline{S_2} &\to 0\overline{S_2} \mid \varepsilon
\end{aligned}$$

Just as one can algorithmically generate a DFA from a regular expression, so too can a PDA be generated from a CFG. There is not, however, an analog of the minimal DFA in the case of PDAs. It is also less clear what repercussions this closure result would have for issues of language processing.

6 Conclusion

Our main contribution in this paper is the first extension of the semantic automata framework to polyadic quantification. After presenting the classical results in this area, and discussing general issues in the connection between automata and processing, we

showed how to model sentences with iterated quantification. We also showed that the regular and context-free languages are closed under iteration in a precise sense. This is the first step in gaining a better understanding of how these machine models might relate more generally to uses of quantifiers in natural language. At this point a number of new questions suggestion themselves for further investigation. Some of these include:

- The extension of semantic automata to iterated quantification gives rise to new empirical predictions, as mentioned above. These predictions should be tested, and more detailed predictions should be explored.
- In light of the discussion in Section 4, it is easy to imagine probabilistic automata or other extensions, reflecting either biases or specific algorithmic verification strategies. This could allow more detailed process models that could be subject to more precise behavioral experimentation.
- Having defined automata for iteration, it is natural to consider other polyadic lifts: resumption, cumulation, branching.[16]
- It may be possible to define automata for irreducibly polyadic quantifiers, which could allow another angle on understanding the elusive Frege boundary (van Benthem 1989, Keenan 1992), through semantic automata.
- A related theoretical question concerns whether minimal DFAs for iterated quantifiers may or must contain non-trivial cycles. This can be phrased more precisely by asking whether these regular languages have non-zero star height (see McNaughton and Papert 1971).
- The close relationship between quantifiers and formal languages in the semantic automata framework allows some results from mathematical linguistics to be used to address semantic issues. For instance, semantic learning may be study in the context of the learnability in the limit framework (Gold 1967). First steps in this direction have been taken by Gierasimczuk (2009) and Clark (2011).

We hope to pursue these questions in future work.

Acknowledgements We thank Johan van Benthem, Christopher Potts, and Jakub Szymanik for helpful discussions.

[16] As a reviewer pointed out, we can already define a machine for cumulation as the sequential composition of $\mathsf{It}(\mathsf{Q}_1, \mathsf{some})$ and $\mathsf{It}(\mathsf{Q}_2, \mathsf{some})$.

References

A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman. *Compilers: Principles, Techniques, and Tools*. Prentice Hall, 2nd edition, 2006.

J. Barwise and R. Cooper. Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, 4(2):159–219, 1981.

J. van Benthem. *Essays in Logical Semantics*. D. Reidel Publishing Company, Dordrecht, 1986.

J. van Benthem. Polyadic quantifiers. *Linguistics and Philosophy*, 12(4):437–464, Aug. 1989.

N. Chater and M. Oaksford. The Probability Heuristics Model of Syllogistic Reasoning. *Cognitive Psychology*, 38(2):191–258, Mar. 1999.

N. Chomsky. On Certain Formal Properties of Grammars. *Information and Control*, 2(2):137–167, June 1959.

R. Clark. On the Learnability of Quantifiers. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 911–923. Elsevier B.V., second edition, 2011.

S. Dehaene. *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, Oxford, 1997.

N. Gierasimczuk. Identification through Inductive Verification Application to Monotone Quantifiers. In P. Bosch, D. Gabelaia, and J. Lang, editors, *7th International Tbilsi Symposium on Logic, Language, and Computation, TbiLLC 2007*, volume 5422 of *Lecture Notes in Artificial Intelligence*, pages 193–205. 2009.

E. M. Gold. Language identification in the limit. *Information and Control*, 10(5): 447–474, May 1967.

J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison Wesley, Boston, 2nd edition, 2001.

E. L. Keenan. Beyond the Frege Boundary. *Linguistics and Philosophy*, 15(2):199–221, 1992.

E. L. Keenan. The Semantics of Determiners. In S. Lappin, editor, *The Handbook of Contermporary Semantic Theory*, number 1989, chapter 2, pages 41–63. Blackwell, Oxford, 1996.

E. L. Keenan and D. Westerståhl. Generalized Quantifiers in Linguistics and Logic. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 859–910. Elsevier, second edition, 2011.

D. Lewis. General semantics. *Synthese*, 22:18–67, 1970.

J. Lidz, P. Pietroski, J. Halberda, and T. Hunter. Interface transparency and the psychosemantics of most. *Natural Language Semantics*, 19(3):227–256, Apr. 2011.

P. Lindström. First order predicate logic with generalized quantifiers. *Theoria*, 32: 186–195, Dec. 1966.

D. Marr. *Vision*. Freeman, San Francisco, 1982.

C. T. McMillan, R. Clark, P. Moore, C. Devita, and M. Grossman. Neural basis for generalized quantifier comprehension. *Neuropsychologia*, 43(12):1729–1737, Jan. 2005.

C. T. McMillan, R. Clark, P. Moore, and M. Grossman. Quantifier comprehension in corticobasal degeneration. *Brain and Cognition*, 62(3):250–260, Dec. 2006.

R. McNaughton and S. A. Papert. *Counter-free Automata*, volume 65 of *MIT Research Monographs*. The MIT Press, 1971.

A. Mostowski. On a generalization of quantifiers. *Fundamenta Mathematicae*, 44: 12–36, 1957.

M. Mostowski. Divisibility Quantifiers. *Bulletin of the Section of Logic*, 20(2):67–70, 1991.

M. Mostowski. Computational semantics for monadic quantifiers. *Journal of Applied Non-Classical Logics*, 8:107–121, 1998.

C. Peacocke. Explanation in Computational Psychology: Language, Perception and Level 1.5. *Mind and Language*, 1(2):101–123, 1986.

S. Peters and D. Westerståhl. *Quantifiers in Language and Logic*. Clarendon Press, Oxford, 2006.

P. Pietroski, J. Lidz, T. Hunter, and J. Halberda. The Meaning of 'Most': Semantics, Numerosity and Psychology. *Mind and Language*, 24(5):554–585, 2009.

P. Smolensky. On the proper treament of connectionism. *Behavioral and Brain Sciences*, 11:1–23, 1988.

S. Soames. Linguistics and Psychology. *Linguistics and Philosophy*, 7(2):155–179, 1984.

P. Suppes. Procedural Semantics. In R. Haller and W. Grassl, editors, *Language, Logic, and Philosophy: Proceedings of the 4th International Wittgenstein Symposium*, pages 27–35. Hölder-Pichler-Tempsy, Vienna, 1980.

A. Szabolcsi. *Quantification*. Research Surveys in Linguistics. Cambridge University Press, Cambridge, 2009.

J. Szymanik. A comment on a neuroimaging study of natural language quantifier comprehension. *Neuropsychologia*, 45(9):2158–2160, May 2007.

J. Szymanik. Computational complexity of polyadic lifts of generalized quantifiers in natural language. *Linguistics and Philosophy*, 33(3):215–250, Nov. 2010.

J. Szymanik and M. Zajenkowski. Comprehension of simple quantifiers: empirical evaluation of a computational model. *Cognitive Science*, 34(3):521–532, Apr. 2010a.

J. Szymanik and M. Zajenkowski. Quantifiers and Working Memory. *Lecture Notes in Artificial Intelligence*, 6042:456–464, 2010b.

J. Szymanik and M. Zajenkowski. Contribution of working memory in parity and proportional judgments. *Belgian Journal of Linguistics*, 25(1):176–194, Jan. 2011.

D. Westerståhl. Quantifiers in Formal and Natural Languages. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic (vol. 4)*, pages 1–132. D. Reidel Publishing Company, Dordrecht, 1989.

Automata and Complexity in Multiple-Quantifier Sentence Verification

Jakub Szymanik, Shane Steinert-Threlkeld, Marcin Zajenkowski, and Thomas F. Icard, III

Institute for Logic, Language and Computation, University of Amsterdam,
Department of Philosophy, Stanford University,
Faculty of Psychology, University of Warsaw
J.K.Szymanik@uva.nl, Shanest@stanford.edu, Zajenkowski@psych.uw.edu.pl, Icard@stanford.edu

Abstract

We study possible algorithmic models for the picture verification task with double-quantified sentences of the form 'Some X are connected with every Y'. We show that the ordering of quantifiers, either Some ∘ Every or Every ∘ Some, influences the cognitive difficulty of the task. We discuss how computational modeling can account for the varying cognitive load in quantifier verification.

1 Introduction

A central area of cognitive science is the study of our linguistic abilities, including the understanding and evaluation of natural language sentences. Given the richness and the variety of natural language constructions it is almost an impossible task to model those cognitive abilities in their full generality. However, there are some fragments of language which have formal semantics and which are well understood and rigorously described by linguists. Those fragments are good candidates for cognitive computational modeling building upon the formal results. In particular, linguistic expressions of quantities deserve special attention. This is because the study of such expressions (determiner phrases) in the framework of Generalized Quantifier Theory (GQT) marks

one of the most well-developed branches of formal semantics. Recently, it has been shown how GQT can give rise to a computational model delivering neuropsychological predictions on verification tasks (see, e.g, McMillan et al. 2005, Szymanik and Zajenkowski 2010a). However, the model could only account for sentences with a single quantifier, like 'More than 5 boys played the game'. In this paper we discuss an extension of the model that covers sentences with multiple-quantifiers, like 'Some boy kissed every girl'. We also test empirically some predictions of the model about the cognitive complexity of various sentences with embedded quantifiers.

1.1 Generalized quantifiers

Generalized quantifiers (GQs) are one of the basic tools of today's linguistics; their mathematical properties have been extensively studied since the 1950s (see, e.g., Peters and Westerståhl 2006). GQT assigns meanings to statements by defining the semantics of the quantifiers in terms of relations between subsets of the universe. Let us consider sentence ((1)) as an example:

(1) Every poet has low self-esteem.

GQT takes 'every' as a binary relation between (in this case) the set of poets and the set of people having low self-esteem. Following the natural linguistic intuition we will say that sentence ((1)) is true if and only if the set of poets is included in the set of people having low self-esteem. Hence, the quantifier 'every' corresponds in this sense to the inclusion relation.

Mathematically, such notion of GQs may be captured by identifying sentences of the form $\mathsf{Q}AB$ with the situations (models) in which those sentences are true (Lindström 1966). For instance, we want to uniformly express the meaning of 'most' independently from the situation. Let us explain this approach further by giving a few examples. Sentence ((1)) is of the form $\mathsf{Every}\ A$ is B, where A stands for poets and B for people having low self-esteem. As we explained above the sentence is true if and only if $A \subseteq B$. Therefore, the quantifier 'every' corresponds to the class of models (M, A, B) in which $A \subseteq B$. For the same reasons the quantifier 'some' corresponds to the class of models in which the set $A \cap B$ is nonempty. Finally, let us consider the quantifier 'most'. The sentence $\mathsf{Most}\ As$ are B is true if and only if the cardinality of set $(A \cap B)$ is greater than the cardinality of set $(A - B)$. Therefore, formally speaking:

$$\mathsf{Every} = \{(M, A, B) \mid A, B \subseteq M \text{ and } A \subseteq B\}.$$

$$\mathsf{Some} = \{(M, A, B) \mid A, B \subseteq M \text{ and } |(A \cap B)| > 0\}.$$

$$\mathsf{Most} = \{(M, A, B) \mid A, B \subseteq M \text{ and } |(A \cap B)| > |(A - B)|\}.$$

Hence, if we fix a model $\mathbb{M}$, then we can treat a generalized quantifier as a relation between relations over the universe, and this is the familiar notion from natural language semantics. For instance, in a given model $\mathbb{M}$ the statement $\mathsf{Most}_M(A, B)$ says that $|(A^M \cap B^M)| > |(A^M - B^M)|$, where $A^M, B^M \subseteq M$.

The above formalization links tightly with many computational models and results on GQs (see, e.g., Szymanik 2009). Even though it has not been designed as a cognitive model, it can raise precise processing predictions (see Szymanik 2007). In the reminder of the paper we study the computational model of the verification tasks for sentences containing combinations of 'every' and 'some' quantifiers.

2 Model

In the semantic automata approach to GQs, quantifiers are modeled as automata of various types. This works by associating each finite model[1] – corresponding, say, to a visual scene – with a string in a formal language and designing an automaton that accepts a string if and only if the model which was translated into that string makes the quantifier sentence true. As an example, consider Every, the model-theoretic interpretation of which is given above. Given a model $\mathbb{M}$, we can write a word in $\{a, b\}^{|\mathbb{M}|}$ by enumerating A and writing an a for every element in $A \setminus B$, and a b for every element in $A \cap B$.[2] To verify whether $\mathbb{M} \in$ Every, we must then simply check whether the word consists entirely of bs since $A \subseteq B$ iff $A \setminus B = \emptyset$. An automaton that recognizes strings corresponding to models in Some should accept any string in which at least one b appears. These two automata are pictured below in Figure 1.

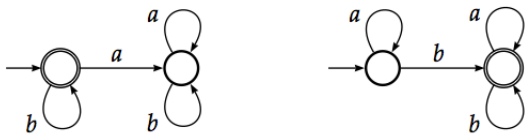


Figure 1: Finite state automata for Every (left) and Some (right)

[1] As we work with natural language quantifier the restriction to finite models is arguably innocent (see Szymanik 2009).

[2] To use a two-letter alphabet, the quantifiers must satisfy certain properties, like topic neutrality, domain independence and conservativity (see, e.g., Peters and Westerståhl 2006). All the quantifiers considered in this study do.

A detailed theory of automata for GQs of this form has been developed by van Benthem (1986) (but see also Mostowski (1998)). Certain quantifiers, such as Most can only be modeled by pushdown automata. These automata add to the ones above a stack data structure, a.k.a. a form of memory. A slew of both neuroimaging and behavioral studies (see, e.g., McMillan et al. 2005, Szymanik and Zajenkowski 2010a;b, Zajenkowski et al. 2011) have shown that this formal distinction is mirrored in the way patients process quantifiers: proportional quantifiers like Most take longer to process, are processed less accurately, and place more demand on working memory than do simpler quantifiers.

Most of the theoretical and experimental work thus far has focused, however, on sentences of the form QAB. But natural language contains the ability to embed quantifiers, as in "some student read every book" or "most students take three classes". These are modeled by *iterated quantifiers*. For instance:

$$\mathsf{Some} \circ \mathsf{Every} = \{(M, A, B, R) \mid (M, A, \{x \in M \mid (M, B, R_x) \in \mathsf{Every}\}) \in \mathsf{Some}\},$$

where $A, B \subseteq M$, $R \subseteq M \times M$, and $R_x = \{y \in M \mid Rxy\}$. Unpacking this definition, we find that "some student read every book" will come out true just when the set of students (x) who are such that every book was read by them (R_x) is non-empty (see Figure 4 for an example).

Szymanik (2010) has studied the computational complexity of the various readings of multi-quantifier sentences, including iterated quantifiers. He identified the border between tractable and intractable constructions that has been later tested experimentally by Schlotterbeck and Bott (2012). It has turned out that the computational differences approximate cognitive complexity of the corresponding verification task. However, the results of Szymanik for multi-quantifier sentences have been formally disconnected from the automata model for simple quantifier sentences. Recently, that gap has been bridged by Steinert-Threlkeld and Icard (forthcoming) who have extended the semantic automata approach to handle cases of iterated quantification. While we refer the reader to the paper for formal details, we here present one example, a machine for Some ∘ Every (Figure 2)[3].

Let us assume that we want to verify sentence 'Some boy reads every book'. The idea is this: for every boy we check the set of books he reads (the string generated by B and R_x). We run the Every automaton on it to check whether that particular boy actually reads all the books. If he does we push a 1 onto the stack, otherwise we push a 0. Next, we move to the books read by the next boy (⧈ separates the corresponding substrings in the encoding) and run the same algorithm. Once we analyzed books read by every boy in that way, we then run the Some machine but using the stack contents

[3] ⧈ is a special separator symbol needed in the encoding of the models.

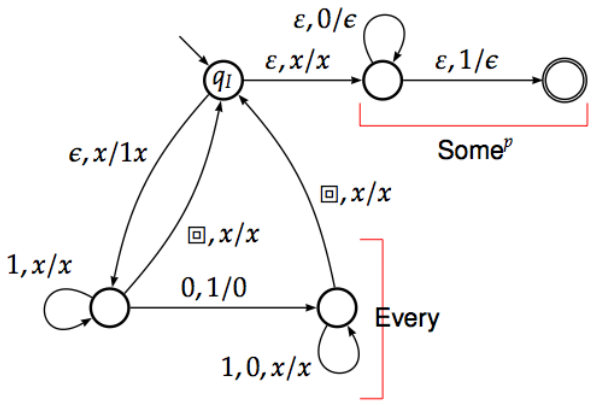


Figure 2: A pushdown automaton for Some ∘ Every

as the tape. So, this machine will reach the accepting state if and only if there is at least one 1 on the stack, which will happen if and only if at least one $x \in A$ is such that the Every automaton accepts the string generated by B and R_x, meaning that there is a boy who reads all books.

This raises an empirical question: do iterated quantifiers place demands on working memory as the pushdown automata model would predict?

2.1 Predictions

We address this question by performing sentence-verification tasks on sentences with Some ∘ Every and Every ∘ Some iterations. First, we must note that it turns out that these iterations can be also modeled by deterministic finite state automata.[4] Figure 3 depicts the minimal DFAs accepting Every ∘ Some and Some ∘ Every.

Nevertheless, we predict that processing iterated quantifiers will place similar demands on working memory to proportional quantifiers. This is because the construction outlined in the previous section provides a general mechanism for converting automata for any two quantifiers into an automaton for their iteration. No such general mechanism is known to exist for generating DFAs for iteration. To see this, compare the two automata in Figure 3. Although the top row of the Some ∘ Every machine does contain a copy of the Every DFA, no such copy of either DFA can be found in Figure 3;

[4] A key result of Steinert-Threlkeld and Icard (forthcoming) is that if Q_1 and Q_2 are computable by deterministic finite state automata, then so too is $Q_1 \circ Q_2$, i.e., a push-down automata from Figure 2 can be actually transformed into a finite-state automata.

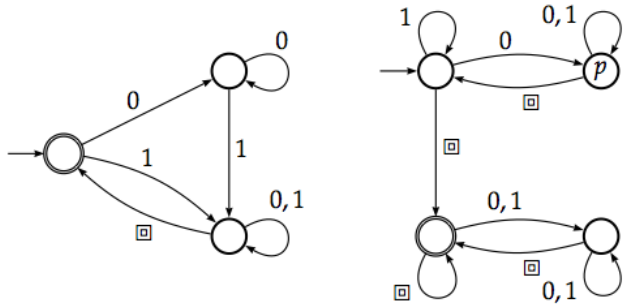


Figure 3: Minimal DFAs accepting the language for Every ∘ Some (left) and Some ∘ Every (right). To compare with PDA let us explain the run of Some ∘ Every automata for the sentence 'Some boy reads every book'. Again, we pick the first boy and look at the books he reads. If he reads all the books we move to accepting state; otherwise, if we find one book he doesn't read we move to state p. From here we can move back to the initial state only by starting to check books read by another boy (marked by ⧈ in the encoding of the model). From the initial state we can go to the double-circled accepting state if and only if we find a boy who reads all books. The bottom row indicates that once we have found one such student, it does not matter whether or not any of the others have read every book.

more importantly, there doesn't appear to be a uniform construction generating the two iterated DFAs. It is plausible that people learn procedures for assessing basic quantifier meanings and then develop a general mechanism for processing embedded quantifiers. Since iteration of quantifiers is one semantic operation which is independent of the two basic quantifiers, we predict that there is a single corresponding mental mechanism for constructing procedures to process iterated quantifiers from basic procedures. This mechanism generates pushdown automata and so we expect to find strong working memory demands when processing iterated quantifiers.

Furthermore, the model allows us to predict that true instances of Every ∘ Some might be more complex to verify than true instances of Some ∘ Every. This is because in the first case subjects have to run through every element of A, whereas in the latter, they needed only find one example; this example might be salient in the image. But even if not, a subject verifying Some ∘ Every can stop processing once he finds one appropriate A and need not continue to the rest. Of course, the model predicts that the situation is opposite for false instances: namely false Every∘Some are simpler to verify than false Some ∘ Every since the former require just finding one counterexample.

3 Experimental results

3.1 Method

To test the theoretical predictions we studied how people verify sentences of the form 'Every X is connected with some Y' and 'Some X is connected with every Y'. We also compared the performance on these sentences with other cognitive tasks. In particular, we measured memory span and cognitive control. According to the multicomponent model of working memory as well as empirical findings, these two cognitive functions reflect central aspects of working memory (see, e.g., Logie 2011). Additionally, we looked at proportional judgments of the form 'More than half of the dots are yellow'. According to the model, proportional sentences are only computable by PDAs and, therefore, they engage working memory.

Participants

Seventy-six Polish-speaking students from University of Warsaw (46 females and 30 males) between 19 and 31 years (mean age was 22.64 years, SD=2.65) were recruited for this experiment. Each subject received a small financial reward for participating.

Materials and procedure

Iterations The task tested how subjects verify two types of sentences against simple pictures (see Figure 4).

Figure 4: Examples of stimuli used in the study. Sentence 'Every circle is connected with some square' is true in situation 1. Sentence 'Some circle is connected with every square' is true in situation 2.

Each sentence was repeated eight times. Half of the trials were true. At the beginning of each trial a sentence was displayed. Subjects had as much time as they needed to read it. Next, a picture was presented, and participants were asked to decide within 20000 ms if the proposition accurately describes the picture. All stimuli were counterbalanced and randomly distributed throughout the experiment. For every sentence we

mesured mean reading time, mean verification time, and accuracy (number of correct answers; maximum=8).

Memory span The memory span task was a computerized version of Sternberg's (1966) short-term memory measure. On each trial of the test, the subjects were presented with a random series of different digits, one at a time, for 300 ms, followed by a blank screen and the test digit. Participants had to decide whether the test digit had appeared in the previously displayed string. Sequences of digits of three lengths (four, six, or eight) were repeated eight times each; hence, there were 24 trials overall. The score was the total of correct responses from all conditions (range 0 to 24).

Cognitive control Cognitive control was measured with the short version of the Attention Networks Test (ANT) designed by Fan et al. (2002).

The authors' starting point was the assumption that the attentional system can be divided into three functionally independent networks: alerting, orienting, and executive control. In the present study we focused on the latter network (the monitoring and resolution of conflict between expectation, stimulus, and response) as an index of cognitive control. In the ANT task, on each trial, the participant has to decide, by pressing a button, whether a central arrow stimulus (the target) points left or right. The target is flanked by distractor stimuli, which may be congruent with the target (arrow points in same direction) or incongruent (arrow points in opposite direction). In each case, two flankers are presented on either side of the target. The control index is calculated by subtracting the RT median of the congruent flanking conditions from the RT median of incongruent flanking conditions.

Proportional judgements This task measured the reaction time and accuracy of proportional judgments, such as 'Less than half of the dots are blue', against color pictures presenting dots. The pictures accompanying sentences differed in terms of the number of objects (15 dots or 17 dots), but not the distance between the cardinalities of two sets of dots (7 vs 8 and 8 vs 9). Within each condition, subjects had to solve eight trials. Half of them were true. Participants were asked to decide, by pressing a button, whether or not the proposition accurately describes the picture. We analyzed mean reaction time (RT) as well as accuracy level (number of correct answers; maximum=8) of each condition.

3.2 Results

Iterations

First we compared the processing of two types of sentences used in the task. ANOVA with type of sentence (2 levels) and statements truth-value (2 levels) as two within-subject factors was used to examine differences in mean verification times and accuracy (see Table 1 and Table 2 for means and standard deviations). The main effect of sentence type was significant indicating that sentences containing quantifiers ordered as every-some were verified significantly longer ($F(1, 75) = 17.01$, $p < 0.001$, η^2=0.19) and less accurately ($F(1, 75) = 22.48$, $p < 0.001$, η^2=0.23) than sentences with some-every order. Moreover, the analysis revealed the significant main effect of the interaction between sentence type and truth-value in case of verification time ($F(1, 75) = 42.02$, $p < 0.001$, η^2=0.36) as well as accuracy ($F(1, 75) = 25.63$, $p < 0.001$, η^2=0.26). Further comparisons among means indicated that true sentences with every-some were processed longer and worse than all other situations. Both false conditions did not differ from one another, and were medium difficult, while true some-every sentences had shortest mean RT and the highest correctness.

Finally, for reading time we analyzed only difference between sentence types. ANOVA reached the tendency level ($F(1, 75) = 2.85$, $p = 0.095$, η^2=0.04) and indicated that participants needed more time for every-some than some-every constructions.

Correlations

Next, we correlated the scores obtained in the iteration verification task with other cognitive measures (see Table 1). Analyzing accuracy, we found that only some-every sentences were highly and positively correlated with scores obtained in the memory task and proportional judgements, while in the case of cognitive control the relationship was negative. The latter result is negative since the high result on control network indicates delay in inhibiting response to competing stimuli, and hence poor executive functioning. Interestingly, similar correlations were obtained between accuracy on proportional judgements and both memory span and control tasks. We also found that the verification times for sentences with two quantifiers are positively associated with the verification times of proportional judgments.

When the correlations are conducted separately for true and false iterated statements, the general pattern of significant correlations remains the same (see Table 2). Specifically, only sentences with some-every order were significantly associated with cognitive control, memory span, and proportional quantifiers. This relationship was independent of truth-value.

Table 1: Means (SD) of all variables and correlations between iteration task and other cognitive measures

	Control	Memory	Prop15 acc	Prop17 acc	Mean (SD)
Every-some read	-.12	-.04	-.14	.02	5019 (2520)
Some-every read	-.05	-.01	-.15	-.07	4738 (2140)
Every-some ver	.11	.11	-.01	.01	2506 (1129)
Some-every ver	.10	-.13	-.02	-.10	2079 (867)
Every-some acc	-.06	.02	.10	-.05	6.48 (1.76)
Some-every acc	-.38**	.29*	.32**	.45**	7.61 (.92)
Control		-.26*	-.33**	-.29*	95.68 (38.22)
Memory			.25*	.30**	20.89 (2.15)
Prop15 acc				.49**	6.86 (1.22)
Prop17 acc					6.88 (1.24)

* $p < 0.05$; ** $p < 0.01$; *Note* Read - reading time; ver - verification time; acc - accuracy; prop15 - proportional quantifiers presented with 15 objects, prop17 - proportional quantifiers presented with 17 objects.

4 Discussion

We have studied the computational model of verifying sentences containing embedded quantifiers. We confirmed the prediction that for true instances Every ∘ Some is harder than Some ∘ Every but we did not find the opposite relation for false instances. Most importantly, while the model suggests that sentences with Some ∘ Every and Every ∘ Some iterations are equally difficult with respect to working memory engagement, we found some differences in subjects' performance: 'Every-some' sentences are more difficult in terms of reaction time and accuracy. On the other hand, only verification of 'some-every' sentences correlates with other tasks engaging working

Table 2: Means (SD) of iterated sentences in true and false conditions, and their correlations with other cognitive measures

	Control	Memory	Prop15 acc	Prop17 acc	Mean (SD)
Every-some ver false	.12	.05	-.02	.05	2231 (870)
Every-some ver true	.09	.20	-.02	-.06	2781 (1569)
Some-every ver false	.03	-.12	.10	.05	2468 (1315)
Some-every ver true	.16	-.20	-.21	-.19	1690 (752)
Every-some acc false	-.06	.18	.03	.03	3.5 (0.80)
Every-some acc true	-.05	-.09	.05	-.11	2.96 (1.35)
Some-every acc false	-.30**	.24*	.23*	.41**	3.72 (0.62)
Some-every acc true	-.38**	.28*	.31**	.36**	3.90 (0.42)

memory resources, like cognitive control and memory span, as well as with accuracy of proportional judgments. Moreover, the latter are also associated with both working memory aspects. These findings point towards an alternative model under which Some ∘ Every gets associated with a canonical push-down automata from Figure 2 and Every ∘ Some iterations are processed with a strategy resembling a finite-state automaton from Figure 3. That could explain, on the one hand, the qualitatively different engagement of working memory in the verification of 'Some X is connected with every Y', and on the other hand, the longer reaction time and higher error-rate in the judgments of 'Every X is connected with some Y'. The idea here would be that even though the push-down automata strategy engages more cognitive resources, it is more effective than the corresponding finite-state automata. A related empirical finding is that the reading time (comprehension) for 'every-some' sentences is longer than for 'some-every' sentences. Therefore, an alternative model should also predict that deriving the push-down automata verification strategy for Some ∘ Every iteration is easier than constructing the finite-state automata strategy for Every ∘ Some iteration. This seems to be a natural direction for future research.

5 Outlook

We think that one of the best strategies for subsequent research would be to embed the formal theory in a proper computational cognitive model or implement it within some cognitive architecture, like ACT-R. The general aim of the project would be to build a psychologically and neurally plausible theory of quantifier meaning and compare it with other proposals, such as Johnson-Laird's mental models (1983) or Clark's Comparison Theory (1976). There are many questions about the correspondence between the formal models of quantifier verification and the cognitive resources the subjects need to use in order to solve the task. Building a computational cognitive model will lead to new experimental predictions that can be consequently tested. Moreover, none of the empirical research so far has looked into actual strategies the subjects are applying in order to verify quantifier sentences. Eye-tracking studies could fill the gap and provide additional data to assess whether models of quantifier verification postulate psychologically plausible strategies. We hope that such tasks could be successfully carried out in a collaboration between cognitive modelers and logicians studying GQT.

6 Conclusions

The paper describes an abstract and purely quantitative model of quantifier verification motivated by logical investigations in the semantics of natural language. From a cognitive computational perspective, this is a sort of conceptual pre-modeling, mathematically delimiting the class of all possible cognitive strategies that could be further implemented in a proper cognitive computational model, giving raise to qualitative predictions. In other words, our approach is to analyze human cognitive behavior by investigating formal computational properties of the task (cf. Marr 1983, Anderson 1990). One could say that our work is positioned between the computational and algorithmic levels of Marr at level 1.5 explanation (see, e.g., Steinert-Threlkeld and Icard 2013, Isaac et al. 2014) We do not specify the actual verification strategies people use but we do more than only formal computational characterization of the task, namely we delimit a class of 'reasonable' strategies. Our toolbox in doing that is modern logic and computation theory which focuses on processes rather than 'logical correctness'. One natural application of this toolbox – that we have explored in the paper – is in estimating cognitive difficulty of a task. We believe that the formal insights logic and computation theory have to offer are instrumental for building plausible cognitive computational models.

Acknowledgements JS acknowledges NWO Veni Grant 639.021.232. The work of MZ was supported by a grant no. 2011/01/D/HS6/01920 funded by the National Science Centre in Poland.

References

J. Anderson. *The Adaptive Character of Thought*. Studies in Cognition. Lawrence Erlbaum, 1990. ISBN 9780805804195.

J. van Benthem. *Essays in Logical Semantics*. D. Reidel, Dordrecht, 1986.

H. Clark. *Semantics and Comprehension*. Mouton, 1976.

J. Fan, B. D. McCandliss, T. Sommer, A. Raz, and M. I. Posner. Testing the Efficiency and Independence of Attentional Networks. *Journal of Cognitive Neuroscience*, 14 (3):340–347, Apr. 2002.

A. Isaac, J. Szymanik, and R. Verbrugge. Logic and complexity in cognitive science. In A. Baltag and S. Smets, editors, *Johan van Benthem on Logical/Informational Dynamics*. Springer, 2014.

P. N. Johnson-Laird. *Mental Models: Toward a Cognitive Science of Language, Inference and Consciousness*. Harvard University Press, 1983.

P. Lindström. First order predicate logic with generalized quantifiers. *Theoria*, 32: 186–195, 1966.

R. Logie. The functional organisation and the capacity limits of working memory. *Current Directions in Psychological Science*, 20:240–245, 2011.

D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing Visual Information*. W.H. Freeman, San Francisco, 1983.

C. T. McMillan, R. Clark, P. Moore, C. Devita, and M. Grossman. Neural basis for generalized quantifier comprehension. *Neuropsychologia*, 43:1729–1737, 2005.

M. Mostowski. Computational semantics for monadic quantifiers. *Journal of Applied Non-Classical Logics*, 8:107–121, 1998.

S. Peters and D. Westerståhl. *Quantifiers in Language and Logic*. Clarendon Press, Oxford, 2006.

F. Schlotterbeck and O. Bott. Easy solutions for a hard problem? The computational complexity of reciprocals with quantificational antecedents. In J. Szymanik and R. Verbrugge, editors, *Proceedings of the Logic and Cognition Workshop at ESSLLI 2012*, pages 60–72. CEUR Workshop Proceedings, 2012.

S. Steinert-Threlkeld and I. Icard, ThomasF. Iterating semantic automata. *Linguistics and Philosophy*, 36(2):151–173, 2013.

S. Sternberg. High-speed scanning in human memory. *Science*, 153:652–654, 1966.

J. Szymanik. A comment on a neuroimaging study of natural language quantifier comprehension. *Neuropsychologia*, 45(9):2158–2160, 2007.

J. Szymanik. *Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language*. PhD thesis, University of Amsterdam, Amsterdam, 2009.

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

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

J. Szymanik and M. Zajenkowski. Quantifiers and working memory. In M. Aloni and K. Schulz, editors, *Amsterdam Colloquium 2009, Lecture Notes In Artificial Intelligence 6042*, pages 456–464. Springer, 2010b.

M. Zajenkowski, R. Styła, and J. Szymanik. A computational approach to quantifiers as an explanation for some language impairments in schizophrenia. *Journal of Communication Disorders*, 44(6):595 – 600, 2011.

Using Intrinsic Complexity of Turn-Taking Games to Predict Participants' Reaction Times

Jakub Szymanik, Ben Meijering, and Rineke Verbrugge

Institute for Logic, Language and Computation, University of Amsterdam
Institute of Artificial Intelligence, University of Groningen
J.K.Szymanik@uva.nl, B.Meijering@rug.nl, L.C.Verbrugge@rug.nl

Abstract

We study structural properties of a turn-based game called the Marble Drop Game, which is an experimental paradigm designed to investigate higher-order social reasoning. We show that the cognitive complexity of game trials, measured with respect to reaction time, can be predicted by looking at the structural properties of the game instances. In order to do this, we define complexity measures of finite dynamic two-player games based on the number of alternations between the game players and on the pay-off structure. Our predictions of reaction times and reasoning strategies, based on the theoretical analysis of complexity of Marble Drop game instances, are compared to subjects' actual reaction times. This research illustrates how formal methods of logic and computer science can be used to identify the inherent complexity of cognitive tasks. Such analyses can be located between Marr's computational and algorithmic levels.

1 Introduction

In recent years, questions have been raised about the applicability of logic and computer science to model cognitive phenomena (see, e.g., Frixione 2001, Stenning and Van Lambalgen 2008, Van Rooij 2008). One of the trends has been to apply formal methods to study the complexity of cognitive tasks in various domains, for instance: syllogistic reasoning (Geurts 2003), problem solving (Gierasimczuk et al. 2013), and

natural language semantics (Szymanik and Zajenkowski 2010). It has been argued that with respect to its explanatory power, such analysis can be located between Marr's (1983) computational and algorithmic levels.

More recently, there has also been a trend to focus on similar questions regarding social cognition, more specifically, theory of mind. Especially, higher-order reasoning of the form 'I believe that Ann knows that Peter thinks ...' became an attractive topic for logical analysis (Verbrugge 2009). Here, the logical investigations often go hand in hand with game theory (see, e.g., Osborne and Rubinstein 1994). In this context, one of the common topics among researchers in logic and game theory has been backward induction (BI), the process of reasoning backwards, from the end of the game, to determine a sequence of optimal actions (van Benthem 2002). Backward induction can be understood as an inductive algorithm defined on a game tree. The BI algorithm tells us which sequence of actions will be chosen by agents that want to maximize their own payoffs, assuming common knowledge of rationality. In game-theoretical terms, backward induction is a common method for determining sub-game perfect equilibria in the case of finite extensive-form games.[1]

Games have been extensively used to design experimental paradigms aiming at studying social cognition (Camerer 2003), recently with a particular focus on higher-order social cognition: the matrix game (Hedden and Zhang 2002), the race game (Gneezy et al. 2010, Hawes et al. 2012), the road game (Flobbe et al. 2008, Raijmakers et al. 2013), and the Marble Drop Game (henceforth, MDG) (Meijering et al. 2010; 2011; 2012). All the mentioned paradigms are actually game-theoretically equivalent. They are all finite extensive-form games that can be solved by applying BI. As an example in this paper we will consider MDG (see Fig. 1).

Many studies have indicated that application of higher-order social reasoning among adults is far from optimal (see, e.g., Hedden and Zhang 2002, Verbrugge and Mol 2008). However, Meijering et al. (2010; 2011) report on a near ceiling performance of subjects when their reasoning processes are facilitated by, for example, a step-wise training. Still, an eye-tracking study of the subjects solving the game suggests that backward induction is not necessarily the only strategy used Meijering et al. (2012).

[1] Backward induction is a generalization of the minimax algorithm for extensive form games; the subgame-perfect equilibrium is a refinement of the Nash equilibrium, introduced to exclude equilibria with implausible threats (Osborne and Rubinstein 1994).

We still do not know exactly what reasoning strategies[2] the subjects are applying when playing this kind of dynamic extensive form games. One way to use formal methods to study this question has been proposed by Ghosh et al. (2010), Ghosh and Meijering (2011): to formulate all reasoning strategies in a logical language, and compare ACT-R models based on each reasoning strategy with a subject's actual performance in a sequence of games, based on reaction times, accuracy and eye-tracking data. This corresponds to a study between the computational and algorithmic levels of Marr's Marr (1983) hierarchy.

Here, we aim to tackle the problem from a somewhat more generic, complexity-theoretic viewpoint: we propose to study the problem on the computational level. Specifically, we will identify inherent, structural properties of the game that make certain MDG trials harder than others.

2 Alternation type

Every instance of a finite extensive form game can be presented as a decision tree. The second-order trials of MDG have the abstract tree form presented in Fig. 2.

How to approximate the complexity of a single instance of MDG? In the worst-case scenario, the backward induction algorithm, based on breadth-first search from the leaves of the tree upwards, will have to travel through all the nodes of the decision tree. Thus, it will find the rational solution (Nash Equilibrium) in time and space proportional to the number of nodes plus the number of edges in the tree, $O(|V| + |E|)$. However, the size of the tree does not seem to be a psychologically plausible complexity measure. To see this, consider two trees of equal size, but in the first one all the nodes are controlled by Player 1 while in the second tree, the players alternate. Obviously, the problem posed by the second tree is much more complex. This suggests that one of the key aspects of the problem is the structure of the move alternation in the game tree. Let us then categorize game trees with respect to such alternations. In the following, we restrict the analysis to two-player games, although it would be possible to extend the ideas to finite dynamic games for more than two players.

Definition 2.1. Let us assume that the players $\{1, 2\}$ strictly alternate in the game; Let player $i \in \{1, 2\}$. Then:

[2]The term 'strategy' is used here more broadly than in game theory, where it is just a partial function from the set of histories (sequences of events) at each stage of the game to the set of actions of the player when it is supposed to make a move. We are interested in human reasoning strategies that can be used to solve the cognitive problems posed by the game.

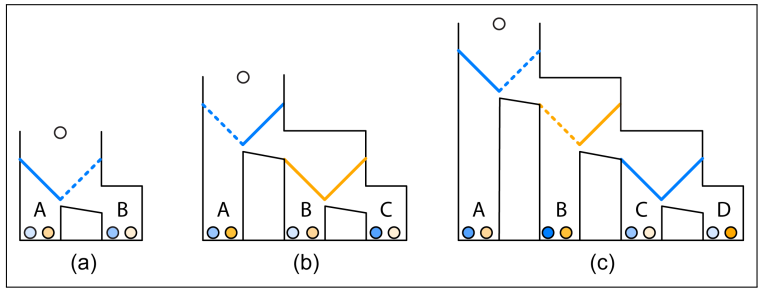


Figure 1: Examples of a zero-, first-, and second-order Marble Drop game. The blue marbles, on the left-hand side in the bins, are the participant's payoffs and the orange marbles, on the right-hand side, are the computer's payoffs. The marbles can be ranked from the lightest to the darkest. For each player, the goal is to get the white marble to drop 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 controls the orange ones (the second set of trapdoors from the left). The participants are told that the computer aims at maximizing its pay-off. The dashed lines represent the trapdoors that both players should remove to attain the darkest possible marble of their color. See `http://www.ai.rug.nl/~meijering/marble_drop.html` for an interactive demo. Backward induction reasoning proceeds from the last decision, which in 1c is Player 1's decision between the blue marbles in payoff-pairs C and D. Player 1 would decide to remove the left trapdoor because C contains the darker blue marble. Backward induction would then proceed with the second-to-last decision, which is Player 2's decision between the orange marbles in payoff-pairs B and C. Player 2 would decide to remove the left orange trapdoor, because B contains the darker orange marble. Backward induction reasoning stops at the third-to-last decision, which is Player 1's decision between the blue marbles in payoff-pairs A and B. Player 1 would remove the right blue trapdoor, because B contains the darker blue marble.

- In a Λ_1^i tree, all the nodes are controlled by Player i.

- A Λ_{k+1}^i tree, a tree of k-alternations for some $k \geq 0$, starts with a Player i node.[3]

For instance, the tree in Fig. 2 is Λ_3^1, a 1-game tree of 2 alternations, because Player 1 has the first move at the root, followed by an alternation of Player 1 to Player 2 and another alternation of Player 2 to Player 1.

[3]From the computational complexity theory perspective, this corresponds to a hierarchy of computational problems of increasing complexity (see, e.g., Arora and Barak 2009).

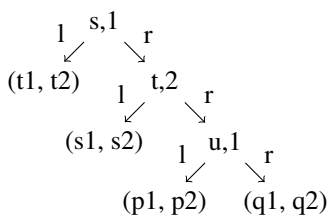


Figure 2: Nodes s and u are controlled by Player 1. t is controlled by Player 2. If a player controls a node then in that node he can choose whether to go left, l, or right, r. Every leaf is labeled with the pay-offs for Players 1 and 2.

3 Pay-off structure and cognitive difficulty

From the psychological perspective, it seems really crucial to take pay-offs into account when comparing the difficulty of particular MDG tasks. For instance, the two trees from Fig. 3 are Λ^1_3, because they both start with Player 1 and both have two alternations, from Player 1 to Player 2 and back again. However, clearly, the first game, represented by T_1, is much easier for Player 1 than the second game, represented by T_2. In the first game it is enough for Player 1 to realize that 999 is the highest possible pay-off, and then he can instantly move left and finish the game.

To explain the eye-tracking data of the subjects solving the Marble Drop game, Meijering et al. (2012) suggest that subjects may be using forward reasoning with backtracking (henceforth FRB), based on statistical analysis of eye gaze sequences. For instance, in the game from Fig. 1c, Player 1 will find out that B contains the darkest blue marble. He has to ask himself whether that marble is attainable. In other words, he has to reason about whether Player 2 would remove the left orange trapdoor. Therefore, Player 1 has to look at the orange marbles in bins B, C and D to find out that bin D contains Player 2's darkest orange marble. The reasoning continues with Player 1 asking himself whether Player 2 thinks that her orange marble in bin D is attainable. In other words, Player 1 has to reason about whether she thinks that he would remove the right blue trapdoor of the rightmost set of trapdoors. Player 1 knows that he would not remove that trapdoor, but that he would remove the left one instead. He also knows that she is aware of this, as both players are aware of each other's goals. Therefore, Player 1 knows that Player 2 knows that her darkest orange marble in D is unattainable. Therefore, Player 1 has to go back to the second decision point (i.e., the orange trapdoors). There, Player 2 would compare the orange marbles in B and C and decide to remove the left orange trapdoor, because the orange marble in B is the

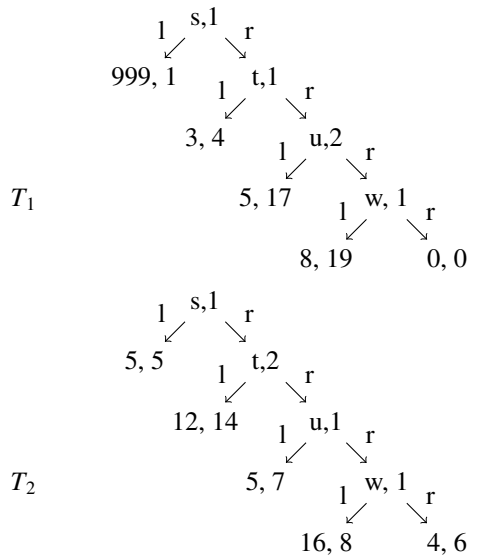


Figure 3: Two Λ_3^1 trees

darkest orange marble that she can still attain. To conclude, Player 1 knows that his darkest blue marble in B is attainable, and will thus remove the right blue trapdoor of the leftmost set of trapdoors.

As it is relatively hard to conclude from the eye-tracking data whether subjects apply exactly the above described forward reasoning with backtracking, we propose an orthogonal idea. We aim to identify the properties of the games that make certain trials harder than others and see whether such an explanation is congruent with forward reasoning plus backtracking. In order to do that, we put forward the following definitions. The idea here is that subjects may be looking for the highest possible pay-off and then try to reach it.

Definition 3.1. A game T is generic, if for each player, distinct end nodes have different pay-offs.

Note, for instance, that the game in Figure 1c is generic: the four bins contain marbles of four different hues of blue and four different hues of orange.

Definition 3.2. Suppose $i \in \{1, 2\}$. If T is a generic game tree with the root node controlled by Player i and n is the highest possible pay-off for Player i, then T^- is the minimal subtree of T containing the root node and the node with pay-off n for Player i.

To illustrate this definition, Figure 4 shows the restricted T^- trees for the two trees shown in Figure 3.

Hypothesis 1. *Let us take two MDG trials T_1 and T_2. T_1 is easier for participants than T_2 if and only if T_1^- is lower in the tree alternation hierarchy than T_2^-.*

Hypothesis 1 takes into account pay-off structures. According to it, the first tree from Fig. 3, T_1, should be easier for participants than the right tree, T_2, as T_1^- is a Λ_1^1 tree while T_2^- is still Λ_3^1, see Fig. 4. Moreover, it is possible that some subjects may try to apply the procedure iteratively: check if the maximum pay-off is reachable, if not then check for the second-best pay-off, and so on.

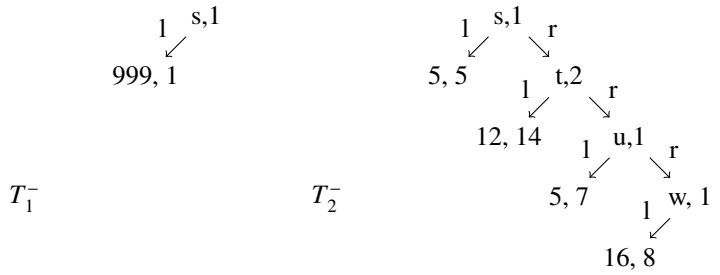


Figure 4: The maximum pay-off restricted trees corresponding to the trees in Fig. 3

As an additional question, we ask whether the following predictions agree with the proposal of Meijering and colleagues (Meijering et al. 2012) that the subjects in the game are applying forward reasoning, with backtracking when necessary (FRB). First of all, why would subjects ever apply FRB?

Hypothesis 2. *For an average random game with 3 decision points structured as the Λ_3^1 game of Figure 2, the forward reasoning plus backtracking algorithm needs fewer computation steps to yield a correct solution than backward induction.*

Furthermore, if subjects used forward reasoning, then we could observe the following by running FRB algorithm on the game trees:

Hypothesis 3. *Let us take two MDG trials T_1 and T_2. The forward induction with backtracking algorithm yields a correct solution for T_1 faster than for T_2 if and only if T_1^- is lower in the tree alternation hierarchy than T_2^-.*

4 Experimental results

To experimentally corroborate our hypotheses, we analyzed performance and reaction time data from (Meijering et al. 2012). Twenty-three first-year psychology students (14 female) with a mean age of 20.8 years (ranging from 18 to 24 years) participated in the experiment and were asked to solve Marble Drop trials, in the sense that they had to make a decision 'left' or 'right' at the first decision point. All experimental game trials had payoff structures that required Player 1 to reason about the decision at each of the three decision points, structured as the Λ_3^1 game of Figure 2. Therefore, the experiment was constructed in a way to be diagnostic for second-order theory of mind (see Meijering et al. 2012, for more information on the experimental design).

We divided experimental trials into two sets: **Accessible** ones, in which the highest possible pay-off for Player 1 is obtainable for him and **Inaccessible** ones, where his highest possible pay-off is not obtainable. For example, the game of Figure 1c is accessible, because Player 1 can reach the marble of the darkest hue of blue, which is located in bin B, by opening the right trapdoor; after all, Player 2 will also choose to stay there. Note that in general, if T_1 represents an accessible game and T_2 an inaccessible one, then T_1^- is lower in the alternation hierarchy than T_2^-.

Therefore, according to Hypothesis 1, our prediction was that the shortest reasoning times will be recorded in the condition "Accessible", where the highest pay-off was obtainable for Player 1.

Furthermore, by simulating forward reasoning with backtracking on experimental trials and computing the number of reasoning steps, we investigated hypotheses 2 and 3. Again, our prediction was that the number of steps should be smaller in "accessible" cases, where the highest-possible pay-off for Player 1 was obtainable.

4.1 Hypothesis 1: pay-offs and alternation type

To investigate the first hypothesis, we compared reaction times (RTs) in games in which the highest payoff was accessible against RTs in games in which the highest payoff was not accessible. The RTs were log-transformed to approximate the normal distribution.

A paired-samples t-test indicated a significant (within-subjects) difference, $t(12) = 4.07, p < .01$. The RTs decrease if the maximum payoff is accessible, which can be seen in Figure 5.

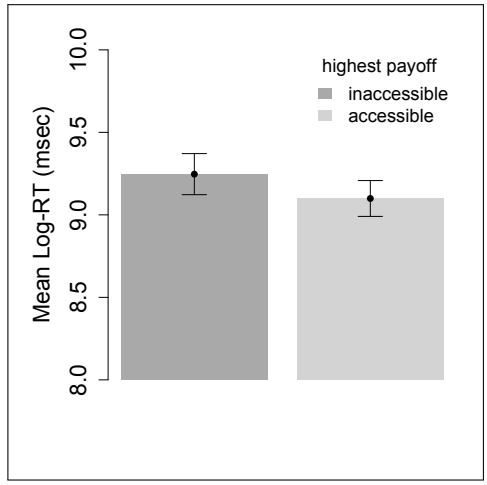


Figure 5: Players' reaction times with respect to accessibility, namely the attainability of the highest payoff for Player 1

4.2 Hypothesis 2: simulating the algorithms

When looking at all possible payoff-structures in Marble Drop games with two alternations (or three decision points), we implemented the forward reasoning plus backtracking algorithm as a set of heuristics based on several cases that can occur in the Marble Drop game; we used the same algorithm that we derived in (Meijering et al. 2012) from the participants' eye-tracking data.[4]

When using the algorithm on all 576 possible pay-off structures, we see that forward reasoning with backtracking in general requires fewer steps than backward induction, e.g., in 288 cases only 1 step is enough. More specifically, forward reasoning with backtracking requires on average 3 steps, whereas backward induction would always require 6 steps, irrespective of payoff structure. Table 1 provides a cross-table of payoff structures and number of steps. This simulation supports our Hypothesis 2.

These simulation results imply that, on average, it pays off to use a forward reasoning strategy. In fact, Meijering et al. (2012) found a strong prevalence of forward reasoning with backtracking, even though participants were presented with a subset of hard-to-solve games in which backward induction would actually be more efficient

[4]Thus, we did not use a generic implementation of forward reasoning with backtracking that would work for any possible game tree.

Table 1: Cross-table of payoff structures and the necessary number of steps when using forward reasoning with backtracking

# of steps	1	2	4	5	6	8
# of payoff structures	288	72	48	56	16	96

on average. However, participants did not know that they were presented with this particular subset of very difficult games.

4.3 Hypothesis 3: FRB and structural complexity

The implementation of the forward reasoning plus backtracking (FRB) algorithm was applied to the subset of actually presented experimental games to determine the number of reasoning steps required for each game. In the following analyses, number of steps was included as a predictor of the reaction times. We label the factor simply as 'forward reasoning with backtracking'.

The log-RTs were analysed by means of linear mixed-effects (LME) models (Baayen et al. 2008) to account for random effects of participants and unequal numbers of observations across all experimental conditions. Traditional (repeated measures) ANOVAs could not be performed as they require equal numbers of observations.

Fitting LMEs on the log-transformed reaction times, we see that forward reasoning plus backtracking (FRB) is a good predictor. The model with FRB cannot be rejected in favor of a simpler model without FRB as a predictor, $\chi^2(1) = 8.4$, $p = 0.004$. We discuss the best model below.

Again, the reaction times significantly decrease if the maximum Player 1 payoff is accessible (Table 2a). In case of games in which the maximum payoff is not accessible, the reaction times do not significantly increase with each additional reasoning step (Table 2b). Those games require in between 6 and 8 reasoning steps, which is too small a difference to find a significant effect on the RTs. In contrast, the RTs do significantly increase with each additional reasoning step in games in which the maximum payoff is accessible (Table 2c).

5 Discussion

We have investigated the structural properties of the Marble Drop Game, an experimental paradigm designed to study higher-order social reasoning. Using theoretical approaches from logic and complexity theory, we identified inherent properties of the

Table 2: Output of full-factorial linear mixed-effects model with factors Accessibility (A), Steps of forward reasoning with backtracking (FRB)

Parameter	Estimate	St. Error	t-value	p-value
a) Accessible	-0.689147	0.271256	-2.54	.000
b) FRB	0.008767	0.034930	0.25	.418
c) A:FRB	0.084336	0.037277	2.26	.000

game trials responsible for the cognitive difficulty of the task. Meijering and colleagues' (2012) reaction time data can be explained by looking at the alternation type and pay-off distribution of the particular game items. It turned out that the game items are harder if the maximum possible pay-off for Player 1 is not accessible for him. This observation is consistent with the assumption that participants were mostly applying forward reasoning with backtracking to solve the games. By simulating forward reasoning with backtracking on the experimental items, we have shown that the reaction times and the number of necessary comparisons significantly decrease if the maximum Player 1 payoff is accessible. As MDG is game-theoretically equivalent to many other experimental paradigms making use of turn-based games (see, e.g., Hedden and Zhang 2002, Gneezy et al. 2010, Hawes et al. 2012, Flobbe et al. 2008, Raijmakers et al. 2013), we would expect that our results generalize to those cases.

One could wonder why the subjects did not use backward induction in the first place, as it is the method that always delivers the optimal pay-off (Osborne and Rubinstein 1994). One possible answer is that they avoided backward induction in order to simplify the underlying reasoning. Recall, that while backward induction reasoning always takes 6 steps in the Marble Drop game with 3 decision points, forward reasoning and backtracking takes on average only 3 steps, corresponding with the phenomenon that T^- is usually lower in the tree alternation hierarchy than T itself. Moreover, iterating the forward reasoning strategy by backtracking in case the highest pay-off is not obtainable will finally lead to the optimal solution. Therefore, some subjects may choose to use that strategy to avoid higher-order reasoning, even though keeping the intermediate results in mind during backtracking is expected to tax working memory more than applying backward induction.

Subjects may as well use other heuristics that do not guarantee reaching the prescribed backward induction result, namely a Nash equilibrium of the game. For instance, as suggested by Hedden and Zhang (2002), subjects may assume that their opponents are playing according to some fixed patterns. Instead of assuming that the opponent is rational and correctly predicts Player 1's choice at the last decision point,

Player 1 may take his opponent to be risk-averse or risk-taking. Such heuristics, essentially based on considering sub-trees of the initial game-tree, will also lead to simplified reasoning.

Of course, assuming that the opponent is of some specific type changes the game drastically and can lead to a very bad outcome, in case of wrong judgement of the other player's type. Still, people notoriously apply similar heuristics in strategic situations, for example, when joining a poker table, many players try to evaluate whether the opponents play 'loose' or 'tight'.[5] An important question is what are the good alternative strategies. They should be not only easy to compute for people but also relatively safe to apply. It seems that the forward reasoning plus backtracking strategy in MDG might be a cognitively attractive strategy for people asked to solve turn-based games. First of all, it does not ask for the second-order social reasoning that is known to be very hard even for many adults (Verbrugge 2009), and moreover, on average it demands fewer comparisons. One may even think that competent players know a collection of various strategies and their strategic abilities could be partially equated with the skill of choosing the right one, i.e., a strategy that may be safely applied in a given context to simplify the underlying reasoning.

6 Outlook

Inspired by the logical study of backward induction and the cognitive science experiments with the Marble Drop Game, we investigated structural properties of turn-taking dynamic games and we provided a more refined analysis of the complexity of particular game trials, which takes into account alternation type of the game and pay-off distribution. We compared our predictions to actual reaction time data from (Meijering et al. 2012).

Of course, there are many further topics to be resolved. For instance, it would be interesting to extend our analysis to account for imperfect information games. Also it would be fruitful to explore connections with various related logical formalisms and to investigate further epistemic phenomena. In parallel, we would like to confront Hypotheses 1 and 3 with the available eye-tracking data from (Meijering et al. 2012), as well as with eye-tracking data to be gathered from a wider class of turn-based two-player games. Moreover, we plan to investigate other reasonable reasoning strategies that subjects may successfully adapt in game-plays.

[5] A similar phenomenon is well-recognized in natural language semantics. People often shift the meaning of sentence φ from $[\![\varphi]\!]$ to a more restricted meaning $[\![\psi]\!] \subseteq [\![\varphi]\!]$. And again, one of the factors triggering such meaning-shifts might be related to the computational complexity of φ (see, e.g., Szymanik 2010).

Acknowledgements The authors are grateful for the support of Vici Grant NWO-277-80-001awarded to RV. JS also acknowledges NWO Veni Grant 639.021.232.

References

S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

R. Baayen, D. Davidson, and D. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412, 2008.

J. van Benthem. Extensive games as process models. *Journal of Logic, Language and Information*, 11(3):289–313, 2002.

C. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, New Jersey, Mar. 2003.

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 (4):417–442, 2008.

M. Frixione. Tractable competence. *Minds and Machines*, 11(3):379–397, 2001.

B. Geurts. Reasoning with quantifiers. *Cognition*, 86(3):223–251, 2003.

S. Ghosh and B. Meijering. On combining cognitive and formal modeling: A case study involving strategic reasoning. In J. Van Eijck and R. Verbrugge, editors, *Proceedings of the Workshop on Reasoning About Other Minds: Logical and Cognitive Perspectives (RAOM-2011)*, Groningen, 2011. CEUR Workshop Proceedings.

S. Ghosh, B. Meijering, and R. Verbrugge. Logic meets cognition: Empirical reasoning in games. In O. Boissier, A. E. Fallah-Seghrouchni, S. Hassas, and N. Maudet, editors, *MALLOW*, volume 627 of *CEUR Workshop Proceedings*, Lyon, 2010. CEUR-WS.org.

N. Gierasimczuk, H. van der Maas, and M. Raijmakers. An Analytic Tableaux Model for Deductive Mastermind Empirically Tested with a Massively Used Online Learning System. Journal of Logic, Language and Information 22(3): 297-314, 2013

U. Gneezy, A. Rustichini, and A. Vostroknutov. Experience and insight in the race game. *Journal of Economic Behavior and Organization*, 75(2):144 – 155, 2010.

D. R. Hawes, A. Vostroknutov, and A. Rustichini. Experience and abstract reasoning in learning backward induction. *Frontiers in Neuroscience*, 6(23), 2012.

T. Hedden and J. Zhang. What do you think I think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1):1 – 36, 2002.

D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing Visual Information*. W.H. Freeman, San Francisco, 1983.

B. Meijering, L. van Maanen, H. van Rijn, and R. Verbrugge. The facilitative effect of context on second-order social reasoning. In R. Catrambone and S. Ohlsson, editors, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1423–1428, Austin (TX), 2010. Cognitive Science Society.

B. Meijering, H. van Rijn, 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 33rd Annual Meeting of the Cognitive Science Society*, pages 2486–2491, Boston, 2011. Cognitive Science Society.

B. Meijering, H. van Rijn, N. A. Taatgen, and R. Verbrugge. What eye movements can tell about theory of mind in a strategic game. *PLoS ONE*, 7(9):e45961, 09 2012.

M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. The MIT Press, Cambridge, MA, 1994.

M. E. Raijmakers, D. J. Mandell, S. E. Es, and M. Counihan. Children's strategy use when playing strategic games. *Synthese*, 2013.

I. van Rooij. The tractable cognition thesis. *Cognitive Science: A Multidisciplinary Journal*, 32(6):939–984, 2008.

K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science*. The MIT Press, Cambridge, MA, 2008.

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

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

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

R. Verbrugge and L. Mol. Learning to apply theory of mind. *Journal of Logic, Language and Information*, 17(4):489–511, 2008.

Backward Induction is PTIME-complete

Jakub Szymanik

Institute for Logic, Language, and Computation, University of Amsterdam
J.K.Szymanik@uva.nl

Abstract

We prove that the computational problem of finding backward induction outcome is PTIME-complete.

1 Introduction

Higher-order reasoning of the form 'I believe that Ann knows that Peter thinks...' is an attractive topic for logical analysis. The logical investigations often go hand in hand with game theory. In this context, one of the common topics among researchers in logic and game theory has been backward induction (henceforth, BI), the process of reasoning backwards, from the end to determine a sequence of optimal actions. BI is a common method for determining sub-game perfect equilibria in the case of finite extensive-form games. BI can be understood as an inductive algorithm defined on a game tree – an algorithm that tells us which sequence of actions will be chosen by agents that want to maximize their own payoffs, assuming common knowledge of rationality.

Games have been also extensively used to design experimental paradigms aiming at studying social cognition, with a particular focus on higher-order social cognition. Often the experimental turn-based games can be modeled as extensive-form games and solved by applying BI. As it is hard to determine what the reasoning strategies used by participants in such games are, formal findings on backward induction have been used to better understand humans' strategic reasonings (Szymanik et al. 2013).

Recently, van Benthem and Gheerbrant (2010) have studied the logical definability of BI. They have observed that it can be defined in the first-order logic extended with the least fixed-point operator as well as in a variety of other dynamic epistemic formalisms. Obviously, from the least fixed-point definability result it follows that BI is in PTIME (Immerman 1998). But is it also hard, and therefore complete for PTIME?

2 Preliminaries

Let us start by recalling that the reachability problem on alternating graphs is PTIME-complete (Immerman 1998).

Definition 2.1. Let an alternating graph $G = (V, E, A, s, t)$ be a directed graph whose vertices, V, are labeled universal or existential. $A \subseteq V$ is the set of universal vertices. $E \subseteq V \times V$ is the edge relation.

Definition 2.2. Let $G = (V, E, A, s, t)$ be an alternating graph. We say that t is reachable from s iff $P_a^G(s, t)$, where $P_a^G(x, y)$ is the smallest relation on vertices of G satisfying:

(1) $P_a^G(x, x)$.

(2) If x is existential and $P_a^G(z, y)$ holds for some edge (x, z) then $P_a^G(x, y)$.

(3) If x is universal, there is at least one edge leaving x, and $P_a^G(z, y)$ holds for all edges (x, z) then $P_a^G(x, y)$.

The idea here is that for t to be reachable from an existential node x there must exist a path from x to t, while the condition for a universal node y is stronger: t is reachable from y if and only if every path from y leads to t. One can think about alternating reachability in terms of a competitive game, where the player controlling existential vertices wants to get to t and the player controlling universal vertices is trying to prevent that. For example, in the alternating graph of Figure 1, t is not reachable from s (i.e., there is no winning strategy for the existential player). To see it just imagine that the first player will move from s to v. Then the second player has only one choice leading to the dead-end. It means, that not every move of the first player controlling the universal node s is on the path to t.

Now, we can define the alternating reachability problem, that is a class of alternating graphs in which t is reachable from s. One can think about that as a decision problem: given an alternating graph G and nodes s, t check whether t is reachable from s.

Definition 2.3. $REACH_a = \{G | P_a^G(s, t)\}$

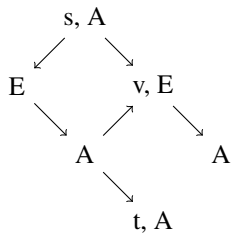


Figure 1: t is not reachable from s

The following computational complexity result will be crucial for us.

Theorem 1 (Immerman 1981)**.** *$REACH_a$ is PTIME-complete via first-order reductions.*

Proof. The original proof of Immerman simulates directly an alternating Turing machine (ATM) to show that the problem is complete for ATM logarithmic space, known to be equal to P (Chandra et al. 1981). □

As the proof of Theorem 1 simulates ATM computation tree it follows that:

Corollary 1. *$REACH_a$ is PTIME-complete on trees.*

Note, that given a game tree T and an existential node s, the problem $REACH_a$ over T intuitively corresponds to the question: 'Is s a winning position for the first player in the zero-sum game T, i.e., can the first player force the game from node s towards node t against all possible counterstrategies of the second player?' (see Greenlaw et al. 1995, Problem A.11.1).

3 Backward Induction Problem

Now we are ready to define the computational decision problem corresponding to BI for extensive form, non zero-sum games. Intuitively: can the first player force the game from node s towards node t against all possible *rational* (= pay-off maximizing) counterstrategies of the second player? The difference here is that we consider only rational strategies as the non zero-sum games do not have to be strictly competitive.

Definition 3.1. A two-player finite extensive form game
$T = (V, E, V_1, V_2, V_{end}, f_1, f_2, s, t)$, where V is the set of nodes, $E \subseteq V \times V$ is the edge

relation (available moves). For $i = 1, 2$, $V_i \subseteq V$ is the set of nodes controlled by Player i, and $V_1 \cap V_2 = \emptyset$. V_{end} is the set of end nodes. Finally, $f_i : V_{end} \longrightarrow \mathbb{N}$ assigns pay-offs for Player i.

Without loss of generality let us restrict attention to *generic games*:

Definition 3.2. A game T is generic, if for each player, distinct end nodes have different pay-offs.

Definition 3.3. Let T be a two-player game. We define the backward induction accessibility relation on T. Let $P^T_{bi}(x, y)$ be the smallest relation on vertices of T such that:

(1) $P^T_{bi}(x, x)$

(2) Take $i = 1, 2$. Assume that $x \in V_i$ and $P^T_{bi}(z, y)$. If the following two conditions hold, then also $P^T_{bi}(x, y)$ holds:

   (a) $E(x, z)$;

   (b) there is no w, v such that $E(x, w)$, $P^T_{bi}(w, v)$, and $f_i(v) > f_i(y)$.

For example, in the tree of Figure **??** t is not a backward induction solution for the game starting from s. Player 2 will rather start the game by going to the state w than v. And, t is not reachable from w.

We can again define the corresponding decision problem – whether in the game represented by tree T and starting in node s the first player can force the output t – as a class of game trees where s and t belong to the backward induction accessibility relation on T?

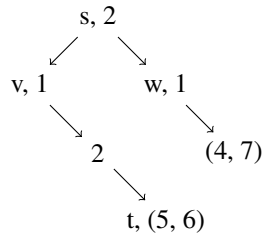


Figure 2: t is not reachable from s

Definition 3.4. $\mathbb{BI} = \{T | P^T_{bi}(s,t)\}$

The problem $\mathbb{BI}$ intuitively corresponds to the question whether t is a sub-game perfect equilibrium in game T starting at node s (Osborne and Rubinstein 1994).

4 Complexity of BI

The definability result of Van Benthem and Gheerbrant implies that it can be decided in polynomial time whether node t is a subgame perfect equilibrium of the game, i.e., the result of a gameplay following a BI strategy. First of all, note that it also follows that given an arbitrary finite extensive game with starting node s one can find a BI solution of the game in polynomial time. Simply, it is enough to run the polynomial decision algorithm for every node of the game. In this section we prove that computing backward induction relation is not only in PTIME but it is actually a PTIME-complete problem.

Theorem 2. $\mathbb{BI}$ *is PTIME-complete via first-order reductions.*

Proof. First of all, $\mathbb{BI}$ is in PTIME by providing FO(LFP) definition (Immerman 1998). Now, it suffices to show PTIME-hardness. For that we will reduce the $REACH_a$ problem on trees (cf. Corollary 1) to the $\mathbb{BI}$ problem. We take any alternating tree $T = (V, E, A, s, t)$. Without loss of generality let us assume that s is existential. We construct a two player game, $T' = (V', E', V_1, V_2, V_{end}, f_1, f_2, s', t')$, where: $V = V'$, $t \in V_{end} = \{\text{end nodes of V}\}$, $E = E'$, $V_1 = V - A$, $V_2 = A$, $s = s$, $t = t$, and for every $v \in V'$, if $v \neq t$, then $f_1(t) > f_1(v)$ and $f_2(t) < f_2(v)$.

Now, we need to prove that $T \in REACH_a$ iff $T' \in \mathbb{BI}$. Assume, that $T \in REACH_a$. It means that whatever Player 2 does in the game T', Player 1 has a strategy to force outcome t. As t gives strictly the best pay-off for Player 1, then $P^{T'}_{bi}(s,t)$. Hence, $T' \in \mathbb{BI}$. For the other direction, assume for contradiction that $T \notin REACH_a$. This means that there is a node $v \neq t$ such that Player 2 can guarantee the game T' to end in v. From the pay-off construction for T', v is more attractive to Player 2 than t. Therefore, it is not the case that $P^{T'}_{bi}(s,t)$. Hence, $T' \notin \mathbb{BI}$. □

What does this tell us about the complexity of backward induction? First of all, problems in PTIME are usually taken to be tractable (Edmonds 1965), so relatively easy to solve, also for humans (Frixione 2001). Furthermore, given assumptions on non-collapse, PTIME-completeness suggests that the problem of deciding whether a

given node is a sub-game perfect equilibrium of the game is difficult to effectively parallelize (it lies outside NC[1]) and solve in limited space (it lies outside LOGSPACE).

Acknowledgements The research was supported by Veni Grant NWO-639-021-232. The author also wish to thank Rineke Verbrugge for many comments and suggestion as well as her Vici project NWO-277-80-001.

References

J. van Benthem and A. Gheerbrant. Game solution, epistemic dynamics and fixed-point logics. *Fundamenta Informaticae*, 100(1-4):19–41, 2010.

A. K. Chandra, D. C. Kozen, and L. J. Stockmeyer. Alternation. *J. ACM*, 28(1): 114–133, 1981.

J. Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965.

M. Frixione. Tractable competence. *Minds and Machines*, 11(3):379–397, 2001.

R. Greenlaw, J. H. Hoover, and W. L. Ruzzo. *Limits to Parallel Computation: P-Completeness Theory*. Oxford University Press, USA, 1995.

N. Immerman. Number of quantifiers is better than number of tape cells. *Journal of Computer and System Sciences*, 22(3):384 – 406, 1981.

N. Immerman. *Descriptive Complexity*. Texts in Computer Science. Springer, New York, NY, 1998.

M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. The MIT Press, Cambridge, MA, 1994.

J. Szymanik, B. Meijering, and R. Verbrugge. Using intrinsic complexity of turn-taking games to predict participants' reaction times. In M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, editors, *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 1426–1432, Austin, TX, 2013. Cognitive Science Society.

[1] A problem is in NC if there exist constants c and k such that it can be solved in time $O(log^c n)$ using $O(n^k)$ parallel processors.

Coherence

Branden Fitelson

Department of Philosophy, Rutgers University
branden@fitelson.org

Abstract

Taking inspiration from the arguments for probabilism advanced by de Finetti (1970) and Joyce (1998, 2009), we develop a general framework for grounding synchronic, epistemic coherence requirements for various types of judgment. Then, we show how to apply our general framework to yield coherence requirements for both full belief and comparative confidence.

1 Introduction: de Finetti, Joyce, and probabilism

Joyce's argument for probabilism as a formal, epistemic, synchronic coherence requirement for (numerical) *degrees* of confidence (*viz.*, credences) was inspired (formally) by an elegant geometrical argument of de Finetti. In this section, we will briefly explain how the simplest case of their (formal) argument goes. Then, we will show how their argument can be generalized to yield a framework for grounding (synchronic, epistemic) coherence requirements, which can be applied to various types of judgment. Finally, we will discuss two applications of this general framework: full belief and comparative confidence.

De Finetti (1970) showed that if a credence function b is non-probabilistic, then there exists another credence function b' that is (in one precise sense) *strictly more accurate — in all possible worlds*.[1] Let's keep things maximally simple. Consider a toy agent S who is forming judgments over a very simple Boolean algebra generated

[1] De Finetti did not interpret the Brier score (his favored scoring rule) as a measure of "inaccuracy". Joyce (1998, 2009) was the first to give this *epistemic* interpretation to de Finetti's argument.

by a propositional language containing one atomic sentence (P). That is, S's doxastic space contains only four propositions $\{\top, P, \neg P, \bot\}$. For simplicity, we will assume that S assigns credence 1 to $\top$ and credence 0 to $\bot$. Thus, the question of S's *probabilistic* coherence reduces to the question of whether S's credence function b satisfies the following two probabilistic constraints regarding P and $\neg P$:

- $b(P) \in [0, 1]$ and $b(\neg P) \in [0, 1]$;
- $b(P) + b(\neg P) = 1$.

Next, let's think about how we might "score" a credence function, in terms of its "*distance from the truth* (or *inaccuracy*) in a possible world". For our toy agent, there are only two relevant possible worlds: w_1 in which P is false, and w_2 in which P is true. If we use the number 1 to "numerically represent" the truth-value true (at a world) and the number 0 to "numerically represent" the truth-value false (at a world), then we can "score" a credence function b using a *scoring rule* which is some function of (i) the values b assigns to P and $\neg P$, and (ii) the "numerical truth-values" of P and $\neg P$ at the two relevant possible worlds w_1 and w_2. It is standard in this context (beginning with de Finetti) to use what is called the *Brier score* (of a credence function b, at a world w), which is the sum of squared differences between credences and truth values. For our toy agent S, it is defined in the following way (think *Euclidean distance* between the agent's credences and the numerical truth-values of P and $\neg P$ in worlds w_1 and w_2, respectively).

- The Brier score of b in $w_1 =_{\text{df}} (0 - b(P))^2 + (1 - b(\neg P))^2$.
- The Brier score of b in $w_2 =_{\text{df}} (1 - b(P))^2 + (0 - b(\neg P))^2$.

The idea behind all such scoring rules is that the *inaccuracy* of a credence function b (at a world w) is measured in terms of b's *distance (at w) from the numerical truth-values* of the set of propositions in the agent's doxastic space.

With these basics in mind, we can now explain how the simplest case of the de Finetti/Joyce argument proceeds. Figure 1 depicts an elegant "geometrical proof" of the following theorem regarding our toy agent S.

> **Theorem** (de Finetti 1970). S's credence function b is non-probabilistic (*i.e.*, S's b violates at least one of the two probabilistic constraints above) *if and only if* ($\Leftrightarrow$) there exists another credence function b' *which has lower Brier score in every possible world* (*i.e.*, b' is *closer to the truth-values* of $P, \neg P$ than b is *in every possible world*, as measured *via* Euclidean distance).

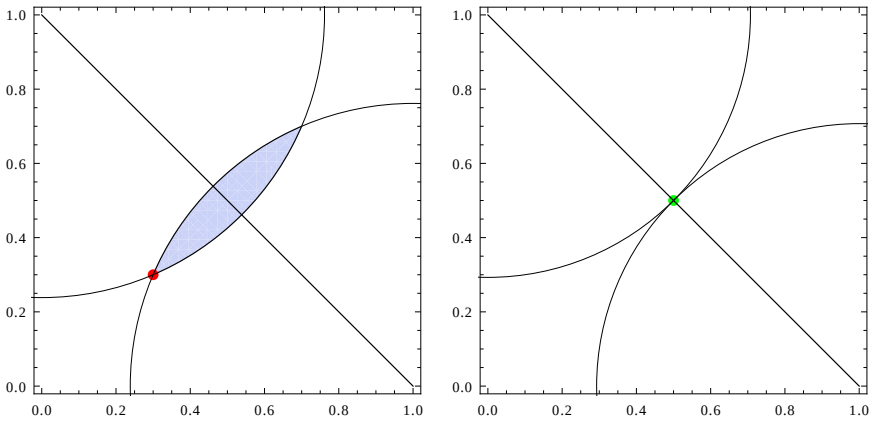


Figure 1: Geometrical proof of the two directions of (the simplest case of) de Finetti's theorem (on the simplest Boolean algebra)

The x-axes in these plots represent $b(P)$ and the y-axes represent $b(\neg P)$. The diagonal lines in the plots represent the set of all of *probabilistic* credence functions b such that $b(P) + b(\neg P) = 1$. The *only if* direction ($\Rightarrow$) of de Finetti's theorem is illustrated in the plot on the left side of Figure 1. The dot (at approximately $b(P) = b(\neg P) = 1/3$) represents a non-probabilistic credence function b. The two curves drawn through the dot represent the sets of credence functions that are the same (Euclidean) distance as b from w_1 and w_2, respectively. The credence functions in the shaded region (which will be non-empty, so long as b is non-probabilistic) are *guaranteed to be closer* (in Euclidean distance) *to the truth-values* of $P, \neg P$ *in both possible worlds*. The *if* direction ($\Leftarrow$) of de Finetti's theorem is illustrated in the plot on the right side of Figure 1. This time, the dot represents a *probabilistic* credence function b. The two curves drawn through the dot represent the sets of credence functions that are the same (Euclidean) distance as b from w_1 and w_2, respectively. This time, the curves are *tangent*, which means *there is no* credence function b' that is closer (in Euclidean distance, *come what may*) to the truth than b is. This simple geometrical argument for the simplest case of de Finetti's theorem can be generalized to finite Boolean algebras of arbitrary size. Unlike the traditional (pragmatic) arguments for probabilism (Ramsey 1926; Hájek 2008), Joyce's interpretation of de Finetti's argument is *epistemic*. This is because it trades solely on the *accuracy* of a credence function b, which is a *distinctively epistemic* aspect of b. In the next section, we explain how to turn Joyce's argument for probabilism into a

general recipe for generating and grounding coherence requirements for different types of judgment.

2 Generalizing Joyce's argument for probabilism

Abstracting away from the details of Joyce's argument for probabilism (as an epistemic coherence requirement for *numerical credences*) reveals a general strategy for grounding coherence requirements for judgment sets of various types. This general argumentative strategy involves *three steps*.

> **Step 1**: Define the *vindicated (viz., perfectly accurate) judgment set* (of type J), at world w. Call this set $\mathring{\mathbf{J}}_w$.
>
> – Think of $\mathring{\mathbf{J}}_w$ as the judgments (of type J) that "the *omniscient/ideal* agent" would make (at world w).
>
> **Step 2**: Define a notion of "distance between $\mathbf{J}$ and $\mathring{\mathbf{J}}_w$". That is, define a measure of *distance from vindication*: $\mathfrak{D}(\mathbf{J}, \mathring{\mathbf{J}}_w)$.
>
> – Think of $\mathfrak{D}$ as measuring how far one's judgment set $\mathbf{J}$ is (in w) from the vindicated or ideal set of judgments $\mathring{\mathbf{J}}_w$ (in w).
>
> **Step 3**: Adopt a *fundamental epistemic principle*, which *uses* $\mathfrak{D}(\mathbf{J}, \mathring{\mathbf{J}}_w)$ to ground a formal coherence requirement for $\mathbf{J}$.
>
> – Think of the fundamental epistemic principle as articulating an *evaluative connection* between $\mathfrak{D}$ and $\mathbf{J}$-coherence.

In the case of Joyce's argument for probabilism, the three steps were as follows

> **Step 1**: The *vindicated credence function* $\mathring{b}_w$ is the *indicator function* $v_w(\cdot)$ which assigns 1 to the truths in w and 0 to the falsehoods in w.
>
> **Step 2**: The measure of *distance from vindication* $\mathfrak{D}(b, \mathring{b}_w)$ is the *Brier score* (*i.e.*, *Euclidean distance* $d(b, \mathring{b}_w)$ between the credal vectors b and $\mathring{b}_w$).

> **Step 3**: The *fundamental epistemic principle* is *strict accuracy dominance avoidance* (SADA), which requires that there does *not* exist a credence function b' such that $d(b', \mathring{b}_w) < d(b, \mathring{b}_w)$, for all possible worlds w.

In the next section, we discuss the application of our general framework to the case of full belief.

3 Application #1: Full Belief[2]

3.1 Deductive consistency: The contemporary dialectic

When it comes to (formal, synchronic, epistemic) coherence requirements for full belief, philosophers have traditionally devoted a lot of attention to the requirement of *deductive consistency*, which can be stated informally as follows:

(CB) **Consistency Requirement for Belief**. Epistemically rational agents should (at any given time) have logically consistent belief sets.

One popular motivation for imposing such a requirement is the presupposition that epistemically rational agents should, in fact, obey the following norm:

(TB) **Truth Norm for Belief**. Epistemically rational agents should (at any given time) believe propositions that are true.

These two norms differ in one fundamental respect: (TB) is *local* in the sense that an agent complies with it only if each particular belief the agent holds (at a given time) has some property (in this case: truth). On the other hand, (CB) is a *global* norm: whether or not an agent's doxastic state (at a given time) is in accordance with (CB) is a more holistic matter, which trades essentially on properties of their entire belief set. While these two epistemic norms differ in this respect, they are also intimately related, logically. We may say that one norm n *entails* another norm n' just in case everything that is permissible according to n is permissible according to n'. In this sense, (TB) *asymmetrically entails* (CB). That is, if an agent is in accordance with (TB), then they must also be in accordance with (CB), but not conversely.

Although (CB) accords well with (TB) there is a strong case to be made that (CB) conflicts with other plausible local norms, in particular:

[2]This section draws heavily on joint work with Kenny Easwaran (Easwaran and Fitelson 2013) as well as Rachael Briggs and Fabrizio Cariani (Briggs et al. 2014). Those papers contain a much more detailed discussion of the application of our general framework to the case of full belief. Moreover, I am currently writing a book (Fitelson 2014) which will go into even more detail about the various applications of our general framework. In this article, I will (basically) be presenting a *précis* of that more complete story.

(EB) **Evidential Norm for Belief**. Epistemically rational agents should (at any given time) believe propositions that are supported by their evidence.

It is plausible to interpret preface cases as revealing a tension between (EB) and (CB). Here is a rendition of the preface that we find particularly compelling (Easwaran and Fitelson 2013).

> **Preface Paradox**. John is an excellent empirical scientist. He has devoted his entire (long and esteemed) scientific career to gathering and assessing the evidence that is relevant to the following first-order, empirical hypothesis: (H) all scientific/empirical books of sufficient complexity contain at least one false claim. By the end of his career, John is ready to publish his masterpiece, which is an exhaustive, encyclopedic, 15-volume (scientific/empirical) book which aims to summarize (all) the evidence that contemporary empirical science takes to be relevant to H. John sits down to write the Preface to his masterpiece. Rather than reflecting on his own fallibility, John simply reflects on the contents of (the main text of) his book, which constitutes *very strong inductive evidence in favor of* H. On this basis, John (inductively) infers H. But, John also believes each of the individual claims asserted in the main text of the book. Thus, because John believes (indeed, knows) that his masterpiece instantiates the antecedent of H, the (total) set of John's (rational/justified) beliefs is inconsistent.

We take it that, in suitably constructed preface cases (such as this one), it would be epistemically permissible for S to satisfy (EB) but violate (CB). That is, we think that some preface cases are *counterexamples* to the claim that (CB) is a requirement of (ideal) epistemic rationality. It is not our aim here to investigate whether this is the correct response to the preface paradox.[3] Presently, we simply take this claim as a *datum*. Our aim here will be to explain how to ground *alternatives* to (CB), using our general framework above.

3.2 Setting up our formal framework for full belief

To streamline our discussion, we will restrict our attention to the simplest application of our general framework to the case of full belief. To wit, let

$B(p) =_{\text{df}} S$ believes that p.

$D(p) =_{\text{df}} S$ disbelieves that p.

[3] We think Christensen (2004) has given compelling arguments for the epistemic rationality of certain preface cases (*i.e.*, for the rationality of some inconsistent belief sets).

For simplicity, we suppose that S is opinionated, and that S forms judgments involving propositions drawn from a finite Boolean algebra of propositions. More precisely, let $\mathcal{A}$ be an agenda, which is a (possibly proper) subset of some finite boolean algebra of propositions. For each $p \in \mathcal{A}$, S either believes p or S disbelieves p, and not both.[4] In this way, an agent can be represented by her "belief set" **B**, which is just the set of her beliefs (B) and disbeliefs (D) over some salient agenda $\mathcal{A}$. More precisely, **B** is a set of proposition-attitude pairs, with propositions drawn from $\mathcal{A}$ and attitudes taken by S toward those propositions (at a given time). Similarly, we think of propositions as sets of possible worlds, so that a proposition is true at any world that it contains, and false at any world it doesn't contain.[5] With these background assumptions in mind, we can now go through the three steps required for the application of our general framework to the case of (opinionated) full belief.

3.3 Step 1: The vindicated belief set

Step 1 is easy. It is clear what it means for a belief set **B** to be perfectly accurate/vindicated. The vindicated set $\mathring{\mathbf{B}}_w$ is given by the following definition:

> $\mathring{\mathbf{B}}_w$ contains $B(p)$ [$D(p)$] just in case p is true [false] at w.

This is clearly the best explication of $\mathring{\mathbf{B}}_w$, since $B(p)$ [$D(p)$] is accurate just in case p is true [false]. So, in this context, Step 1 is uncontroversial. It follows from our (widely accepted) correctness/accuracy conditions for belief/disbelief.

3.4 Step 2: Measuring distance between belief sets

Step 2 is less straightforward, because there are many ways one could measure "distance between judgment sets". For simplicity, we adopt perhaps the most naïve distance measure, which is given by

[4] The assumption of opinionation results in no significant loss of generality for present purposes. This is for two reasons. First, as Christensen (2004) convincingly argues, suspension of judgment is (ultimately) not a compelling way for defenders of (CB) to respond to the preface paradox (or other similar paradoxes of consistency). Second, we are not assuming that agents never suspend judgment (*i.e.*, that agents are opinionated *across the board*). Rather, we are focusing on specific agendas on which (epistemically rational) agents happen to exhibit opinionation. Of course, in general, we would want to be able to model suspension of judgment in our framework. See (Easwaran 2013) for just such a generalization of the present framework.

[5] It is implicit in this formalism that agents satisfy a weak sort of logical omniscience, in the sense that if two propositions are logically equivalent, then they are in fact the same proposition, and so the agent can't have distinct attitudes toward them. However, it is not assumed that agents satisfy a stronger sort of logical omniscience — an agent may believe some propositions while disbelieving some other proposition that is entailed by them (*i.e.*, our logical omniscience assumption does not imply closure).

$d(\mathbf{B}, \mathbf{B}') =_{\text{df}}$ the number of judgments on which $\mathbf{B}$ and $\mathbf{B}'$ disagree.[6]

In particular, if you want to know how far your judgment set $\mathbf{B}$ is from vindication (at w) just count the number of mistakes you have made (at w). To be sure, this is a very naïve measure of distance from vindication. In this paper, I will not delve into the dialectic concerning measures of distance between judgment sets. Presently, I will simply assume the most naïve, counting measures of distance (for a detailed discussion of this issue, see Easwaran and Fitelson 2013, Fitelson 2014).[7]

3.5 Step 3: The fundamental epistemic principle for B

Step 3 is the philosophically crucial step. Given our setup, there is *a* choice of fundamental epistemic principle that yields (CB) as a coherence requirement for full belief. Specifically, consider the following principle

> **Possible Vindication** (PV). There exists *some* possible world w at which *all* of the judgments in $\mathbf{B}$ are accurate. Or, to put this more formally, in terms of our distance measure d: $(\exists w)[d(\mathbf{B}, \mathring{\mathbf{B}}_w) = 0]$.

Given our setup, it is easy to see that (PV) is equivalent to (CB). As such, a defender of (TB) would presumably find (PV) attractive as a fundamental epistemic principle. However, in light of preface cases (and other paradoxes of consistency), many philosophers would be inclined to say that (PV) is too strong to yield a (plausible, binding) coherence requirement for full belief. Indeed, we ultimately opt for fundamental principles that are strictly weaker than (PV). But, as we mentioned above, our rejection of (PV) was not (initially) motivated by prefaces and the like. Rather, our adoption of fundamental principles that are weaker than (PV) was motivated (initially) by analogy with Joyce's arguments for probabilism as a coherence requirement for credences.

In the case of credences, the analogue of (PV) is clearly inappropriate. The vindicated set of credences (*i.e.*, the credences an omniscient agent would have) are such that they assign maximal credence to all truths and minimal credence to all falsehoods (Joyce 1998). As a result, in the credal case, (PV) would require that all of one's credences be *extremal*. One doesn't need preface-like cases (or any other subtle or paradoxical cases) to see that this would be an unreasonably strong requirement. It is for

[6] This is called the *Hamming distance* between the binary vectors $\mathbf{B}$ and $\mathbf{B}'$ (Deza and Deza 2009).

[7] As it turns out, we only need to assume that our measures of distance between judgment sets are *additive* in a rather weak sense. This is explained in (Easwaran and Fitelson 2013, Fitelson 2014). We omit those discussions here, in the interest of maintaining the brevity of this *précis*. For further useful recent discussions concerning measures of distance between judgment sets, (see, e.g. Pigozzi 2006, Miller and Osherson 2009, Duddy and Piggins 2012).

this reason that Joyce (and all others who argue in this way for probabilism) back away from the analogue of (PV) to strictly weaker epistemic principles — specifically, to accuracy-dominance avoidance principles, which are credal analogues of the following fundamental epistemic principle.

> **Weak Accuracy-Dominance Avoidance** (WADA). **B** is *not weakly*[8] *dominated* in distance from vindication. Or, to put this more formally (in terms of d), there does *not* exist an alternative belief set $\mathbf{B}'$ such that:
>
> (i) $(\forall w)[d(\mathbf{B}', \mathring{\mathbf{B}}_w) \leq d(\mathbf{B}, \mathring{\mathbf{B}}_w)]$, and
>
> (ii) $(\exists w)[d(\mathbf{B}', \mathring{\mathbf{B}}_w) < d(\mathbf{B}, \mathring{\mathbf{B}}_w)]$.

(WADA) is a very natural principle to adopt, if one is not going to require that it be possible to achieve perfect accuracy. Backing off (PV) to (WADA) is analogous to what one does in decision theory, when one adopts a weak dominance principle rather than a principle of *maximizing (actual) utility*.

Initially, it may seem undesirable for an account of epistemic rationality to allow for doxastic states that cannot be perfectly accurate. But, as Richard Foley (1992) explains, an epistemic strategy that is guaranteed to be imperfect is sometimes preferable to one that leaves open the possibility of vindication.

> . . . if the avoidance of recognizable inconsistency were an absolute prerequisite of rational belief, we could not rationally believe each member of a set of propositions and also rationally believe of this set that at least one of its members is false. But this in turn pressures us to be unduly cautious. It pressures us to believe only those propositions that are certain or at least close to certain for us, since otherwise we are likely to have reasons to believe that at least one of these propositions is false. At first glance, the requirement that we avoid recognizable inconsistency seems little enough to ask in the name of rationality. It asks only that we avoid certain error. It turns out, however, that this is far too much to ask.

We agree with Foley's assessment that (PV) is too demanding. (WADA), however, seems to be a better candidate fundamental epistemic principle. As we will explain below, if S violates (WADA), then S's doxastic state *must* be defective — from both alethic and evidential points of view.

[8] Strictly speaking, Joyce *et al.* opt for *strict* dominance-avoidance principles. However, in the credal case (assuming continuous, strictly proper scoring rules), there is no difference between weak and strict dominance (Schervish *et al.* 2009). So, there is no serious disanalogy here.

3.6 Principled alternatives to deductive consistency

If an agent S satisfies (WADA), then we say S is *non-dominated* (we'll also apply the term 'non-dominated' to belief sets). The above considerations suggest the following new coherence requirement for full belief

(NDB) Epistemically rational agents should (at any given time) be non-dominated.

Interestingly, (NDB) is strictly weaker than (CB). Moreover, (NDB) is weaker than (CB) in the right way, in light of the preface case (and other similar paradoxes of consistency). Our first two theorems help to explain why.[9]

The first theorem states a necessary and sufficient condition for (*i.e.*, a characterization of) non-dominance: we call it *Negative* because it identifies certain objects, the *non*-existence of which is necessary and sufficient for non-dominance. The second theorem states a sufficient condition for non-dominance: we call it *Positive* because it states that in order to show that a certain belief set **B** is non-dominated, it's enough to construct a certain type of object.

Definition 3.1 (Witnessing Sets)**.** **S** is a *witnessing set* iff (a) at every world, at least half of the judgments[10] in **S** are inaccurate; and, (b) at some world, more than half of the judgments in **S** are inaccurate.

If **S** is a witnessing set and no proper subset of it is a witnessing set, then **S** is a *minimal witnessing set*. Notice that if **S** is a witnessing set, then it must contain a minimal witnessing set. Theorem 1 shows that the name "witnessing set" is apt, since these entities provide a witness to incoherence.

Theorem 1 (Negative)**.** **B** *is non-dominated if and only if no subset of* **B** *is a witnessing set.*

It is an immediate corollary of this first theorem that if **B** is logically consistent [*i.e*, if **B** satisfies (PV)], then **B** is non-dominated. After all, if **B** is logically consistent, then there is a world w such that no judgments in **B** are inaccurate at w. However, while consistency guarantees coherence, the converse is not the case. That is, coherence does not guarantee consistency. This will be most perspicuous as a consequence of our second central theorem:

[9] In the interest of brevity, we omit all proofs. All theorems reported here are proven in (Easwaran and Fitelson 2013, Fitelson 2014, Fitelson and McCarthy 2013).

[10] Throughout the paper, we rely on naïve counting. This is unproblematic since all of our algebras are finite.

Definition 3.2. A probability function Pr *represents* a belief set **B** iff for every $p \in \mathcal{A}$:

(i) **B** contains $B(p)$ iff $\Pr(p) > 1/2$.

(ii) **B** contains $D(p)$ iff $\Pr(p) < 1/2$.

Theorem 2 (Positive)**.** **B** *is non-dominated if*[11] *there is a probability function* Pr *that represents* **B**.

To appreciate the significance of Theorem 2, it helps to think about a standard lottery case.[12] Consider a fair lottery with n tickets, exactly one of which is the winner. For each $j \leqslant n$ (for $n \geqslant 3$), let p_j be the proposition that the j^{th} ticket is not the winning ticket. And, let q be the proposition that some ticket is the winner. Finally, let Lottery be the following belief set:

$$\left\{B(p_j) \mid 1 \leqslant j \leqslant n\right\} \cup \{B(q)\}\,.$$

Lottery is clearly non-dominated (just consider the probability function that assigns each ticket equal probability of winning), but it is not logically consistent. This explains why (NDB) is strictly weaker than (CB). Moreover, this example is a nice illustration of the fact that (NDB) is weaker than (CB) in a desirable way. More precisely, we can now show that (NDB) is entailed by both alethic considerations [(TB)/(CB)] and evidential considerations [(EB)].

While there is much disagreement about the precise content of (EB), there is widespread agreement that the following is a necessary condition for (EB).

> **Necessary Condition for Satisfying (EB)**. S satisfies (EB), *i.e.*, all of S's judgments are *justified*, ***only*** *if*:
>
> ($\mathcal{R}$) There exists *some* probability function that probabilifies (*i.e.*, assigns probability greater than 1/2 to) each of S's beliefs and dis-probabilifies (*i.e.*, assigns probability less than 1/2 to) each of S's disbeliefs.

Many evidentialists agree that probabilification — relative to some probability function — is a necessary condition for justification. Admittedly, there is a lot of dis-

[11] For counterexamples to the converse of Theorem 2, see Easwaran & Fitelson 2013.

[12] We are *not endorsing* the belief set Lottery in this example as *epistemically rational*. Indeed, we think that the lottery paradox is not as compelling — as a counterexample to (CB) — as the preface paradox is. On this score, we agree with Pollock (1990) and Nelkin (2000). We are just using this lottery example to make a formal point about the logical relationship between (CB) and (NDB).

agreement about which probability function is implicated in ($\mathcal{R}$).[13] But, because our Theorem 2 only requires the existence of some probability function that probabilifies *S*'s beliefs and dis-probabilifies *S*'s disbeliefs, it is sufficient to ensure (on most evidentialist views) that (EB) entails (NDB). And, given our assumptions about prefaces (and perhaps even lotteries), this is precisely the entailment that fails for (CB). Thus, by grounding coherence for full beliefs in the same way Joyce grounds probabilism for credences, we are naturally led to a coherence requirement for full belief that is a plausible alternative to (CB). This gives us a principled way to reject (CB), and to offer a new type of response to preface cases (and other similar paradoxes of consistency). Figure 2 depicts the logical relations between the requirements and norms discussed in this section.

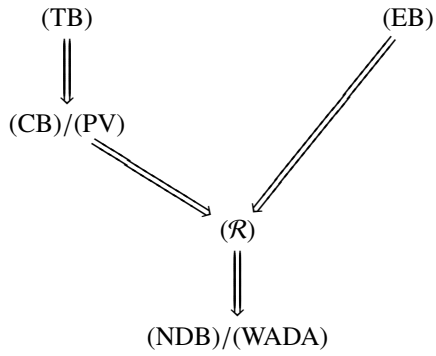


Figure 2: Logical relations between the requirements and norms for full belief

Here, I have presented a highly abridged rendition of the application of our framework to the case of full belief. A much more detailed and comprehensive version of our account (and its applications) can be found in (Easwaran and Fitelson 2013, Briggs et al. 2014, Fitelson 2014). In the next section, we explain how our framework can be fruitfully applied to comparative confidence judgments.

[13] Internalists like Fumerton (1995) require that the function Pr(·) which undergirds (EB) should be "internally accessible" to the agent (in various ways). Externalists like Williamson (2000) allow for "inaccessible" evidential probabilities. And, subjective Bayesians like Joyce (2005) say that Pr(·) should reflect the agent's subjective degrees of belief (*viz.*, credences). Despite this disagreement, most evidentialists agree that (EB) entails ($\mathcal{R}$), which is all we need for present purposes.

4 Application #2: Comparative confidence[14]

4.1 Setting up our framework for comparative confidence

In this section, we will be concerned with a relational epistemic attitude that we will call *comparative confidence*. We will use the notation $\ulcorner p \geq q \urcorner$ to express the comparative confidence relation, which can be glossed as $\ulcorner S$ is *at least as confident* in the truth of p as they are in the truth of $q \urcorner$ (where S is some epistemic agent a some time t).[15] For the purposes of the present paper, we will make various simplifying assumptions about $\geq$. These simplifying assumptions are not essential to our overall approach, but they will make it easier to introduce the main ideas involved in our justifications of epistemic coherence requirements for $\geq$.

Our first simplifying assumption is that our agents S form judgments regarding (pairs of) propositions drawn from a *finite Boolean algebra* of propositions $\mathcal{B}$. In other words, one can think of the p's and q's in judgments of the form $\ulcorner p \geq q \urcorner$ as *classical, possible-worlds propositions*. We will also assume (for simplicity) a weak form of *logical omniscience*, according to which agents always make the same judgments regarding logically equivalent propositions. Finally, we will assume that the relation $\geq$ constitutes a *total preorder* on the Boolean algebra $\mathcal{B}$. That is, we will assume that $\geq$ satisfies the following two *ordering conditions*.[16]

Totality. For all $p, q \in \mathcal{B}$, *either* $p \geq q$ *or* $q \geq p$.

Transitivity. For all $p, q, r \in \mathcal{B}$, *if* $p \geq q$ and $q \geq r$, *then* $p \geq r$.

With $\geq$ in hand, we can define a "*strictly* more confident than" relation $>$, as follows

$$p > q =_{\text{df}} p \geq q \text{ and } q \ngeq p.$$

[14] This section draws heavily on joint work with David McCarthy (Fitelson and McCarthy 2013).

[15] It is difficult to articulate the intended meaning of $\ulcorner p \geq q \urcorner$ without implicating that the $\geq$-relation reduces to (or essentially involves) some *non-relational* comparison of degrees of confidence b of the agent S (*e.g.*, $b(p) \geq b(q)$). But, it is important that no such reductionist assumption be made in the present context. Later in the paper, we will discuss issues of numerical *representability* of $\geq$-relations. But, the reader should assume that $\geq$ is an *autonomous* relational attitude, which may not (ultimately) reduce to (or essentially involve) something non-relational. Other glosses on $\ulcorner p \geq q \urcorner$ have been given in the literature, *e.g.*, $\ulcorner S$ judges p to be no less believable/plausible than $q \urcorner$.

[16] We are well aware of the fact that each of these total preorder assumptions have been a source of controversy in the literature on coherence requirements for comparative confidence relations. See (Forrest 1989, Fishburn 1986, Lehrer and Wagner 1985) for discussion. But, we have chosen (in this initial investigation) to simplify things by bracketing controversies about the *order structure* of $\geq$. We will address those issues (which we think will require a different sort of treatment in any event) in future work (Fitelson 2014).

And, we can define an "equally confident in" (or "epistemically indifferent between") relation $\sim$, as follows

$$p \sim q =_{\text{df}} p \succeq q \text{ and } q \succeq p.$$

Because $\succeq$ is a total preorder on $\mathcal{B}$, it will follow that $\succ$ is an *asymmetric, transitive, irreflexive* relation on $\mathcal{B}$; and, it will also follow that $\sim$ is an *equivalence relation* on $\mathcal{B}$. In other words, for each pair of propositions $p, q \in \mathcal{B}$ our agent(s) will be such that *either* $p \succ q$ *or* $q \succ p$ *or* $p \sim q$, where these three relations have the usual ordering properties one would naturally be inclined to attribute to them.

What we're interested in presently is providing an *epistemic justification* for various sorts of *coherence requirements — above and beyond* the (not uncontroversial *fn.* 16) assumption that $\succeq$ is a total preorder — that have been proposed for $\succeq$ in the contemporary literature. But, before we do that, we'll need to say a little bit more about how we're going to *represent* $\succeq$-relations.

One convenient way to represent a $\succeq$-relation on a Boolean algebra $\mathcal{B}_n$ containing n propositions is *via* its *adjacency matrix*. Let $p_1, \ldots, p_n$ be the n propositions contained in some Boolean algebra $\mathcal{B}_n$. The adjacency matrix $A^{\succeq}$ of a $\succeq$-relation on $\mathcal{B}_n$ is an $n \times n$ matrix of zeros and ones such that $A^{\succeq}_{ij} = 1$ iff $p_i \succeq p_j$.

It's instructive to look at a simple example. Consider the simplest Boolean algebra $\mathcal{B}_4$, which is generated by a single contingent claim P. This algebra $\mathcal{B}_4$ contains the following four propositions: $\langle p_1, p_2, p_3, p_4 \rangle = \langle \top, P, \neg P, \bot \rangle$. To make things concrete, let P be the claim that a fair coin (which is about to be tossed) will land heads (so, $\neg P$ says the coin will land tails). Suppose our agent S is equally confident in (*viz.*, epistemically indifferent between) P and $\neg P$. And, suppose that S is strictly more confident in $\top$ than in any of the other propositions in $\mathcal{B}_4$, and that S is strictly less confident in $\bot$ than in any of the other propositions in $\mathcal{B}_4$. This description fully characterizes a $\succeq$-relation on $\mathcal{B}_4$, which has the adjacency matrix representation (and the graphical representation) depicted in Figure 3. In the adjacency matrix $A^{\succeq}$ of $\succeq$, a 1 appears in the $\langle i, j \rangle$-th cell just in case $p_i \succeq p_j$. In the graphical representation of $\succeq$, an arrow is drawn from p_i to p_j just in case $p_i \succeq p_j$. With our basic formal framework in hand, we are ready to proceed.

In the next section, we'll discuss a fundamental coherence requirement for $\succeq$ that has been accepted by (nearly) everyone in the contemporary literature. Then, we will layout our general framework for grounding $\succeq$-coherence requirements, and we will explain how our framework can be used to provide a compelling epistemic justification for this fundamental coherence requirement for $\succeq$.

$\succeq$	$\top$	P	$\neg P$	$\bot$
$\top$	1	1	1	1
P	0	1	1	1
$\neg P$	0	1	1	1
$\bot$	0	0	0	1

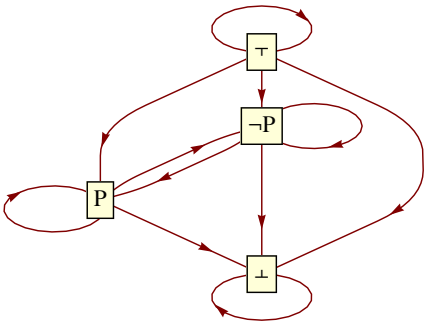


Figure 3: Adjacency matrix $A^{\succeq}$ and graphical representation of an intuitive $\succeq$-relation on the smallest Boolean algebra $\mathcal{B}_4$

4.2 The fundamental coherence requirement for $\succeq$

The literature on coherence requirements for $\succeq$ has become rather extensive. A plethora of coherence requirements of varying degrees of strength, *etc.*, have been proposed and defended. We will not attempt to survey all of these requirements here.[17] Instead, we will focus on the most fundamental of the existing coherence requirements, which is common to all the approaches we have seen.

A *plausibility measure* (*a.k.a.*, a *capacity*) on a Boolean algebra $\mathcal{B}$ is real-valued function $\mathrm{Pl} : \mathcal{B} \mapsto [0, 1]$ which maps propositions from $\mathcal{B}$ to the unit interval, and which satisfies the following three axioms (Halpern 2003, *p.* 51)

(Pl_1) $\mathrm{Pl}(\bot) = 0$.

(Pl_2) $\mathrm{Pl}(\top) = 1$.

(Pl_3) For all $p, q \in \mathcal{B}$, if p entails q then $\mathrm{Pl}(q) \geq \mathrm{Pl}(p)$.

The *fundamental coherence requirement* for $\succeq$ ($\mathfrak{C}$) can be stated in terms of *representability by a plausibility measure*. That is, here is one way of stating ($\mathfrak{C}$).

($\mathfrak{C}$) It is a requirement of ideal epistemic rationality that an agent's $\succeq$-relation (assumed to be a total preorder on a finite Boolean algebra $\mathcal{B}$) be *representable by*

[17] See (Halpern 2003) for an up-to-date and comprehensive survey. See, also, (Wong et al. 1991, Capotorti and Vantaggi 2000, Spohn 2012) and references therein.

some plausibility measure. That is, a $\succeq$-relation is *coherent only if* there exists some plausibility measure Pl such that for all $p, q \in \mathcal{B}$

$$p \succeq q \text{ iff } \mathrm{Pl}(p) \geq \mathrm{Pl}(q).$$

It is well known (Capotorti and Vantaggi 2000) that ($\mathfrak{C}$) can be stated *via axiomatic constraints* on $\succeq$, as follows:

($\mathfrak{C}$) It is a requirement of ideal epistemic rationality that an agent's $\succeq$-relation (assumed to be a total preorder on a finite Boolean algebra $\mathcal{B}$) satisfy the following two axiomatic constraints

(A_1) $\top \succ \bot$.

(A_2) For all $p, q \in \mathcal{B}$, if p entails q then $q \succeq p$.

In words, ($\mathfrak{C}$) requires (A_1) that an agent's $\succeq$-relation ranks tautologies *strictly above* contradictions, and (A_2) that an agent's $\succeq$-relation lines up with the "is deductively entailed by" (*viz.*, the "logically follows from") relation.

As far as we know, despite the (nearly) universal acceptance of ($\mathfrak{C}$) as a coherence requirement for $\succeq$, no *epistemic* justification has been given for ($\mathfrak{C}$). Various *pragmatic* justifications of requirements like ($\mathfrak{C}$) have been given. Starting with Ramsey (1926), the most well-known arguments for these sorts of constraints on $\succeq$ as a formal, synchronic, coherence requirements for comparative confidence have been *pragmatic*. For instance, "Money Pump" arguments and "Representation Theorem" arguments (Savage 1972, Halpern 2003) aim to show that agents with $\succeq$-relations that violate ($\mathfrak{C}$) must exhibit some sort of "pragmatic defect". Following Joyce (1998, 2009), we will be focusing on *non*-pragmatic (*viz.*, *epistemic*) defects implied by the synchronic incoherence (in a precise sense to be explicated below) of an agent's $\succeq$-relation. To be more specific, we will be concerned with the *accuracy* of an agent's $\succeq$-relation (in a precise sense to be explicated below), which we will take to be *distinctively epistemic*.

Next, we'll explain our general (broadly Joycean) strategy for grounding epistemic coherence requirements for $\succeq$. This will allow us to explain *why* ($\mathfrak{C}$) is a requirement of ideal *epistemic* rationality. Moreover, our explanation will be a *unified and principled* one, which dovetails with the similar explanations for credence and full belief rehearsed above. As in the cases of credence and full belief, applying our general framework requires completing the three steps.

4.3 Grounding the fundamental requirement (𝔈)

Step 1: The vindicated ≥-relation

We will adopt the (Joycean) idea that the vindicated confidence ordering $\overset{\circ}{\geq}_w$ ranks all truths (in w) *strictly above* all falsehoods (in w); and, we will also assume that $\overset{\circ}{\geq}_w$ is *indifferent* between propositions having the same truth-value (in w). That is, we will assume the following definition

$$p \overset{\circ}{\geq}_w q =_{\text{df}} \begin{cases} p \succ q & \text{if } p \text{ is true in } w \text{ and } q \text{ is false in } w \\ p \sim q & \text{if } p \text{ and } q \text{ have the same truth-value in } w \end{cases}$$

It is easy to see that this definition determines a *unique vindicated total preorder* in each possible world. Here, we follow Joyce in adopting an *extensional* definition of the vindicated/perfectly accurate judgment set in w — *i.e.*, we assume that $p \overset{\circ}{\geq}_w q$ *is determined solely by the truth-values of* p *and* q *in* w. We think the $\succ$ clause of the definition of $\overset{\circ}{\geq}_w$ is less controversial than the $\sim$ clause. Indeed, our main focus here will be on grounding coherence requirements for $\succ$.[18]

Step 2: Distance from the vindicated ≥-relation

In the case of Joyce's argument for probabilism, this step has proved to be the most controversial. It turns out that Joyce's argument is very sensitive to his choice of measure of distance from vindication (Maher 2002). We won't get into that controversy here. Moreover, as in the case of full belief above, we will adopt a very naïve measure of distance between comparative confidence orderings.[19] For simplicity, we will present our argument for (𝔈) using the simplest (and most well known and widely used) measure of distance between finite binary relations: *Kemeny distance*. Kemeny and Snell (1962) give an axiomatic argument in favor of a measure $\mathfrak{d}(\geq_1, \geq_2)$ of distance between, which is equivalent to the following definition

$$\mathfrak{d}(\geq_1, \geq_2) =_{\text{df}} \text{the number of cells } \langle i, j \rangle \text{ such that } A_{ij}^{\geq_1} \neq A_{ij}^{\geq_2}.$$

[18] Grounding coherence requirements for $\sim$ turns out to be a subtle and tricky affair. For simplicity, we will largely ignore the question of explicating the proper epistemic requirements for $\sim$ in this *précis*. See *fn.* 21 and (Fitelson and McCarthy 2013) for discussion.

[19] As in the case of full belief, our arguments here will not (ultimately) depend that sensitively on our particular (naïve) choice of distance measure. We omit that dialectic here, but see (Fitelson and McCarthy 2013) and (Fitelson 2014) for discussion.

That is, $\mathfrak{d}(\succeq_1, \succeq_2)$ just *counts the number of (point-wise) differences between the adjacency matrices of* $\succeq_1$ and $\succeq_2$. This is equivalent to counting the number of pairs $\langle i, j\rangle$ such that the two relations $\succeq_1$ and $\succeq_2$ *disagree* regarding whether $p_i \succeq p_j$. An illustrative example is helpful here. Recall our toy agent who forms judgments on the simplest Boolean algebra $\mathcal{B}_4$. Let S's $\succeq$-relation be given by the intuitive ordering depicted in Figure 3. Now, because there are only two salient possible worlds in this case, we only have two vindicated $\succeq$-relations to consider. Let $\overset{\circ}{\succeq}_1$ be the vindicated $\succeq$-relation in world w_1 in which P is false, and let $\overset{\circ}{\succeq}_2$ be the vindicated $\succeq$-relation in world w_2 in which P is true. The adjacency matrices of $\overset{\circ}{\succeq}_1$ and $\overset{\circ}{\succeq}_2$ are depicted in Figure 4.

$\overset{\circ}{\succeq}_1$	$\top$	P	$\neg P$	$\bot$
$\top$	1	1	1	1
P	0	1	0	1
$\neg P$	1	1	1	1
$\bot$	0	1	0	1

$\overset{\circ}{\succeq}_2$	$\top$	P	$\neg P$	$\bot$
$\top$	1	1	1	1
P	1	1	1	1
$\neg P$	0	0	1	1
$\bot$	0	0	1	1

Figure 4: The adjacency matrices of the vindicated $\succeq$-relations (over $\mathcal{B}_4$) in worlds w_1 (P false) and w_2 (P true), respectively

With all three of the salient adjacency matrices in front of us (in Figures 3 and 4), it is easy to calculate the values of $\mathfrak{d}(\succeq, \overset{\circ}{\succeq}_1)$ and $\mathfrak{d}(\succeq, \overset{\circ}{\succeq}_2)$. For $\mathfrak{d}(\succeq, \overset{\circ}{\succeq}_1)$, all we have to do is count the number of cells in $A^{\succeq}$ and $A^{\overset{\circ}{\succeq}_1}$ that differ. Inspection of these matrices reveals that there are three cells: $\langle 3, 1\rangle$, $\langle 4, 2\rangle$, and $\langle 2, 3\rangle$ at which $A^{\succeq}$ and $A^{\overset{\circ}{\succeq}_1}$ differ. Thus, $\mathfrak{d}(\succeq, \overset{\circ}{\succeq}_1) = 3$. A similar inspection reveals that there are also three cells: $\langle 2, 1\rangle$, $\langle 3, 2\rangle$, and $\langle 4, 3\rangle$ at which $A^{\succeq}$ and $A^{\overset{\circ}{\succeq}_2}$ differ. Thus, $\mathfrak{d}(\succeq, \overset{\circ}{\succeq}_2) = 3$. In other words, $\succeq$ is equidistant from $\overset{\circ}{\succeq}_1$ and $\overset{\circ}{\succeq}_2$, according to our (Kemeny) measure of distance from vindication. This brings us to our third and final Step.

Step 3: The fundamental epistemic principle for $\succeq$

For the purposes of grounding ($\mathfrak{C}$), we will adopt the same fundamental epistemic principle that Joyce used — *strict accuracy dominance avoidance* (SADA), *i.e.*, the following principle

> **Strict Accuracy-Dominance Avoidance** (SADA). $\succeq$ should *not be strictly dominated* in distance from vindication. Or, to put this more formally (in terms of $\mathfrak{d}$), there should *not* exist a relation $\succeq'$ on $\mathcal{B}$ such that
>
> $$(\forall w)\left[\mathfrak{d}(\succeq', \overset{\circ}{\succeq}_w) < \mathfrak{d}(\succeq, \overset{\circ}{\succeq}_w)\right].$$

As in the case of full belief, the avoidance of dominance in distance from vindication is a very basic principle of epistemic utility theory. If one violates (SADA), then one is (ideally, *a priori*) in a position to know that *one must not be living up to one's epistemic aim of having accurate judgments*. Interestingly, (SADA) is sufficient to ground (𝔈). That is, we have the following fundamental theorem.

Theorem 3. *If* $\geq$ *violates (*𝔈*), then* $\geq$ *violates (SADA). That is, (SADA) entails (*𝔈*).*

Next, we will peek beyond (𝔈) to stronger coherence requirements for $\geq$ that have appeared in the literature. As we'll see, (SADA) can be used to provide epistemic justifications for a large family of coherence requirements for $\geq$.

4.4 Coherence requirements stronger than (𝔈)

The fundamental requirement (𝔈) is but one member of a family of coherence requirements has been proposed for $\geq$. We will not survey all of the requirements in this family here. We'll focus on a handful of members of the family. Before stating the other requirements in the family, we'll first need to define two more numerical functions that will serve as *representers* of comparative confidence relations.

A *mass function* on a Boolean algebra $\mathcal{B}$ is real-valued function $m : \mathcal{B} \mapsto [0, 1]$ which maps propositions from $\mathcal{B}$ to the unit interval, and which satisfies the following two axioms.

(M_1) $m(\bot) = 0$.

(M_2) $\sum_{p \in \mathcal{B}} m(p) = 1$.

A *belief function* on a Boolean algebra $\mathcal{B}$ is a real-valued function $\mathrm{Bel} : \mathcal{B} \mapsto [0, 1]$ which maps propositions from $\mathcal{B}$ to the unit interval, and which is generated by an underlying mass function m in the following way

$$\mathrm{Bel}_m(p) =_{\mathrm{df}} \sum_{\substack{q \in \mathcal{B} \\ q \text{ entails } p}} m(q).$$

It is easy to show that all belief functions are plausibility functions (but not conversely). In this sense, the concept of a belief function is a refinement of the concept of a plausibility function. The class of Belief functions, in turn, contains the class of *probability* functions, which can be defined in terms of a special type of mass function. Let

$\mathfrak{s} \in \mathcal{B}$ be the *states* of a Boolean algebra $\mathcal{B}$ (or the *state descriptions* of a propositional language $\mathcal{L}$ which generates $\mathcal{B}$). A *probability* mass function is real-valued function $\mathfrak{m} : \mathcal{B} \mapsto [0, 1]$ which maps *states* of $\mathcal{B}$ to the unit interval, and which satisfies the following two axioms.

$(\mathfrak{M}_1)$ $\mathfrak{m}(\bot) = 0$.

$(\mathfrak{M}_2)$ $\sum_{\mathfrak{s} \in \mathcal{B}} \mathfrak{m}(\mathfrak{s}) = 1$.

A *probability function* on a Boolean algebra $\mathcal{B}$ is a real-valued function $\mathrm{Pr} : \mathcal{B} \mapsto [0, 1]$ which maps propositions from $\mathcal{B}$ to the unit interval, and which is generated by an underlying probability mass function $\mathfrak{m}$ in the following way

$$\mathrm{Pr}_{\mathfrak{m}}(p) =_{\mathrm{df}} \sum_{\substack{\mathfrak{s} \in \mathcal{B} \\ \mathfrak{s} \text{ entails } p}} \mathfrak{m}(\mathfrak{s}).$$

It is easy to show that all probability functions are belief functions (but not conversely). So, probability functions are special kinds of belief functions (and belief functions are, in turn, special kinds of plausibility measures).

There are various senses in which a real-valued function f may be said to *represent* a comparative confidence relation $\succeq$. We will say that f *fully agrees* with a comparative confidence relation $\succeq$ just in case, for all $p, q \in \mathcal{B}$, $p \succeq q$ *if and only if* $f(p) \geq f(q)$. Thus, the fundamental coherence requirement ($\mathfrak{C}$) requires that there exist a plausibility measure Pl which *fully agrees* with $\succeq$. There is a weaker kind of numerical representability that will play an important role for us. A real-valued function f is said to *partially agree* with a comparative confidence relation $\succeq$ just in case, for all $p, q \in \mathcal{B}$, $p \succ q$ *only if* $f(p) > f(q)$. If f partially agrees with $\succeq$, then we will say that f *partially represents* $\succeq$. And, if f fully agrees with $\succeq$, then we will say that f *fully represents* $\succeq$. It is easy to see that full representability (of $\succeq$ by f) is *strictly stronger* than partial representability (of $\succeq$ by f).

It is well known (Wong et al. 1991) that a total preorder $\succeq$ is partially represented by some belief function Bel just in case $\succeq$ satisfies (A_2). The following theorem is, therefore, an immediate corollary of Theorem 3.

Theorem 4. *(SADA) entails that $\succeq$ is partially represented by some belief function Bel. That is, (SADA) entails that there exists a belief function Bel such that, for all $p, q \in \mathcal{B}$, $p \succ q$ only if $Bel(p) > Bel(q)$.*

A natural question to ask at this point is whether (SADA) ensures that $\succeq$ is *fully* represented by some belief function Bel. Interestingly, the answer to this question is *no*. In order to see this, it helps to recognize that full representability by a belief function has a simple axiomatic characterization (Wong et al. 1991). Specifically, a total preorder $\succeq$ is fully represented by some belief function just in case $\succeq$ satisfies (A_1), (A_2), and

(A_3) If p entails q and q, r are mutually exclusive, then:

$$q \succ p \Longrightarrow q \vee r \succ p \vee r.$$

Theorem 3 establishes that (SADA) entails both (A_1) and (A_2). However, it turns out that (SADA) is not quite strong enough to entail (A_3). That is, we have the following (negative) theorem regarding (SADA).

Theorem 5. *(SADA) does not entail (A_3). [As a result, (SADA) is not strong enough to ensure that $\succeq$ is fully represented by some belief function (Wong et al. 1991).]*

Let's take stock. So far, we have encountered the following three coherence requirements for $\succeq$, in increasing order of strength.

($\mathfrak{C}_0$) $\succeq$ should be partially representable by some belief function Bel. This is equivalent to requiring that $\succeq$ (a total preorder) satisfies (A_2).

($\mathfrak{C}$) $\succeq$ should be fully representable by some plausibility measure Pl. This is equivalent to requiring that $\succeq$ (a total preorder) satisfies (A_1) and (A_2).

($\mathfrak{C}_1$) $\succeq$ should be fully representable by some belief function Bel. This is equivalent to requiring that $\succeq$ (a total preorder) satisfies (A_1), (A_2), and (A_3).

Moving beyond ($\mathfrak{C}_1$) takes us into the realm of *comparative probability*. A total preorder $\succeq$ is said to be a *comparative probability* relation just in case $\succeq$ satisfies (A_1) and the following two additional axioms.

(A_4) For all $p \in \mathcal{B}$, $p \succeq \bot$.

(A_5) For all $p, q, r \in \mathcal{B}$, if p, q are mutually exclusive and p, r are mutually exclusive, then:

$$q \succeq r \Longleftrightarrow p \vee q \succeq p \vee r.$$

It is easy to show that $\{(A_1), (A_2)\}$ jointly entail (A_4). So, $\geq$ (a total preorder) is a comparative probability relation just in case $\geq$ satisfies the three axioms (A_1), (A_2) and (A_5). Now, consider the following coherence requirement.

$(\mathfrak{C}_2)$ $\geq$ should be be a comparative probability relation. This is equivalent to requiring that $\geq$ (a total preorder) satisfies (A_1), (A_2) and (A_5).

It is well known (and not too difficult to prove) that (A_5) is *strictly stronger* than (A_3), in the presence of (A_1) and (A_2). Therefore, $(\mathfrak{C}_2)$ is *strictly stronger* than $(\mathfrak{C}_1)$. The following axiomatic constraint is a weakening of (A_5).

$(A_5^\star)$ For all $p, q, r \in \mathcal{B}$, if p, q are mutually exclusive and p, r are mutually exclusive, then:

$$q \succ r \Longrightarrow p \vee r \not\succ p \vee q$$

The following coherence requirement is a (corresponding) weakening of $(\mathfrak{C}_2)$.

$(\mathfrak{C}_2^\star)$ $\geq$ should (be a total preorder and) satisfy (A_1), (A_2) and $(A_5^\star)$.

De Finetti (1937, 1951) famously conjectured that all comparative probability relations are fully representable by some probability function. As it turns out, this conjecture is false. In fact, Kraft et al. (1959) showed that some comparative probability relations are *not even partially* representable by any probability function.[20] That brings us to our final two coherence requirements for $\geq$, which we add (in increasing order of strength) to our family of $\geq$-requirements.

$(\mathfrak{C}_3)$ $\geq$ should be be partially representable by some probability function.

$(\mathfrak{C}_4)$ $\geq$ should be be fully representable by some probability function.

In light of the counterexample of Kraft et al. (1959), $(\mathfrak{C}_4)$ is *strictly stronger* than $(\mathfrak{C}_2)$, and $(\mathfrak{C}_3)$ does not follow from $(\mathfrak{C}_2)$. In fact, it is straightforward to show that $(\mathfrak{C}_3)$ entails $(\mathfrak{C}_2^\star)$, but $(\mathfrak{C}_3)$ is *independent* of $(\mathfrak{C}_2)$ and $(\mathfrak{C}_1)$.

Figure 5 depicts the logical relations between the $\geq$-coherence requirements we've been discussing. The superscripts on the coherence requirements in Figure 5 have the

[20] We won't enter into the fascinating subsequent historical dialectic here. But, we discuss it in some detail in (Fitelson and McCarthy 2013).

following meanings. If a coherence requirement is known to follow from (SADA), then it gets a "✓". If a coherence requirement is known *not* to follow from (SADA), then it gets an "✗". If it is an open question whether (SADA) entails a coherence requirement, then it gets a "?". The "✓"s on $(\mathfrak{C}_0)$ and $(\mathfrak{C})$ are implied by Theorem 3, above. And, the other "✓" [for $(\mathfrak{C}_2^\star)$] is implied by the following theorem.

Theorem 6. *If $\succeq$ violates $(\mathfrak{C}_2^\star)$, then $\succeq$ violates (SADA). That is, (SADA) entails $(\mathfrak{C}_2^\star)$.*

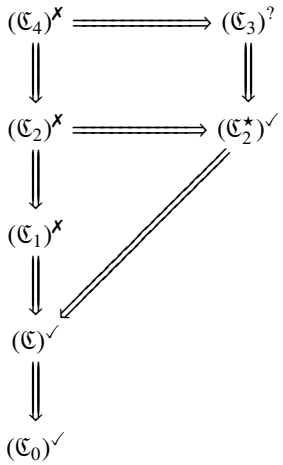


Figure 5: Logical relations between $\succeq$-coherence requirements

Because probability functions are refinements of belief functions, the existence of a *probabilistic* (partial or full) representation of $\succeq$ is a *strictly stronger* than the existence of a (partial or full) belief-function representation of $\succeq$. It is, therefore, an immediate corollary of Theorem 5 that (SADA) is not strong enough to ensure that $\succeq$ is fully represented by a probability function (indeed, our proof of Theorem 5 establishes *all three* of the "✗"s in Figure 5).[21] However, this leaves open the question of whether

[21] As we mentioned above (*fn.* 18), determining the proper vindication constraints and coherence requirements for $\sim$ is tricky. We have adopted a naïve, extensional definition of vindication for $\sim$. This is the reason why we have only been able to establish *partial* numerical representability results for $\succeq$ (*viz.,* complete numerical representability results for $\succ$) *via* (SADA). We don't have the space here to delve into the subtleties of $\sim$-coherence. But, we have more to say about these thorny questions in (Fitelson and McCarthy 2013) and (Fitelson 2014).

(SADA) is sufficient to ensure that $\succeq$ is *partially representable* by a probability function (which explains the "?" superscript on ($\mathfrak{C}_3^\star$) in Figure 5). We have made some headway toward settling this question (Fitelson and McCarthy 2013), but it remains open.

5 Conclusions

Inspired by the arguments of de Finetti and Joyce (for probabilism), we have developed a general framework for grounding (formal, synchronic, epistemic) coherence requirements for various types of judgment. We have shown how to fruitfully apply this framework to the cases of full belief and comparative confidence. This has been but a *précis* of a larger project. For a more thorough discussion of our framework and its applications, see (Easwaran and Fitelson 2013, Briggs et al. 2014, Fitelson and McCarthy 2013, Fitelson 2014).

Acknowledgements I would like to thank my collaborators Rachael Briggs, Fabrizio Cariani, Kenny Easwaran and David McCarthy. The work I have presented here is largely joint work with them. I would also like to thank Sonja Smets and the rest of the ILLC team for inviting me to present this material (in a LIRa Seminar and in this yearbook), and also for inviting me to teach a master class (in June 2014) on my book manuscript (Fitelson 2014), which contains a much more comprehensive and detailed treatment of the topics discussed here. Finally, I would like to thank Nina Gierasimczuk for her editorial assistance.

References

R. Briggs, F. Cariani, K. Easwaran, and B. Fitelson. Individual coherence and group coherence. In J. Lackey, editor, *Essays in Collective Epistemology*. Oxford University Press, 2014. To appear.

A. Capotorti and B. Vantaggi. Axiomatic characterization of partial ordinal relations. *International Journal of Approximate Reasoning*, 24(2):207–219, 2000.

D. Christensen. *Putting logic in its place*. Oxford University Press, 2004.

B. de Finetti. La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937.

B. de Finetti. *La 'Logica del Plausible' secondo la concezione di Polya*. Società Italiana per il Progresso delle Scienze, 1951.

B. de Finetti. *Theory of probability*. Wiley, 1970.

M. Deza and E. Deza. *Encyclopedia of distances*. Springer, 2009.

C. Duddy and A. Piggins. A measure of distance between judgment sets. *Social Choice and Welfare*, 39(4):855–867, 2012.

K. Easwaran. Dr. Truthlove, or how I learned to stop worrying and love Bayesian probabilities. Manuscript, 2013.

K. Easwaran and B. Fitelson. Accuracy, coherence and evidence. to appear in *Oxford Studies in Epistemology, Vol. 5*, 2013.

P. C. Fishburn. The axioms of subjective probability. *Statistical Science*, 1(3):335–345, 1986.

B. Fitelson. *Coherence*. 2014. Book manuscript (in progress).

B. Fitelson and D. McCarthy. New foundations for comparative confidence. Manuscript, 2013.

R. Foley. *Working without a net*. Oxford University Press, 1992.

P. Forrest. The problem of representing incompletely ordered doxastic systems. *Synthese*, 79(2):279–303, 1989.

R. Fumerton. *Metaepistemology and skepticism*. Rowman & Littlefield, 1995.

A. Hájek. Arguments for–or against–probabilism? *The British Journal for the Philosophy of Science*, 59(4):793–819, 2008.

J. Y. Halpern. *Reasoning about uncertainty*. MIT Press, 2003.

J. M. Joyce. A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65 (4):575–603, 1998.

J. M. Joyce. How probabilities reflect evidence. *Philosophical Perspectives*, 19(1): 153–178, 2005.

J. M. Joyce. Accuracy and coherence: prospects for an alethic epistemology of partial belief. In F. Huber and C. Schmidt-Petri, editors, *Degrees of Belief*. Springer, 2009.

J. G. Kemeny and J. L. Snell. *Mathematical models in the social sciences*, volume 9. Ginn, 1962.

C. H. Kraft, J. W. Pratt, and A. Seidenberg. Intuitive probability on finite sets. *The Annals of Mathematical Statistics*, 30(2):408–419, 1959.

K. Lehrer and C. Wagner. Intransitive indifference: The semi-order problem. *Synthese*, 65(2):249–256, 1985.

P. Maher. Joyce's argument for probabilism. *Philosophy of Science*, 69(1):73–81, 2002.

M. K. Miller and D. Osherson. Methods for distance-based judgment aggregation. *Social Choice and Welfare*, 32(4):575–601, 2009.

D. K. Nelkin. The lottery paradox, knowledge, and rationality. *The Philosophical Review*, 109(3):373–409, 2000.

G. Pigozzi. Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. *Synthese*, 152(2):285–298, 2006.

J. L. Pollock. *Nomic probability and the foundations of induction*. Oxford Univ. Press, 1990.

F. P. Ramsey. Truth and probability. In R. Braithwaite, editor, *Foundations of Mathematics and other Logical Essays*, pages 156–198. Kegan, Paul, Trench, Trubner & Co., 1926.

L. Savage. *The foundations of statistics*. Dover, 1972.

M. Schervish, T. Seidenfeld, and J. Kadane. Proper scoring rules, dominated forecasts, and coherence. *Decision Analysis*, 6(4):202–221, 2009.

W. Spohn. *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford University Press, 2012.

T. Williamson. *Knowledge and its limits*. Oxford University Press, 2000.

S. M. Wong, Y. Yao, P. Bollmann, and H. Burger. Axiomatization of qualitative belief structure. *IEEE Transactions on Systems, Man and Cybernetics*, 21(4):726–734, 1991.

Thinking about Knowledge. Some Fresh Views

Rohit Parikh

Brooklyn College and CUNY Graduate Center
rparikh@gc.cuny.edu

Abstract

The discussion of knowledge has been dominated among computer scientists and logicians by Kripke structures and their progeny. This tendency has led to exciting technical developments but often falls short of reaching actual human practice, both at the individual and social level. Also, the question, "But where do these Kripke structures come from?" tends to be thrown under the rug. We offer some ideas.

1 Introduction

I will start with two examples from literature. One is Shakespeare's Hamlet, in a play with the same name. Hamlet, while unwise in some other ways, does have some epistemic skills.

According to the play, it is given out that Hamlet's father, the king of Denmark, was bitten by a serpent while sleeping in his garden. His brother succeeds to the throne and marries Hamlet's mother. Hamlet arrives back to Denmark and wonders if he has come to his father's funeral or to his mother's wedding.

But then Hamlet is told by his father's ghost (or what appears to be his father's ghost) a rather different story. The ghost says,

> *Now, Hamlet, hear: 'Tis given out that, sleeping in my orchard,*
> *A serpent stung me; so the whole ear of Denmark*
> *Is by a forged process of my death*

Rankly abused: but know, thou noble youth,
The serpent that did sting thy father's life
Now wears his crown.

Hamlet does not know what to do. If the ghost is right then it is Hamlet's duty to kill his uncle and avenge his father's death. But what if the ghost is lying?

Hamlet decides to put on a little play to check if the story told him by his father's ghost is indeed true.

I'll have these players
Play something like the murder of my father
Before mine uncle: I'll observe his looks;
I'll tent him to the quick: if he but blench,
I know my course. The spirit that I have seen
May be the devil: and the devil hath power
To assume a pleasing shape; yea, and perhaps
Out of my weakness and my melancholy,
As he is very potent with such spirits,
Abuses me to damn me: I'll have grounds
More relative than this: the play 's the thing
Wherein I'll catch the conscience of the king.

Hamlet does catch his uncle in a state of unpleasant surprise and decides to avenge his father's murder.

A play by Shaw, *The Man of Destiny* also has some epistemic considerations. A letter has been received by Napolean, presumably detailing his wife Josephine's infidelities. A pretty woman, acting on Josephine's behalf, appears before Napolean reads the letter and tries to use her beauty and her intelligence to talk him out of reading it. Napolean sees through her tricks and responds,

NAPOLEON (with coarse familiarity, treating her as if she were a vivandiere). Capital! Capital! (He puts his hands behind him on the table, and lifts himself on to it, sitting with his arms akimbo and his legs wide apart.) Come: I am a true Corsican in my love for stories. But I could tell them better than you if I set my mind to it. Next time you are asked why a letter compromising a wife should not be sent to her husband, answer simply that the husband would not read it. Do you suppose, little innocent, that a man wants to be compelled by public opinion to make a scene, to fight a duel, to break up his household, to injure his career by a scandal, when he can avoid it all by taking care not to know?

There are fascinating epistemic insights shown by these two great writers which logicians would do well to try and formalize.

Let us consider these two examples. First Hamlet.

Claudius never *tells*, Hamlet, "I killed your father." Rather, there is a default reaction on the part of Claudius if he is innocent and a default reaction (or at least a likely reaction) if he is guilty.

If Claudius is innocent, then his reaction to the play might be to enjoy it, to criticize it or to be bored. But his actual reaction is dismay and shock. He leaves the play halfway through. This may not be proof that he is guilty but as a good Bayesian, Hamlet must realize that now he has evidence independent of his father's ghost.

Napolean of course is much more savvy than Claudius. The issue here is, "Can a man *choose* not to know p"? Perhaps he does not want to *not know* p but merely not have others know that he does know p. If others know that Napolean knows that Josephine is unfaithful, then he would expected to act on the basis of this infidelity and he does not want to bear the cost. But he does actually want to know. After protesting that he does not plan to read the letter, he reads it when he is alone in the garden.

Let p stand for "Josephine is unfaithful", n for Napolean, and w for the world, thought of as one person. Then Napolean achieves

$$K_n(p) \text{ and } \neg K_w K_n(p)$$

These two are very sophisticated uses of epistemic reasoning by these two great writers, Shakespeare and Shaw (the only person to win both the Nobel prize and an Oscar!)

2 Putting your money where your mouth is

In a joint paper with Aranzazu San Gines, we consider the following story.

> *Imagine, for instance, the following situation. We are in Spain. Today is the morning of December 22nd, the day of the Christmas lottery. Most people have the radio on, hoping for their numbers to be winners. (A) and (B) are respectively a woman and her boyfriend who are having a coffee in a cafeteria. (A) receives a call. After the woman hangs up, the following dialog takes place:*
> *(A) We have won the lottery!!*
> *(B) What? How do you know that?*
> *(A) It was my father. He seemed quite excited. He said he has good news. He wants us to be at home in half an hour to tell us the news and celebrate. You see? I know it! We have won the lottery!*

(B) Really? The lottery? Would you call your boss right now and tell her that you quit your job?
(A) Mmmmm. OK, OK, you're right. I don't know for certain ...

Here B challenges A asking her to put her money where her mouth is. Is she prepared to resign her job based on her belief that her father has won the lottery? She backs down ending with I don't know for certain. If the father has some *other* piece of good news, then resigning her job would be a big mistake.

This case parallels the one discussed by Jason Stanley in his prize winning work (Stanley 2005).

Hannah and her husband are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. But as they drive past the bank, they notice that the lines inside are very long, as they often are on Friday afternoons. Thinking that it isn't very important that their paychecks are deposited right away, Hannah says "I know the bank will be open tomorrow, since I was there just two weeks ago on Saturday morning. So we can deposit them tomorrow morning."

But then Hannah's husband reminds her that a very important bill comes due on Monday, and that they have to have enough money in their account to cover it. He says, "Banks do change their hours. Are you certain that's not what is going to happen tomorrow?" Hannah concedes, uttering "I guess I don't really know that the bank will be open tomorrow."

In both cases, Hannah has the same evidence and let us also assume that the truth values are the same. Since there are no Gettier type issues here we can assume that knowledge coincides (here) with justified true belief. So whence the difference?

Clearly the difference in the two cases, "knowledge" and "not knowledge" arises from a difference in belief. But in fact the belief must be the same since the evidence is also the same. It is the *weight* placed on the belief which is at issue.

This line of attack by Stanley is a little like that due to Ramsey (1954) and later Savage (1972). The subjective probability of an event is determined by looking at the risks which the agent is willing to take. The risk posed by the non-deposit of the check is small initially but rises once the husband points out that there is a bill due. The probability of "The bank will be open on Saturday" is high enough to justify the first risk but not the second.

Ever since Ramsey (and perhaps earlier) we have accepted the idea that whether someone has a belief is revealed by the actions they take. If someone picks up an umbrella as she is going out we assume that she believes that it is raining or soon will.

Moreover, we tend to attribute knowledge, not necessarily on the basis of evidence, but on the basis of successful behavior. In an earlier paper I wrote about a mouse being

asked to choose between two boxes, one of which contains cheese and the other does not. I said,

> *Suppose now that the mouse invariably goes to that box which contains the cheese. We would then say, "he somehow knows where the cheese is", and would look for some evidence that might have given him a clue. Perhaps we would find such a clue, but our judgment that he knew where the cheese was would not depend on such a clue, but rather on our perception of successful behaviour on the part of the mouse which we were unable to otherwise understand.*

Presumably the belief of the mouse is revealed by the choices he makes. And if the mouse does find the cheese then there is truth as well. But what about *justification?* As Wittgenstein says in a similar context, "No such thing was in question here." Repeated success on the part of the mouse indicates knowledge (Peirce 1931-5).

3 Socrates' problem

The *justified true belief* account of knowledge is that knowing something is no more than having a justified belief that it is true, and indeed its being true. There is a common impression that the justified true belief (JTB) definition of knowledge is due to Plato and was undermined by Gettier in his 1963 paper (Gettier 1963).

Gettier himself says, "Plato seems to be considering some such definition at Theaetetus 201, and perhaps accepting one at Meno 98."

The *Stanford Encyclopedia of Philosophy* article on the Analysis of Knowledge [IS] says,

> *Socrates articulates the need for something like a justification condition in Plato's Theaetetus, when he points out that 'true opinion' is in general insufficient for knowledge. For example, if a lawyer employs sophistry to induce a jury into a belief that happens to be true, this belief is insufficiently well-grounded to constitute knowledge.*

Others who have attributed the JTB theory to Plato include Artemov and Nogina (2005).

However, a cursory look at the *Theaetetus* shows that Socrates at least did *not* endorse the JTB theory.

It is the boy Theaetetus (who was a mere 16 years old at the time) and not Socrates who proposes the JTB account after proposing two others, *knowledge as perception* and *knowledge as true belief*.

> *Oh, yes, Socrates, that's just what I once heard a man say; I had forgotten but it is now coming back to me. He said that it is true judgement with an account that is knowledge; true judgment without an account falls outside of knowledge.*

Socrates subjects Theatetus' assertion to rigorous analysis and finally undermines this third, JTB account, ending with the words,

> *therefore, knowledge is neither perception, nor true judgement, nor an account added to true judgement.*

The JTB account of knowledge, rather than being endorsed by Socrates, is explicitly rejected.

But what is of interest to us in this paper is Socrates' objection to JTB which is different from that of Gettier and arguably deeper.

Gettier's own undermining of JTB went the following route. Someone justifiably believes A.

He deduces B from A and indeed A implies B.

And B is true. So the belief in B is both justified and true.

However, unfortunately, A is false so that the belief in B, while justified, can't really be considered knowledge.

3.1 The nature of justification

Socrates does not go this route but instead asks what a justification might be like (the Greek term here for justification is *logos*, which translates roughly as 'account'.)

An analogy to justification here is an analysis of the first syllable SO of his own name. SO is composed of the two letters S and O and that spelling out is rather like a justification.

Both an analysis and a justification have structure and Socrates points out that the letters S and O, not having structure, cannot have an analysis.

But is it possible to know the syllable SO without knowing the letters S and O? And if not, then how can we rest a knowledge of SO on a knowledge of S and a knowledge of O?

This issue is also addressed by Wittgenstein who says, "Explanations come to an end somewhere".[1]

[1] There is actually a piece of music with that title.

Our knowledge of some facts is parasitical on our knowledge of some other facts. But eventually we hit bottom and then there is no more to be said. Like G.E. Moore, we know a hand when we see it.

4 Strategizing

In the movie *Basic Instinct 2*, the psychiatrist Michael Glass asks Catherine Tramell (played by Sharon Stone), "Did you kill Adam Towers?" and Tramell counters, "Would you believe me if I said I didn't?"

This issue has not been discussed much in epistemic literature but is prominent in the literature on cheap talk. Much current literature in epistemic logic assumes that messages sent are truthful. But the cheap talk literature considers the possibility of messages which might not be truthful and asks why someone would send an untruthful message and when such a message might be believed.

Such issues have been discussed at length by Stalnaker in his very readable paper (Stalnaker 2005).

Stalnaker does not assume that the parties are speaking the truth but rather that they will talk in such a way as to maximize their payoff in the game they are playing with the listener.

Stone is presumably not familiar with Stalnaker's work, but what she is saying is that a message "I didn't kill Johnny Boz" would not be credible[2] as the other message "I did kill Johnny Boz" would not be sent regardless of the situation.

What is interesting is that Tramell does not utter the message which would not be credible, but a meta-message. Why? Perhaps it serves a purpose a little like that of John Snow, below.

We tend to believe the speaker if we know that she has nothing to gain by lying (and is assumed to be well informed). But we might not believe Mr. Obama when he says that the NSA is not listening to our conversations. The NSA is *his* agency and he has nothing to gain by giving away the truth about it.

In the cited paper, Stalnaker discusses the case of the US Treasury Secretary John Snow, who in response to a question said, "When the dollar is at a lower level it helps exports, and I think exports are getting stronger as a result." What Snow said was perfectly true, but it caused a precipitous drop in the dollar causing the *Wall Street Journal* to scold Snow for "dumping on his own currency."

Did Snow intend or not intend for the dollar to drop?

[2]Stalnaker uses *credible* as a technical term. The message "I did not kill Johnny Boz" would be pf-rational (*prima facie* rational) as Tramell would prefer it to be believed, but not *credible*, for she would send the same message even if she *had* killed Johnny Boz.

Such issues are too deep for the present paper and we will postpone detailed discussion to a later publication.

5 Conclusions

We have looked in this paper at various ways in which we could follow a wider approach to knowledge. The pattern which Plaza and others (Baltag and Moss 2004, van Benthem et al. 2006) have followed is certainly an important one. It consists of hearing a formula from a trusted source.

But there are many many other ways of acquiring knowledge and knowledge plays a very wide role both in personal and social life. I encourage logicians to follow this path.

Acknowledgements This research was partly supported by a grant from the PSC-CUNY Faculty research assistance program. Thanks to Aranazazu San Gines and Çağıl Taşdemir for comments.

References

A. Baltag and L. Moss. Logics for epistemic programs. *Synthese*, 139:165–224, 2004.

J. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

E. Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.

C. Peirce. The fixation of belief. In C. Hartshorne and P. Weiss, editors, *Collected Papers of Charles Sanders Perice*. Harvard University Press, 1931-5.

F. P. Ramsey. Truth and probability. In *The Foundations of Mathematics and Other Logical Essays*. Routledge, 1954.

L. Savage. *The Foundations of Statistics*. Dover, second edition, 1972.

R. Stalnaker. Saying and meaning, cheap talk and credibility. In J. Benz and R. van Rooij, editors, *Game Theory and Pragmatics*. Palgrave MacMillan, 2005.

J. Stanley. *Knowledge and Practical Interests*. Oxford University Press, 2005.

Critical Comparisons between the Nash Noncooperative Theory and Rationalizability

Tai-Wei Hu and Mamoru Kaneko

Northwestern University, Waseda University
t-hu@kellogg.northwestern.edu, mkanekoepi@waseda.jp

Abstract

The theories of Nash noncooperative solutions and of rationalizability intend to describe the same target problem of *ex ante* individual decision making, but they are distinctively different. We consider what their essential difference is by giving a unified approach and parallel derivations of their resulting outcomes. Our results show that the only difference lies in the use of quantifiers for each player's predictions about the other's possible decisions; the universal quantifier for the former and the existential quantifier for the latter. Based on this unified approach, we discuss the statuses of those theories from the three points of views: Johansen's postulates, the free-will postulate vs. complete determinism, and prediction/decision criteria. One conclusion we reach is that the Nash theory is coherent with the free-will postulate, but we would meet various difficulties with the rationalizability theory.

1 Introduction

We make critical comparisons between the theory of *Nash noncooperative solutions* due to Nash (1951) and the theory of *rationalizable strategies* due to Bernheim (1984) and Pearce (1984). Each theory is intended to be a theory of *ex ante* individual decision making in a game, and thus focuses on the decision-making process before the actual play of the game. The difference in their resulting outcomes has been well analyzed and known. However, their conceptual difference has not been much discussed. In

this paper, we evaluate these two theories while considering certain conceptual bases of game theory and addressing the question of logical coherence of these theories with them.

We begin with a brief review of these theories. It is well known that Nash (1951) provides the concept of Nash equilibrium and proves its existence in mixed strategies. However, it is less known that the main focus of (Nash 1951) is on *ex ante* individual decision making. In that paper, various other concepts are developed, including interchangeability, solvability, subsolutions, symmetry, and values; those concepts are ingredients of a theory of *ex ante* individual decision making, though the aim is not explicitly stated in (Nash 1951). This view is discussed in Nash's (1950) dissertation (p. 23) and a few other papers such as Johansen (1982) and Kaneko (1999).[1] We call the entire argumentation the *Nash noncooperative theory*.[2]

On the other hand, in the literature, the theory of rationalizability is typically regarded as a faithful description of *ex ante* individual decision making in games, expressing the common knowledge of "rationality". Mas-Colell *et al*. (1995, p. 243) wrote: "*The set of rationalizable strategies consists precisely of those strategies that may be played in a game where the structure of the game and the player's rationality are common knowledge among the players.*" This view is common in many standard game theory/micro-economics textbooks.

The literature exhibits a puzzling feature: Both theories target *ex ante* individual decision making, and both are widely used by many researchers. However, their formal definitions, predicted outcomes, and explanations differ considerably. This puzzling feature raises the following questions: How should we make comparisons between these theories? Then, what are their main differences? How would the difference be evaluated? What are bases for such an evaluation? This paper attempts to answer these questions.

We formulate the two theories in terms of prediction/decision criteria, which gives a unified framework for comparisons of these theories. For the Nash theory, the criterion is given by the following requirements:

N1^o: player 1 chooses his best strategy against all of his predictions about player 2's choice based on N2^o;

N2^o: player 2 chooses his best strategy against all of his predictions about player 1's choice based on N1^o.

[1] Millham (1974) and Jansen (1981) study the mathematical structure of the solution and subsolutions, but do not touch the view.

[2] The mathematical definition of Nash equilibrium allows different interpretations such as a steady state in a repeated situation (one variant is the "mass-action" interpretation due to Nash 1950, pp. 21-22), but we do not touch other interpretations (see Kaneko 2004).

We may say that player 1 makes a decision if it satisfies N1^o; however, to determine this decision, N1^o requires a prediction about 2's possible decisions, which are determined by N2^o. The symmetric form N2^o determines a decision for 2 if he predicts 1's decisions. In this sense, these requirements are circular. Also, they can be regarded as a system of simultaneous equations with players' decisions/predictions as unknown. In Section 3, we show that the system N1^o-N2^o characterizes the Nash noncooperative solution as the greatest set satisfying them if the game is solvable (the set of Nash equilibria is interchangeable); and for an unsolvable game, a maximal set satisfying them is a subsolution.

The rationalizable strategies are characterized by another prediction/decision criterion R1^o-R2^o:

R1^o: player 1 chooses his best strategy against some of his predictions about player 2's choice based on R2^o;

R2^o: player 2 chooses his best strategy against some of his predictions about player 1's choice based on R1^o.

These are obtained from N1^o-N2^o simply by replacing the quantifier "for all" by "for some" before predictions about the other player's decisions. These requirements are closely related to the BP-property ("best-response property" in (Bernheim 1984) and (Pearce 1984)), and the characterization result is given in Section 3.

The above prediction/decision criteria and characterization results unify the Nash noncooperative theory and rationalizability theory, and pinpoint their difference: It is the choice of the universal or existential quantifiers for predictions about the other player's possible decisions. To evaluate this difference, we first review the discussion of *ex ante* decision making in games given in (Johansen 1982). In his argument, a theory of *ex ante* decision making in games should describe a player's active inferences based on certain axioms about his own and the other's decision-making. Johansen gives four postulates for the Nash solution, although his argument there is still informal and contains some ambiguities.

Our formulation of N1^o-N2^o may be viewed as an attempt to formalize his postulates in the language of classical game theory. The pinpointed difference between the two theories clarifies the precise requirements in those postulates to obtain the Nash theory. One of Johansen's postulates requires that any possible decision be a best response to the predicted decisions, which is violated by the "for some" requirement in R1^o-R2^o. His postulates help to clarify N1^o-N2^o, and *vice versa*. Nevertheless, his postulates contain some subtle concepts, which go beyond the language of classical game theory.

One such concept is "rationality". In the theory of rationalizability, "rationality" is typically regarded as equivalent to payoff maximization. In Johansen's postulates, how-

ever, payoff maximization is separated from "rationality", and is only one component of "rationality". We also take this broader view of "rationality"; in our formulation, it includes, but not limited to, the prediction/decision criterion and logical abilities to understand their implications. This broader view allows further research on decision criterion (such as the additional principles needed for specific classes of unsolvable games) and investigations of how players' logical abilities affect their decisions.

To evaluate the difference further, we go to deeper methodological assumptions: the *free-will postulate* vs. *complete determinism*. The former, stating that each player has free will, is automatically associated with decision making. The quantifier "for all" in N1^o-N2^o is coherent with the application of the free-will postulate between the players. On the other hand, as will be argued in Section 4, the theory of rationalizability is better understood from the perspective of complete determinism. Indeed, the epistemic justification for rationalizability begins with a complete description of players' actions as well as mental states, and characterizes classes of those states by certain assumptions.

As a result, our problem is a choice between two methodological assumptions, the free-will postulate and complete determinism. This choice is discussed by Morgenstern (1935) and Heyek (1952) in the context of economics and/or social science in general. Based upon their arguments, we will conclude that the free-will postulate is more coherent with large part of social science than complete determinism. From this perspective, the Nash theory is preferable to rationalizability.

The Nash theory might be less preferred in that it does not recommend definite decisions for unsolvable games. However, it may not be a defect from the perspective that it points out that additional principles, other than the decision criteria given above, are needed for decision making in unsolvable games. A general study of such additional principles is beyond the scope of this paper, but we remark that many applied works appeal to principles such as symmetry (which is already discussed in (Nash 1951)) and the Pareto criterion. As an instance, we will give an argument with the Pareto principle for the class of games of strategic complementarity in Section 3.1. From a theoretical perspective, our approach provides a framework to discuss coherence between basic decision criteria and additional principles.

The paper is written as follows: Section 2 introduces the theories of Nash noncooperative solutions and rationalizable strategies; we restrict ourselves to finite 2-person games for simplicity. Section 3 formulates N1^o-N2^o and R1^o-R2^o, and gives two theorems characterizing the Nash noncooperative theory and rationalizability. In Section 4, we discuss implications from them considering foundational issues. Section 5 gives a summary and states continuation to the companion paper.

2 Preliminary definitions

In this paper, we restrict our analysis to finite 2-person games with pure strategies. In Section 3.3, we discuss required changes for our formulation to accommodate mixed strategies.

We begin with basic concepts in a finite 2-person game. Let $G = (N, \{S_i\}_{i\in N}, \{h_i\}_{i\in N})$ be a finite 2-person game, where $N = \{1, 2\}$ is the set of players, S_i is the finite set of pure strategies and $h_i : S_1 \times S_2 \to R$ is the payoff function for player $i \in N$. We assume $S_1 \cap S_2 = \emptyset$. When we take one player $i \in N$, the remaining player is denoted by j. Also, we write $h_i(s_i; s_j)$ for $h_i(s_1, s_2)$. The property that s_i is a *best-response* against s_j, i.e.,

$$h_i(s_i; s_j) \geq h_i(s_i'; s_j) \text{ for all } s_i' \in S_i, \tag{1}$$

is denoted by $\text{Best}(s_i; s_j)$. Since $S_1 \cap S_2 = \emptyset$, the expression $\text{Best}(s_i; s_j)$ has no ambiguity. A pair of strategies (s_1, s_2) is a *Nash equilibrium* in G iff $\text{Best}(s_i; s_j)$ holds for both $i = 1, 2$. We use $E(G)$ to denote the set of all Nash equilibria in G. The set $E(G)$ may be empty.

Nash noncooperative solutions: A subset E of $S_1 \times S_2$ is *interchangeable* iff

$$(s_1, s_2), (s_1', s_2') \in E \text{ imply } (s_1, s_2') \in E. \tag{2}$$

It is known that this requirement is equivalent for E to have the product form, as stated in the following lemma.

Lemma 1. *Let $E \subseteq S_1 \times S_2$ and let $E_i = \{s_i : (s_i; s_j) \in E \text{ for some } s_j \in S_j\}$ for $i = 1, 2$. Then, E satisfies (2) if and only if $E = E_1 \times E_2$.*

Now, let $\mathbf{E} = \{E : E \subseteq E(G)$ and E satisfies (2)$\}$. We say that E is the *Nash solution* iff E is nonempty and is the greatest set in $\mathbf{E}$, i.e., $E' \subseteq E$ for any $E' \in \mathbf{E}$. We say that E is a *Nash subsolution* iff E is a nonempty maximal set in $\mathbf{E}$, i.e., there is no $E' \in \mathbf{E}$ such that $E \subsetneq E'$.

Table 2.1

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$
$\mathbf{s}_{11}$	$(2, 2)$	$(1, 1)$
$\mathbf{s}_{12}$	$(1, 1)$	$(0, 0)$

Table 2.2

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$
$\mathbf{s}_{11}$	$(1, 1)$	$(1, 1)$
$\mathbf{s}_{12}$	$(1, 1)$	$(0, 0)$

When $E(G) \neq \emptyset$, $E(G)$ is the Nash solution if and only if $E(G)$ satisfies (2). When the Nash solution exists for game G, G is called *solvable*. The game of Table 2.1 is solvable. On the other hand, a game G may be unsolvable for two reasons: either

$E(G) = \emptyset$ or $E(G)$ is nonempty but violates (2). For a game G with $E(G) \neq \emptyset$, a subsolution exists always; specifically, for any $(s_1, s_2) \in E(G)$, there is a subsolution E^o containing (s_1, s_2). This E^o may not be unique: The game of Table 2.2 is not solvable and has two subsolutions: $\{(\mathbf{s}_{11}, \mathbf{s}_{21}), (\mathbf{s}_{11}, \mathbf{s}_{22})\}$ and $\{(\mathbf{s}_{11}, \mathbf{s}_{21}), (\mathbf{s}_{12}, \mathbf{s}_{21})\}$, and both include $(\mathbf{s}_{11}, \mathbf{s}_{21})$.

In Section 3, it will be argued that the Nash solution can be regarded as a theory of *ex ante* decision making in games. Here we give two comments about this argument.

First, for a solvable game, the theory recommends the set of possible decisions for each player, i.e., the set of Nash strategies for him; moreover, the recommendation also includes the set of predicted decisions of the other player. This means that from player 1's perspective, $E_1(G) = \{s_1 \in S_1 : (s_1, s_2) \in E(G) \text{ for some } s_2\}$ describes player 1's possible decisions, while $E_2(G) = \{s_2 \in S_2 : (s_1, s_2) \in E(G) \text{ for some } s_1\}$ is player 1's predictions of player 2's decisions. As shown later, predictions about player 2's decisions are crucial to determine player 1's possible decisions from the perspective of *ex ante* decision making in games.

Second, the Nash theory does not provide a definite recommendation for decisions if the game is unsolvable, even if a subsolution exists. Suppose that G has exactly two subsolutions, say, $F^1 = F_1^1 \times F_2^1$ and $F^2 = F_1^2 \times F_2^2$ with $F_i^1 \neq F_i^2$ for $i = 1, 2$. One may think that the Nash theory would recommend the set $E_i = F_i^1 \cup F_i^2$ for player i as the set of possible decisions to play G. However, this is not valid; we cannot find a set E_1' or E_2' such that $E_1' \times (F_2^1 \cup F_2^2)$ or $(F_1^1 \cup F_1^2) \times E_2'$ satisfies interchangeability.

Rationalizable strategies: Now, we turn to rationalizability. The pure strategy version introduced here is known as *point-rationalizability* due to Bernheim (1984). We begin with the iterative definition of rationalizability. A sequence of sets of strategies, $\{(R_1^\nu(G), R_2^\nu(G))\}_{\nu=0}^\infty$, is inductively defined as follows: for $i = 1, 2$, $R_i^0(G) = S_i$, and

$$R_i^\nu(G) = \{s_i : \text{Best}(s_i; s_j) \text{ holds for some } s_j \in R_j^{\nu-1}(G)\} \text{ for any } \nu \geq 1. \tag{3}$$

We obtain rationalizable strategies by taking the intersection of these sets, i.e., $R_i(G) = \bigcap_{\nu=0}^\infty R_i^\nu(G)$ for $i = 1, 2$; a pure strategy $s_i \in S_i$ is *rationalizable* iff $s_i \in R_i(G)$.

It is shown by induction on ν that $R_i^\nu(G)$ is nonempty for all ν and $i = 1, 2$. Also, each sequence $\{R_i^\nu(G)\}_\nu$ is monotonically decreasing. Because each $R_i^\nu(G)$ is finite and nonempty, $R_i^\nu(G)$ becomes constant after some $\overline{\nu}$; as a result, $R_i(G)$ is nonempty. These facts are more or less known, but we give a proof for completeness.

Lemma 2. *$\{R_i^\nu(G)\}_\nu$ is a decreasing sequence of nonempty sets, i.e., $R_i^\nu(G) \supseteq R_i^{\nu+1}(G) \neq \emptyset$ for all ν.*

Proof. We show by induction over ν that the two sequences $\{R_i^\nu(G)\}_\nu$, $i = 1, 2$, are decreasing with respect to the set-inclusion relation. Once this is shown, since S_i is finite, we have $R_i(G) = \bigcap_{\nu=0}^{\infty} R_i^\nu(G) \neq \emptyset$. For the base case of $\nu = 0$, we have $R_i^0(G) = S_i \supseteq R_i^1(G)$ for $i = 1, 2$. Now, suppose the hypothesis that this inclusion holds up to ν and $i = 1, 2$. Let $s_i \in R_i^{\nu+1}(G)$. By (3), $\mathrm{Best}_i(s_i; s_j)$ holds for some $s_j \in R_j^\nu(G)$. Since $R_j^{\nu-1}(G) \supseteq R_j^\nu(G)$ by the supposition, $\mathrm{Best}_i(s_i; s_j)$ holds for some $s_j \in R_j^{\nu-1}(G)$. This means $s_i \in R_i^\nu(G)$. □

Criterion for prediction/decision making: Our discussion of *ex ante* decision making in games begins with a prediction/decision criterion.[3] While comparison between the Nash theory and rationalizability is our concern, some simpler examples of decision criteria may be helpful. First, utility maximization can be regarded as a decision criterion in a non-interactive context, which recommends the set of decisions maximizing a given utility function. In game theory, a classical example of a decision criterion is the *maximin* criterion due to von Neumann-Morgenstern (1944): It recommends a player to choose a strategy maximizing the guarantee level (that is, the minimum payoff for a strategy). In $G = (N, \{S_i\}_{i\in N}, \{h_i\}_{i\in N})$, let E_i be a nonempty subset of $S_i, i = 1, 2$. The set E_i is interpreted as the set of possible decisions for player i based on the maximin criterion. The criterion is formulated as follows:

NM1: for each $s_1 \in E_1$, s_1 maximizes $\min_{s_2\in S_2} h_1(s_1; s_2)$;

NM2: for each $s_2 \in E_2$, s_2 maximizes $\min_{s_1\in S_1} h_2(s_2; s_1)$.

These are not interactive, since NMi, $i = 1, 2$, can recommend a decision without considering NMj, and player i needs to know only his own payoff function. Thus, no prediction is involved for decision making with this criterion.

A more sophisticated criterion may allow one player to consider the other's criterion. One possibility is the following:

N1: for each $s_1 \in E_1$, $\mathrm{Best}(s_1, s_2)$ holds for all $s_2 \in E_2$;

NM2: for each $s_2 \in E_2$, s_2 maximizes $\min_{s_1\in S_1} h_2(s_2; s_1)$.

The criterion N1 requires player 1 to predict player 2's decisions and to choose his best decision against that prediction, while player 2 still adopts the maximin criterion. In this sense, their interpersonal thinking stops at the second level. In the Nash theory and rationalizability theory, we would meet some circularity and their interpersonal thought goes beyond the second level.

[3] A general concept of a prediction/decision criterion is formulated in an epistemic logic of shallow depths in (Kaneko and Suzuki 2002).

There may be multiple pairs of (E_1, E_2) that satisfies a given decision criterion. Without other information than the criterion and components of the game, a player (and we) cannot make a further choice of particular strategies among those satisfying the criterion. In the case of NM1-NM2, E_i should consist of all strategies maximizing $\min_{s_2 \in S_2} h_1(s_1; s_2)$; that is, E_i is the greatest set satisfying NMi. In the case of N1-NM2, this should also be applied to player 1's predictions about 2's choice: E_2 in N1 should be the greatest set satisfying NM2. We will impose this greatest-set requirement for E_i in Section 3; this is not a mere mathematical requirement, but is very basic for the consideration of *ex ante* decision making, as it will be discussed later.

3 Parallel derivations of the Nash noncooperative solutions and rationalizable strategies

In this section we give two parallel decision criteria, and derive the Nash noncooperative solutions and the rationalizable strategies from those criteria. Our characterization results pinpoint the difference between the two theories. This difference is used as the basis for our evaluation of these two theories, which comes in Section 4. We give remarks on the mixed strategy versions of those derivations in Section 3.3.

3.1 The Nash noncooperative solutions

The decision criterion for the Nash solution formalizes the statements N1^o and N2^o in Section 1. This criterion, N1-N2, is formulated as follows: Let E_i be a subset of $S_i, i = 1, 2$, interpreted as the set of possible decisions based on N1-N2,

N1: for each $s_1 \in E_1$, Best$(s_1; s_2)$ holds for all $s_2 \in E_2$;

N2: for each $s_2 \in E_2$, Best$(s_2; s_1)$ holds for all $s_1 \in E_1$.

These describe how each player makes his decisions; when one player's viewpoint is fixed, one of N1-N2 is interpreted as decision making, and the other is interpreted as prediction making. For example, from player 1's perspective, N1 describes his decision making, and N2 describes his prediction making.

Mathematically, N1 and N2 can be regarded as a system of simultaneous equations with unknown E_1 and E_2. First we give a lemma showing that (E_1, E_2) satisfies N1-N2 if and only if it consists only of Nash equilibria.

Lemma 3. *Let E_i be a nonempty subset of S_i for $i = 1, 2$. Then, (E_1, E_2) satisfies N1-N2 if and only if any $(s_1, s_2) \in E_1 \times E_2$ is a Nash equilibrium in G.*

Proof. (Only-If): Let (s_1, s_2) be any strategy pair in $E_1 \times E_2$. By N1, $h_1(s_1, s_2)$ is the largest payoff over $h_1(s_1', s_2)$, $s_1' \in S_1$. By the symmetric argument, $h_2(s_1, s_2)$ is the largest payoff over s_2''s. Thus, (s_1, s_2) is a Nash equilibrium in G.
(If): Let $(s_1, s_2) \in E_1 \times E_2$ be a Nash equilibrium. Since $h_1(s_1, s_2) \geq h_1(s_1', s_2)$ for all $s_1' \in S_1$, we have N1. We have N2 similarly.□

Regarding N1-N2 as a system of simultaneous equations with unknown E_1 and E_2, there may be multiple solutions; indeed, any Nash equilibrium pair as a singleton set is a solution for N1-N2. However, the sets E_1 and E_2 should be based only on the information of the game structure G. This implies that we should look for the pair of greatest sets (E_1, E_2) that satisfies N1-N2.[4]

The following theorem states that N1-N2 is a characterization of the Nash solution theory.

Theorem 1. *[The Nash Noncooperative Solutions]* ***(0)****: G has a Nash equilibrium if and only if there is a nonempty pair* (E_1, E_2) *satisfying N1-N2.*

(1)*: Suppose that G is solvable. Then E is the Nash solution* $E(G)$ *if and only if the greatest pair* (E_1, E_2) *satisfying N1-N2 exists and* $E = E_1 \times E_2$.

(2)*: Suppose that G has a Nash equilibrium but is unsolvable. Then E is a Nash subsolution if and only if* (E_1, E_2) *is a nonempty maximal pair satisfying N1-N2.*

Proof. **(0)**: If (s_1, s_2) is a Nash equilibrium of G, then $E_1 = \{s_1\}$ and $E_2 = \{s_2\}$ satisfy N1-N2. Conversely, if a nonempty pair (E_1, E_2) satisfies N1-N2, then, by Lemma 3, any pair $(s_1, s_2) \in E_1 \times E_2$ is a Nash equilibrium of G.

(1):(If): Let (E_1, E_2) be the greatest pair satisfying N1-N2. It suffices to show $E(G) = E_1 \times E_2$. By Lemma 3, any $(s_1, s_2) \in E_1 \times E_2$ is a Nash equilibrium. Conversely, let $(s_1', s_2') \in E(G)$ and $E_i' = \{s_i'\}$ for $i = 1, 2$. Since this pair (E_1', E_2') satisfies N1-N2, we have $(s_1', s_2') \in E_1' \times E_2' \subseteq E_1 \times E_2$. Hence, $E(G) = E_1 \times E_2$.
(Only-If): Since E is the Nash solution, it satisfies (2). Hence, E is expressed as $E = E_1 \times E_2$ by Lemma 1. Since it consists of Nash equilibria, (E_1, E_2) satisfies N1-N2 by Lemma 3. Since $E(G) = E = E_1 \times E_2$, (E_1, E_2) is the greatest pair having N1-N2.
(2): (If): Let (E_1, E_2) be a maximal pair satisfying N1-N2, i.e., there is no (E_1', E_2') satisfying N1-N2 with $E_1 \times E_2 \subsetneq E_1' \times E_2'$. By Lemma 3, $E_1 \times E_2$ is a set of Nash equilibria. Let E' be a set of Nash equilibria satisfying (2) with $E_1 \times E_2 \subseteq E'$. Then, E' is also expressed as $E_1' \times E_2'$. Since $E_1' \times E_2'$ satisfies N1-N2 by Lemma 3, we have $E_i' \subseteq E_i$ for $i = 1, 2$ by maximality for (E_1, E_2). By the choice of E', we have

[4]If any additional information is available, then we extend N1-N2 to include it and should consider the pair of greatest sets satisfying the new requirements.

$E_1 \times E_2 = E'$. Thus, E is a maximal set satisfying interchangeability(2).
(Only-If): Since E is a subsolution, it satisfies (2). Hence, E is expressed as $E = E_1 \times E_2$. Also, by Lemma 3, (E_1, E_2) satisfies N1-N2. Since $E = E_1 \times E_2$ is a subsolution, (E_1, E_2) is a maximal set satisfying N1-N2.□

When G has a Nash equilibrium but is unsolvable, there are multiple pairs of maximal sets (E_1, E_2) satisfying N1-N2. We do not have those problems in NM1-NM2 in Section 2.3, for which the greatest pair always exists and is nonempty. The reason for this difference may be the interactive nature of N1-N2, which is lacking in NM1-NM2.

For an unsolvable game G with a Nash equilibrium, there is no single definite recommended set of decisions and predictions based on N1-N2, even though the decision criterion and game structure are commonly understood between the players. Each maximal pair (E_1, E_2) satisfying N1-N2 may be a candidate, but it requires further information for the players to choose among them. Thus, N1-N2 alone is not sufficient to provide a definite recommendation for unsolvable games. Theorem 1 gives a demarcation line between the games with a definite recommendation and those without it.

One possible way to reach a recommendation for an unsolvable game is to impose an additional criterion, such as the symmetry requirement in Nash (1951). The game of Table 2.2 is unsolvable, but it has a unique symmetric equilibrium $(\mathbf{s}_{11}, \mathbf{s}_{21})$. Hence, if we add the symmetry criterion, we convert an unsolvable game to a solvable game.

Another possible criterion is the Pareto-criterion. It may work to choose one subsolution for some class of games. For example, it is known that a finite game of strategic complementarity (or super modularity) has a Nash equilibrium in pure strategies, and under some mild condition, that if it has multiple equilibria, they are Pareto-ranked (see Vives 2005 for an extensive survey of this theory and its applications). For those games, when there are multiple equilibria, each equilibrium constitutes a subsolution. However, when we add the Pareto-criterion, the subsolution which Pareto dominates the other subsolutions is chosen. Since a finite game version of this theory is not well known, we give a brief description of this theory in our context.

Assume that the strategy set S_i is linearly ordered so that S_i is expressed as $\{1, ..., \ell_i\}$ for $i = 1, 2$. Here, $S_1 \cap S_2 = \emptyset$ is violated but is recovered by a light change. We say that a game G has the *SC property* iff (1): for $i = 1, 2$, $h_i(s_i; s_j)$ is *concave* with respect to s_i, i.e., for all $s_i = 1, ..., l_i - 2$ and all $s_j \in S_j$

$$h_i(s_i + 1; s_j) - h_i(s_i; s_j) \geq h_i(s_i + 2; s_j) - h_i(s_i + 1; s_j); \tag{4}$$

and (2): $h_i(s_i; s_j)$ is *strategically complement*, i.e., for all $s_1 \in S_1 \backslash \{\ell_1\}$ and $s_2 \in S_2 \backslash \{\ell_2\}$,

$$h_i(s_i + 1; s_j) - h_i(s_i; s_j) \leq h_i(s_i + 1; s_j + 1) - h_i(s_i; s_j + 1). \tag{5}$$

Then, the following are more or less known results, but we give a proof for self-containedness.[5]

Lemma 4. *Let G be a game with the SC property.*
(1): G has a Nash equilibrium in pure strategies.
(2): Suppose u-single peakedness, i.e., for each $i = 1, 2$ *and* $s_j \in S_j$, $h_i(s_i; s_j)$ *has a unique maximum over* S_i*.Then, when G has multiple equilibria, they are linearly ordered with strict Pareto-dominance.*

Proof. **(1)**: We will use Tarski's fixed point theorem: Let $(A, \leq)$ be a complete lattice, i.e., any subset of A has both infimum and supremum with respect to $\leq$. A function $\varphi : A \to A$ is called *increasing* iff $a \leq b$ implies $\varphi(a) \leq \varphi(b)$. Tarski's theorem states that φ is an increasing function on a complete lattice $(A, \leq)$ to itself, then φ has a fixed point. See (Vives 2005, Appendix) and (Cousot and Cousot 1979) for relevant concepts.

We define the partial order $\leq$ over $S_1 \times S_2$ by: $(s_1, s_2) \leq (s'_1, s'_2) \Longleftrightarrow s_i \leq s'_i$ for $i = 1, 2$. Then, $(S_1 \times S_2, \leq)$ is a complete lattice. We also define the *(least) best-response function* $f : S_1 \times S_2 \to S_1 \times S_2$ as follows: for $i = 1, 2$ and $s_j \in S_j$,

$$f_i(s_j) = \min\{t_i : \text{ Best}(t_i; s_j) \text{ holds}\}. \tag{6}$$

Now, $f(s_1, s_2) = (f_1(s_2), f_2(s_1))$ for each $(s_1, s_2) \in S_1 \times S_2$. We show that this f is increasing. Then, f has a fixed point (s^o_1, s^o_2), which is a Nash equilibrium.

Suppose $s_j < s'_j$. Let $f_i(s_j) = t_i$. By (6) and (5), we have $0 < h_i(t_i; s_j) - h_i(t_i - 1; s_j) \leq h_i(t_i; s'_j) - h_i(t_i - 1; s'_j)$. By (4), we have $0 < h_i(t_i; s'_j) - h_i(t_i - 1; s'_j) \leq h_i(k_i; s'_j) - h_i(k_i - 1; s'_j)$ for all $k_i \leq t_i$. Thus, $h_i(t_i; s'_j) \geq h_i(k_i; s'_j)$ for all $k_i \leq t_i$. This implies that player i's best response to s'_j is at least as small as t_i, i.e., $f_i(s'_j) = t'_i \geq t_i$.

(2): Let $(s_1, s_2), (s'_1, s'_2)$ be two Nash equilibria with $s_i < s'_i$. By the monotonicity of f shown in (1), $s_j = f_j(s_i) \leq f_j(s'_i) = s'_j$. If $s_j = s'_j$, then $h_i(\cdot; s_j)$ takes a maximum at s_i and s'_i. This is not allowed by u-single peakedness.□

When an SC game G with u-single peakedness has multiple equilibria, G is unsolvable by (2). However, if we add one criterion for player i's prediction/decision criterion, then we can choose one solution for any SC game with u-single peakedness. It may be better to state the result as a theorem.

[5]Intervals of reals are typically adopted for these results. But Tarski's fixed point theorem is applied for the existence result in our case, too. In fact we can construct an algorithm to find a Nash equilibrium.

Theorem 2. *Let G be an SC game with u-single peakedness. Suppose that (E_1, E_2) and (E_1', E_2') satisfy N1-N2. Then, if for $i = 1, 2$, $h_i(s) \geq h_i(s')$ for some $s \in E_1 \times E_2$ and $s' \in E_1' \times E_2'$, then $E_1 \times E_2$ consists of the unique NE Pareto-dominating all other NE's.*

Thus, one subsolution is chosen by adding the Pareto-criterion to N1-N2.

3.2 Rationalizable strategies

The decision criterion for rationalizability theory, which formalizes the statements R1^o and R2^o in Section 1, is given as follows: for E_1 and E_2,

R1: for each $s_1 \in E_1$, Best$(s_1; s_2)$ holds for some $s_2 \in E_2$;

R2: for each $s_2 \in E_2$, Best$(s_2; s_1)$ holds for some $s_1 \in E_1$.

This criterion differs from N1-N2 only in that the quantifier "for all" before players' predictions in N1-N2 is replaced by "for some". In fact, R1-R2 is the pure-strategy version of the BP-property given by Bernheim (1984) and Pearce (1984). The greatest pair (E_1, E_2) satisfying R1-R2 exists and coincides with the sets of rationalizable strategies $(R_1(G), R_2(G))$. A more general version of the following theorem is reported in Bernheim (1984) (Proposition 3.1); we include the proof for self-containment.

Theorem 3. *$(R_1(G), R_2(G))$ is the greatest pair satisfying R1-R2.*

Proof. Suppose that (E_1, E_2) satisfies R1-R2. First, we show by induction that $E_1 \times E_2 \subseteq R_1^\nu(G) \times R_2^\nu(G)$ for all $\nu \geq 0$, which implies $E_1 \times E_2 \subseteq R_1(G) \times R_2(G)$. Since $R_i^0(G) = S_i$ for $i = 1, 2$, $E_1 \times E_2 \subseteq R_1^0(G) \times R_2^0(G)$. Now, suppose $E_1 \times E_2 \subseteq R_1^\nu(G) \times R_2^\nu(G)$. Let $s_i \in E_i$. Due to the R1-R2, there is an $s_j \in E_j$ such that Best$(s_i; s_j)$ holds. Because $E_j \subseteq R_j^\nu(G)$, we have $s_j \in R_j^\nu(G)$. Thus, $s_i \in R_i^{\nu+1}(G)$.

Conversely, we show that $(E_1(G), E_2(G))$ satisfies R1-R2. Let $s_i \in R_i(G) = \bigcap_{\nu=0}^\infty R_i^\nu(G)$. Then, for each $\nu = 0, 1, 2, ...$, there exists $s_j^\nu \in R_j^\nu$ such that Best$(s_i; s_j^\nu)$ holds. Since S_j is a finite set, we can take a subsequence $\{s_j^{\nu_t}\}_{t=0}^\infty$ in $\{s_j^\nu\}_{\nu=0}^\infty$ such that for some $s_j^* \in S_j$, $s_j^{\nu_t} = s_j^*$ for all ν_t. Then, s_j^* belongs to $R_j(G) = \bigcap_{\nu=0}^\infty R_j^\nu(G)$. Also, Best$_i(s_i; s_j^*)$ holds. Thus, $(R_1(G), R_2(G))$ satisfies R1-R2. □

Existence of a theoretical prediction: Theorem 3 and Lemma 2 imply that the greatest pair satisfying R1-R2 exists and consists of the sets of rationalizable strategies. Interchangeability is automatically satisfied by construction. In this respect, the rationalizability theory appears preferable to the Nash theory in that it avoids the issues due to emptiness or multiplicity of subsolutions. We take a different perspective to reverse this preference: Difficulties involved in the Nash theory identify situations where additional requirements other than N1-N2 are required for prediction/decision making. In

this sense, the Nash theory is a more precise and potentially richer theory of *ex ante* decision making in an interactive situations.

Set-theoretical relationship to the Nash solutions: It follows from Theorem 3 that each strategy of a Nash equilibrium is a rationalizable strategy. Hence, the Nash solution, if it exists, is a subset of the set of rationalizable strategy profiles. However, the converse does not necessarily hold. Indeed, consider the game of Table 3.4, where the subgame determined by the 2*nd* and 3*rd* strategies for both players is the "matching pennies".

Table 3.4

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$	$\mathbf{s}_{23}$
$\mathbf{s}_{11}$	$(5,5)$	$(-2,-2)$	$(-2,-2)$
$\mathbf{s}_{12}$	$(-2,-2)$	$(1,-1)$	$(-1,1)$
$\mathbf{s}_{13}$	$(-2,-2)$	$(-1,1)$	$(1,-1)$

This game has a unique Nash equilibrium, $(\mathbf{s}_{11}, \mathbf{s}_{21})$. Hence, the set consisting of this equilibrium is the Nash solution.

Both $\mathbf{s}_{11}$ and $\mathbf{s}_{21}$ are rationalizable strategies. Moreover, the other four strategies, $\mathbf{s}_{12}, \mathbf{s}_{13}$ and $\mathbf{s}_{22}, \mathbf{s}_{23}$ are also rationalizable: Consider $\mathbf{s}_{12}$. It is a best response to $\mathbf{s}_{22}$, which is a best response to $\mathbf{s}_{13}$, and $\mathbf{s}_{13}$ is a best response to $\mathbf{s}_{23}$, which is a best response to $\mathbf{s}_{12}$. That is, we have the following relations:

$$\text{Best}(\mathbf{s}_{12};\mathbf{s}_{22}),\ \text{Best}(\mathbf{s}_{22};\mathbf{s}_{13}), \text{Best}(\mathbf{s}_{13};\mathbf{s}_{23}),\ \text{and Best}(\mathbf{s}_{23};\mathbf{s}_{12}).$$

By Theorem 3, those four strategies are rationalizable. In sum, all the strategies are rationalizable in this game.

This example shows that even for solvable games, the Nash solution may differ from rationalizable strategies.[6] As we shall see later, the game of Table 3.4 becomes unsolvable if mixed strategies are allowed, while the rationalizable strategies remain the same.

3.3 Mixed strategy versions

Theorems 1 and 3 can be carried out in mixed strategies without much difficulty. The use of mixed strategies may give some merits and demerits to each theory. Here, we give comments on the mixed strategy versions of the two theories.

[6]When a 2-person game has no Nash equilibria, each player has at least two rationalizable strategies. If a player has a unique rationalizable strategy, it is a Nash strategy. Moreover, when each player has a unique rationalizable strategy, then the pair of them is a unique Nash equilibrium. Example 3.4 states that the converse does not necessarily hold.

The mixed strategy versions can be obtained by extending the strategy sets S_1 and S_2 to the mixed strategy sets $\Delta(S_1)$ and $\Delta(S_2)$, where $\Delta(S_i)$ is the set of probability distributions over S_i. The notion of Nash equilibrium is defined in the same manner with the strategy sets $\Delta(S_1)$ and $\Delta(S_2)$: Once the Nash equilibrium is defined, the Nash solution, subsolution, etc. are defined in the same manner. However, the mixed strategy version of rationalizability requires some modification: A sequence of sets of strategies, $\{(\tilde{R}_1^\nu(G), \tilde{R}_2^\nu(G))\}_{\nu=0}^\infty$, is inductively defined as follows: for $i = 1, 2$, $\tilde{R}_i^0(G) = S_i$, and for any $\nu \geq 1$,

$$\tilde{R}_i^\nu(G) = \{s_i : \text{Best}(s_i; m_j) \text{ holds for some } m_j \in \Delta(\tilde{R}_j^{\nu-1}(G))\}.$$

A pure strategy $s_i \in S_i$ is *rationalizable* iff $s_i \in \tilde{R}_i(G) = \bigcap_{\nu=0}^\infty \tilde{R}_i^\nu(G)$.

Requirements N1-N2 are modified by replacing S_i by $\Delta(S_i)$, $i = 1, 2$;for $E_i \subseteq \Delta(S_i)$, $i = 1, 2$,

N1^m: for each $m_1 \in E_1$, Best$(m_1; m_2)$ holds for all $m_2 \in E_2$,

N2^m: for each $m_2 \in E_2$, Best$(m_2; m_1)$ holds for all $m_1 \in E_1$.

Notice that N1^m-N2^m is the same as N1-N2 with different strategy sets. Moreover, Theorem 1 still holds without any substantive changes.

In a parallel manner, the mixed strategy version of rationalizability can also be obtained: for $E_i \subseteq \Delta(S_i)$, $i = 1, 2$,

R1^m: for each $m_1 \in E_1$, Best$(m_1; m_2)$ holds for some $m_2 \in E_2$,

R2^m: for each $m_2 \in E_2$, Best$(m_2; m_1)$ holds for some $m_1 \in E_1$.

This is a direct counterpart of R1-R2 in a game with mixed strategies. In this case, a player is allowed to play mixed strategies. However, in the original version of rationalizability in (Bernheim 1984) and (Pearce 1984), the players are allowed to use pure strategies only; indeed, mixed strategies are interpreted as a player's beliefs about the other player's decisions. We can reformulate R1^m-R2^m based on this interpretation of mixed strategies: In R1^m, the first occurrence of m_1 is replaced by a pure strategy in the support of E_1, and R2^m is modified in a parallel manner. This reformulation turns out to be mathematically equivalent to R1^m-R2^m.

With the replacement of R1-R2 by R1^m-R2^m in Theorem 3.5, the following statement holds:

Theorem 4. $(\Delta(\tilde{R}_1(G)), \Delta(\tilde{R}_2(G)))$ *is the greatest pair satisfying R1^m-R2^m.*

A simple observation is that a rationalizable strategy in the pure strategy version is also a rationalizable strategy in the mixed strategy version. Similarly, since a Nash

equilibrium in pure strategies is also a Nash equilibrium in mixed strategies, it may be conjectured that if a game G has the Nash solution E in the pure strategies, it might be a subset of the Nash solution in mixed strategies. In fact, this conjecture is answered negatively.

Consider the game of Table 3.4. This game has seven Nash equilibria in mixed strategies:

$$((1,0,0),(1,0,0)),\ ((0,\tfrac{1}{2},\tfrac{1}{2}),(0,\tfrac{1}{2},\tfrac{1}{2})),\ ((\tfrac{4}{18},\tfrac{7}{18},\tfrac{7}{18}),(\tfrac{4}{18},\tfrac{7}{18},\tfrac{7}{18}))$$
$$((\tfrac{1}{8},\tfrac{7}{8},0),(\tfrac{3}{10},\tfrac{7}{10},0)),\ ((\tfrac{1}{8},0,\tfrac{7}{8}),(\tfrac{3}{10},0,\tfrac{7}{10})),\ ((\tfrac{3}{10},\tfrac{7}{10},0),(\tfrac{1}{8},0,\tfrac{7}{8})),\ ((\tfrac{3}{10},0,\tfrac{7}{10}),(\tfrac{1}{8},\tfrac{7}{8},0)).$$

This set does not satisfy interchangeability (2). For example, $((1,0,0),(1,0,0))$ and $((0,\frac{1}{2},\frac{1}{2}),(0,\frac{1}{2},\frac{1}{2}))$ are Nash equilibria, but $((0,\frac{1}{2},\frac{1}{2}),(1,0,0))$ is not a Nash equilibrium. Thus, (2) is violated, and the set of all mixed strategy Nash equilibria is not the Nash solution. This result depends upon the choice of payoffs: In Table 3.5, (s_{11}, s_{21}) is a unique Nash equilibrium even in mixed strategies, while all pure strategies are still rationalizable.

Table 3.5

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$	$\mathbf{s}_{23}$
$\mathbf{s}_{11}$	$(5,5)$	$(\frac{1}{2},\frac{1}{2})$	$(\frac{1}{2},\frac{1}{2})$
$\mathbf{s}_{12}$	$(\frac{1}{2},\frac{1}{2})$	$(1,-1)$	$(-1,1)$
$\mathbf{s}_{13}$	$(\frac{1}{2},\frac{1}{2})$	$(-1,1)$	$(1,-1)$

4 Evaluations of N1-N2 and R1-R2

Our unified approach pinpoints the difference between the Nash and rationalizability theories: the choice of quantifier "for all" or "for some" for each player's predictions. Here we evaluate this difference reflecting upon on the conceptual bases of game theory. We take Johansen's (1982) argument on the Nash theory as our starting point. Then, we make comparisons between the two theories by considering two methodological principles: the free-will postulate and complete determinism. We also consider multiplicity in prediction/decision criteria and how we should take it in our research activities.

4.1 Johansen's argument

Johansen (1982) gives the following four postulates for prediction/decision making in games and asserts that the Nash noncooperative solution is derived from those postulates for solvable games. For this, he assumes (p. 435) that the game has the unique

Nash equilibrium, but notes (p. 437) that interchangeability is sufficient for his assertion.

Postulate J1 (Closed World)[7]**:** A player makes his decision $s_i \in S_i$ on the basis of, and only on the basis of information concerning the action possibility sets of two players S_1, S_2 and their payoff functions h_1, h_2.

Postulate J2 (Symmetry in PD criterion): In choosing his own decision, a player assumes that the other is rational in the same way as he himself is rational.

Postulate J3 (Predictability): If any[8] decision is a rational decision to make for an individual player, then this decision can be correctly predicted by the other player.

Postulate J4 (Optimization against "for all" predictions): Being able to predict the actions to be taken by the other player, a player's own decision maximizes his payoff function corresponding to the predicted actions of the other player.

Notice that the term "rational" occurs in J2 and J3, and "payoff maximization" in J4. The term "rational" in Johansen's argumentation is broader than its typical meaning in the game theory literature referring to "payoff maximization." Indeed, He regards these four postulates together as an attempt to define "rationality"; "payoff maximization" is only one component of "rationality". We may further disentangle it using the concept of prediction/decision criterion, which includes J4 as its component, and the concept of logical abilities. Then, the above four postulates will be well understood, which is now discussed.

Postulate J1 is the starting point for his consideration of *ex ante* decision making. Postulate J2 requires the decision criterion be symmetric between the decision maker and the other player in his mind. Postulate J3 requires each player's prediction about the other's decision be correctly made. Postulate J4 corresponds to the payoff maximization requirement. In the following, we first elaborate Postulates J2 and J3, and then use J1-J4 as a reference point for our critical comparisons between N1-N2 and R1-R2.

Postulate J2 implies that from player 1's perspective, the decision criterion has to be symmetric between the two players. In our context, this is interpreted as applied to the choice of prediction/decision criterion. Both N1-N2 and R1-R2 satisfy this symmetric requirement. The combination N1-NM2 discussed in Section 2 violates symmetry, and so does N1-R2, which will be further discussed in Section 4.3.

Postulate J3 is interpreted in the following manner: First, player 1 thinks about the whole situation, taking player 2's criterion as given, and makes inferences from this

[7]The titles of those postulates are given by the present authors.

[8]This "any" was "some" in Johansen's orginal Posutlate 3. According to logic, this should be "any". However, this is expressed as "some" by many scientists (even mathematicians).

thinking. Based on such inferences, player 1 makes a prediction about 2's decisions. This prediction is *correct* in the sense that player 1's prediction criterion is the same as 2' decision criterion and 1 has the same logical ability as player 2's. In this sense, predictability in J3 is a result of a player's contemplation of the whole interactive situation.[9] In this reasoning, "rationality" in J3 emphasizes symmetry in players' interpersonal logical abilities, while that in J2 emphasizes symmetry in his prediction/decision criterion.

Postulates J1-J3 are compatible with N1-N2 and R1-R2. Only Postulate J4 makes a distinction between the Nash theory and rationalizability theory. If we read Postulate J4 in light of his assertion that interchangeability is a sufficient condition for J1-J4 to lead to the Nash solution, we can interpret J4 as adopting "for all" predicted actions of the other player's possible decisions.

Johansen (1982) does not give a formal analysis of his postulates. Our N1-N2 may be regarded as a formulation of these postulates in the language of classical game theory. In this sense, Theorem 1 formalizes Johansen's assertion that the Nash solution is characterized by J1-J4. If we modify Postulate J4 so that the "for all" requirement is replaced by the "for some" requirement, Theorem 3 for R1-R2 would be a result. We still need to discuss what are bases for the choice of "for all" or "for some".

4.2 The free-will postulate vs. complete determinism

Here, we evaluate the difference between N1-N2 and R1-R2, based on two conflicting meta-theoretical principles: the free-will postulate and complete determinism.

The free-will postulate: This states that players have freedom to make choices following their own will. Whenever the social science involves value judgements for individual beings and/or the society, they rely on the free-will postulate as a foundation.[10] In a single person decision problem, utility maximization may effectively void this postulate.[11] However, in an interactive situation, even if both players are very smart, it is still possible that individual decision making, based on utility maximization alone, may not result in a unique decision. This is first argued in Morgenstern (1935), using the paradox of Moriarty chasing Holmes. This is still a central problem in game theory;

[9] Bernheim's (1986, p. 486) interpretation of J3 in his criticism against these postulates is quite different from our reasoning. In his framework, predictability simply means that the belief about the other player's action, which is exogenously given, coincides with the actual action.

[10] The free-will postulate is needed for deontic concepts such as responsibility for individual choice and also for individual and social efforts for future developments.

[11] This does not imply that utility maximization even for 1-person problem violates the free-will postualte; he has still freedom to ignore his utility.

the free-will postulate constitutes an important part of this problem. In this respect, the free-will postulate still remains relevant to game theory.

Consider applications of the postulate at two different layers in terms of interpersonal thinking:

(*i*): It is applied by the outside observer to the (inside) players;

(*ii*): It is applied by an inside player to the other player.

In application (*i*), the outside theorist respects the free will of each player; the theorist can make no further refinement than the inside player. This corresponds to the greatestness requirement for (E_1, E_2) in Theorems 1.(1) and Theorem 3. In (*ii*), when one player has multiple predictions about the other's decisions, the free-will postulate, applied to interpersonal decision making, requires the player take all possible predictions into account. N1-N2 is consistent with this requirement in that it requires each player's decision be optimal against all predictions.[12]

Criterion R1-R2 involves some subtlety in judging whether it is consistent with application (*ii*). The main difficulty is related to the interpretation of "for some" before the prediction about the other's decision. This leads us to another view, "complete determinism."

Complete determinism: The quantifier "for some" in R1-R2 has two different interpretations:

(*a*): it requires only the mere existence of a rationalizing strategy;

(*b*): it suggests a specific rationalizing strategy predetermined for some other reason.

Interpretation (*a*) is more faithful to the mathematical formulation of R1-R2 as a decision criterion. If we accept (*a*), then arbitrariness of the rationalizing strategy shows no respect to the other player's free will, but we would not find a serious difficulty in R1-R2 with the free-will postulate in that R1-R2 is a prediction/decision criterion adopted by a player. However, this reminds us Aesops' *sour grapes* that the fox finds one convenient reason to persuade himself: For R1-R2, it suffices to find any rationalizing strategy. This interpretation of "rationalization" is at odds with the purpose of a theory of *ex ante* decision-making for games, since such a theory is supposed to provide a rationale for players' decisions as well as predictions. Interpretation (*a*) requires no rationale for each specific rationalizing strategy.

Interpretation (*b*) resolves the arbitrariness in (*a*): According to (*b*), there are some further components, not explicitly included in the game description G and R1-R2, that determine a specific rationalizing strategy. However, a specific rationalizing strategy

[12]There are many other criteria consistent with the requirement. For example, player 1 uses the maximin criterion to choose his action against E_2. Another possibility is to put equal probability on each action in E_2 and to apply expected utility maximization.

for each step has to be uniquely determined, for otherwise the player would have to arbitrarily choose among different strategies or to look for a further reason to choose some of them. Thus, interpretation (*b*) violates Johansen's postulate J1.

Interpretation (*b*) deserves a further analysis, since it is related to complete determinism, which has been regarded as very foundational in natural sciences. To determine a specific rationalizing strategy, one possibility is to refer to a full description of the world including players' mental states; this presumes some form of determinism. We consider only complete determinism for simplicity. Such a full description in a situation with two persons may require an infinite hierarchy of beliefs. Indeed, there is a literature, beginning from Aumann (1987)[13], to justify the rationalizability theory or alike along this line (see Tan-Werlang 1988).

Complete determinism is incompatible with the free-will postulate in that it contains no room for decision; *ex ante* decision making is an empty concept from this perspective. From this view, R1-R2 is regarded as a partial description of a law of causation.

Except for conflicting against the free-will postulate, complete determinism may not be very fruitful as a methodology for social science in general, which is aptly described by Hayek (1952, Section 8.93): *"Even though we may know the general principle by which all human action is causally determined by physical processes, this would not mean that to us a particular human action can ever been recognizable as the necessary result of a particular set of physical circumstances."*

Complete determinism is justified only because of its non-refutability by withdrawing from concrete problems into its own abstract world. In fact, neither complete determinism nor the free-will postulate can be justified by its own basis. Either should be evaluated with coherency of the entire scope and the scientific and/or theoretical discourse.

Our conclusion is that the free-will postulate is needed for the perspective of social sciences, and complete determinism has no such a status in social sciences. The Nash noncooperative theory is constructed coherently with the free-will postulate, but the rationalizability theory meets a great difficulty to reconcile with it.

[13] In the problem of common knowledge in the information partition model due to Robert Aumann, the information partitions themselves are assumed to be common knowledge. He wrote in (Aumann 1976, p. 1237): "*Included in the full description of a state ω of the world is the manner in which information is imparted to the two persons*". This can be interpreted as meaning that the primitive state ω includes every information. A person receives some partial information about ω, but behind this, everything is predetermined. This view is shared with Harsanyi (1967/8) and Aumann (1987).

4.3 Prediction/decision criteria

The characterizations of the Nash and rationalizability theories in terms of prediction/decision criteria are helpful to find their differences as well as to understand Johansen's argument and *vice versa*. However, these characterization results also introduce a new problem: Among all possible prediction/decision criteria, why should we focus particularly on the Nash theory or the rationalizability theory? Here we consider a few exmaples of prediction/decision criteria and their resulting outcomes.

Relativistic view: It may be the case that people adopt different prediction/decision criteria. In addition to N1-N2 and R1-R2, as already indicated, NM1-NM2, N1-NM2 and N1-R2 are also possible candidates, among others. Even restricting our focus to N1-N2 and R1-R2, it is natural to ask why we avoid a mixture, such as N1-R2, of those criteria. Moreover, this combination actually generates a different outcome either from N1-N2 or R1-R2. Consider the game which is obtained from Table 3.4 by changing the payoffs in the first row and first column.

Table 4.1

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$	$\mathbf{s}_{23}$
$\mathbf{s}_{11}$	$(1,1)^{NE}$	$(1,1)^{NE}$	(0,0)
$\mathbf{s}_{12}$	(1,0)	$(1,-1)$	$(-1,1)$
$\mathbf{s}_{13}$	(0,0)	$(-1,1)$	$(1,-1)$

For this game, we can calculate the greatest pairs (E_1, E_2) satisfying N1-N2, R1-R2 and N1-R2 as follows:

N1-N2:	$(\{\mathbf{s}_{11}\}, \{\mathbf{s}_{21}, \mathbf{s}_{22}\})$
R1-R2:	$(\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}, \{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\})$
N1-R2:	$(\{\mathbf{s}_{11}, \mathbf{s}_{12}\}, \{\mathbf{s}_{21}, \mathbf{s}_{22}\})$

As soon as we start considering different combinations, they could provide actually different recommendations.

This relativistic view may turn our target problem into an empirical study of such criteria in real societies. However, a prediction/decision criterion itself is still an analytic concept that serves as a benchmark to understand the prediction/decision-making process in interactive situations. From this perspective, the focus should rather be a study of the underlying structures and rationales for those criteria; if a criterion is incoherent with other bases, people will eventually avoid it. The goal of such study is then to separate some criteria from others, even if we take the relativistic view that people

follow diverse ways of prediction/decision making. For example, Johansen's postulate J2 accepts N1-N2 and R1-R2 but rejects N1-R2 as a legitimate criterion.

This paper analyzes two specific criteria, N1-N2 and R1-R2, taking Johansen's postulates and the current game theory literature as given. However, if we enter the relativistic world of prediction/decision criterion, we may require rationales for the postulates such as J2. A full analysis, which would involve broader conceptual bases for prediction/decision criterion and more explicit study of the underlying thought processes for prediction/decision making, is way beyond the current research. Nevertheless, Section 5 mentions a further research possibility on these problems as a continuation of the present paper.

5 Conclusions

5.1 The unified framework and parallel derivations

We presented the unified framework and parallel derivations of the Nash noncooperative solutions and rationalizable strategies. The difference between them is pinpointed to be the choice of the quantifier "for all" or "for some" for predictions about the other player's possible decisions. In Section 4, we discussed various conceptual issues by viewing the quantifier "for all" and "for some" from the perspectives of Johansen's postulates, the free-will postulate vs. complete determinism, and prediction/decision criteria.

Comparisons with Johansen's postulates help us well understand our unified framework and derivations. The argument from the perspective of the free-will postulate vs. complete determinism concludes that the Nash theory is more coherent to social sciences as a whole than the rationalizability theory. Nevertheless, as a descriptive concept, it would be possible for some people to use a criterion with "for some" for their decision making. Reflections upon our approach in terms of prediction/decision criteria manifest that vast aspects of prediction/decision making in social context are still hidden.

One such problem is the treatment of the assumption of common knowledge. We started this paper with the quotation of Mas-Colell and Green (1995) about the standard interpretation of rationalizability theory in terms of common knowledge. It is also common to interpret the Nash theory as to require common knowledge of the game structure. In this paper, the notion of common knowledge or even knowledge/beliefs remains interpretational. To study the thought process for prediction/decision making explicitly, we meet new issues and additional framework is necessary. The following section discusses these issues.

5.2 Thought process for prediction/decision making

The present paper employs the standard game theory language. In this language, many essential elements remain informal and hidden, including a player's beliefs or knowledge. Those elements are essential for understanding the thought process for prediction/decision-making. Here we consider only N1-N2, but the argument is also applicable to R1-R2.

Prediction making (putting oneself in the other's shoes): N1-N2 is understood as describing both prediction making and decision making: From player 1's perspective, E_1 in N1 is his decision variable, while E_2 in N1 is his prediction variable. Here, player 1 puts himself into player 2's shoes to make predictions. In fact, this argument could not stop here; by putting himself in 2's shoes, 1 needs to think about 2's predictions about 1's decisions. Continuing this argument *ad infinitum,* we meet the infinite regress described in Diagram 5.1, which is made from the viewpoint of player 1. A symmetric argument from player 2's viewpoint can be constructed.

Double uses of N1-N2: In the infinite regress, N1 is a decision criterion for 1 and is a prediction criterion for 2, while N2 is a decision criterion for 2 and a prediction criterion for 1. Thus, N1 and N2 are used both as decision and prediction criteria. This double use makes the infinite regress in Diagram 5.1 collapse into a system of simultaneous equations described by Diagram 5.2. Theorem 1 solves this system of equations.

Diagram 5.1

N1		N1		N1		$\cdots$
$\downarrow$	$\nearrow$	$\downarrow$	$\nearrow$	$\downarrow$	$\nearrow$	
N2		N2		N2		

$\Longrightarrow$

Diagram 5.2

N1
$\downarrow \uparrow$
N2

The language of classical game theory is incapable to explicitly distinguish between player 1's and 2's perspectives; as a result, many foundational problems can only be discussed at interpretational levels. One way to formalize those issues is to reformulate the above problem in the epistemic logic framework. Then, we can avoid the collapses from Diagram 5.1 into Diagram 5.2, and explicitly discuss the relationship between the above infinite regress and the common knowledge of N1-N2. In doing so, we will be able to evaluate the standard interpretations, such as the quotation from (Mas-Colell and Green 1995) in Section 1, of the rationalizability theory as well as the Nash theory. Also, we can more explicitly discuss Johansen's (1982) argument. The research on these problems will be undertaken in the companion paper (Hu and Kaneko 2013).

Acknowledgements The authors are partially supported by Grant-in-Aids for Scientific Research No.21243016 and No.2312002, Ministry of Education, Science and Culture. The authors thank Geir B. Ascheim and Jean-Jacques Herings for valuable comments and discussions.

References

M. W. A. Mas-Colell and J. Green. *Microeconomic Theory*. Oxford University Press, 1995.

R. Aumann. Agreeing to disagree. *Annals of Statistics*, 4:1236–1239, 1976.

R. Aumann. Correlated equilibrium as an expression of bayesian rationality? *Econometrica*, 55:1–18, 1987.

D. Bernheim. Rationalizable strategic behavior. *Econometrica*, 52:1007–1028, 1984.

D. Bernheim. Axiomatic characterizations of rational choice in strategic environments. *Scandinavian Journal of Economics*, 88:473–488, 1986.

P. Cousot and T. Cousot. Constructive versions of tarski's fixed point theorem. *Pacific Journal of Mathematics*, 82:43–57, 1979.

J. Harsanyi. Games with incomplete information played by bayesian agents i-iii. *Management Science*, 14:159–182, 1967/8.

F. Hayek. *The Sensory Order: An Inquiry into the Foundations of Theoretical Psychology*. University of Chicago Press, 1952.

T. Hu and M. Kaneko. Infinite regress arising from prediction/decision making in games. http://www.waseda-pse.jp/ircpea/jp/publish/working-paper/, 2013.

M. Jansen. Maximal nash subsets for bimatrix games. *Naval Research Logistics Quarterly*, 28:147–152, 1981.

L. Johansen. On the status of the nash type of noncooperative equilibrium in economic theory. *Scandinavian Journal of Economics*, 84:421–441, 1982.

M. Kaneko. Epistemic consideration of decision making in games. *Mathematical Social Sciences*, 38:105–137, 1999.

M. Kaneko. *Game Theory and Mutual Misunderstanding*. Springer, 2004.

M. Kaneko and N. Suzuki. Bounded interpersonal inferences and decision making. *Economic Theory*, 19:63–103, 2002.

C. Millham. On nash subsets of bimatrix games. *Naval Research Logistics Quarterly*, 21:307–317, 1974.

O. Morgenstern. Perfect foresight and economic equilibrium. *Zeitschrift für Nationalökonomie*, 6:337–357, 1935.

J. Nash. *Non-cooperative Games*. PhD thesis, Princeton University, 1950.

J. Nash. Non-cooperative games. *Annals of Mathematics*, 54:286–295, 1951.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

D. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52:1029–1050, 1984.

T. Tan and S. Werlang. The bayesian foundations of solution concepts of games. *Journal of Economic Theory*, 45:370–391, 1988.

X. Vives. Complementarities and games: New developments. *Journal of Economic Literature*, 43:437–479, 2005.

A Logic-Based Approach to Pluralistic Ignorance

Jens Ulrik Hansen

Department of Philosophy, Lund University, Sweden
Jens_Ulrik.Hansen@fil.lu.se

Abstract

"Pluralistic ignorance" is a phenomenon mainly studied in social psychology. Viewed as an epistemic phenomenon, one way to define it is as a situation where "*no one believes, but everyone believes that everyone else believes*". In this paper various versions of pluralistic ignorance are formalized using epistemic/doxastic logic (based on plausibility models). The motive is twofold. Firstly, the formalizations are used to show that the various versions of pluralistic ignorance are all consistent, thus there is nothing in the phenomenon that necessarily goes against logic. Secondly, pluralistic ignorance, is on many occasions, assumed to be fragile. In this paper, however, it is shown that pluralistic ignorance need not be fragile to announcements of the agents' beliefs. Hence, to dissolve pluralistic ignorance in general, something more than announcements of the subjective views of the agents is needed. Finally, suggestions to further research are outlined.

Pluralistic ignorance is a term from the social and behavioral sciences going back to the work of Floyd H. Allport and Daniel Katz (1931).[1] Krech and Crutchfield (1948, pp. 388-89) define pluralistic ignorance as a situation where "*no one believes, but everyone believes that everyone else believes*". Elaborated, pluralistic ignorance is the phenomenon where a group of people shares a false belief about the beliefs, norms, actions or thoughts of the other group members. It is a social phenomenon where people make systematic errors in judging other people's private attitudes. This makes it an important notion in understanding the social life. However, pluralistic ignorance is a term used to describe many different phenomena that all share some common features.

[1] See (O'Gorman 1986) for more on the coining of the term "pluralistic ignorance".

Therefore, there are many different definitions and examples of pluralistic ignorance and a few of the most common of these will be presented in Section 1.

Pluralistic ignorance has been approached by formal methods before (Centola et al. 2005, Hendricks 2010), but to the knowledge of the author, (Hendricks 2010) is the only paper that takes a logic-based approach.[2] Hendricks (2010) models pluralistic ignorance using formal learning theory and logic. In this paper, the tool will be classical modal logic in the form of doxastic/epistemic logic. In Section 2, we introduce a doxastic/epistemic logic based on the plausibility models previously studied by van Benthem (2007) and Baltag and Smets (2008). The reason for choosing this framework instead of, for instance, the multi-modal logic **KD45**, is that **KD45** cannot straightforwardly be combined with public announcements.[3] Since one of the aspect of pluralistic ignorance studied in this paper is the question of what it takes to dissolve the phenomenon, we need to be able to talk about the dynamics of knowledge and beliefs. Public announcements are the simplest form of actions that can affect the beliefs and knowledge of the agents and they therefore serve the purpose of this paper perfectly.

After having presented the formal framework in Section 2, it is possible in Section 3 to give a formal analysis of the different versions of pluralistic ignorance. We will give several different formalizations of pluralistic ignorance and discuss whether they are satisfiable or not. Afterwards, we will look at what it takes to dissolve pluralistic ignorance and show that, in general, something more than mere announcements of agents' true beliefs is needed. Since the logical approach to pluralistic ignorance is still very limited, there is ample opportunity for further research and several suggestions will be discussed in Section 4. Following this, a concise conclusion is given in Section 5. Finally, a postscript is added at the very end summing up the research on logic-based models of pluralistic ignorance that have appeared since the first publishing of this paper.

1 Examples of pluralistic ignorance

Examples of pluralistic ignorance are plentiful in the social and behavioral sciences' literature. One example is the drinking of alcohol on (American) college campuses. Several studies have shown that many students feel much less comfortable with drink-

[2] Since the first appearance of this paper in *Future Directions for Logic - Proceedings of PhDs in Logic III*, College Publications, 2012, several other papers with a logic-based approach to pluralistic ignorance have appeared. In Section 6, these papers will be discussed in more details.

[3] Public announcement of a formula φ corresponds, in the model theory of modal logic, to the operation of going to the submodel only containing worlds where φ was true. However, the class of frames underlying the logic **KD45** is not closed under taking submodels, since seriality is not preserved when going to submodels. When combined with public announcement the logic **KD45** actually turns into the logic **S5**.

ing than they believe the average college student does (Prentice and Miller 1993). In other words, the students do not believe that drinking is at all enjoyable, but they still believe that all of their fellow students believe drinking to be quite enjoyable. Another classical example is the classroom example in which, after having presented the students with difficult material, the teacher asks them whether they have any questions. Even though most students do not understand the material they may not ask any questions. All the students interpret the lack of questions from the other students as a sign that they understood the material, and to avoid being publicly displayed as the stupid one, they dare not ask questions themselves. In this case the students are ignorant with respect to some facts, but believe that the rest of the students are not ignorant about the facts.

A classical made-up example is from Hans Christian Andersen's fable "The Emperor's New Clothes" from 1837. Here, two impostors sell imaginary clothes to an emperor claiming that those who cannot see the clothes are either not fit for their office or just truly stupid. Not wanting to appear unfit for his office or truly stupid, the Emperor (as well as everyone else) pretends to be able to see the garment. No one *personally* believes the Emperor to have any clothes on. They do, however, believe that everyone else believes the Emperor to be clothed. Or alternatively, everyone is ignorant to whether the Emperor has clothes on or not, but believes that everyone else is not ignorant. Finally, a little boy cries out: "but he has nothing on at all!" and the pluralistic ignorance is dissolved.

What might be clear from these examples is that pluralistic ignorance comes in many versions. A logical analysis of pluralistic ignorance may help categorize and distinguish several of these different versions. Note that these examples were all formulated in terms of beliefs, but pluralistic ignorance is often defined in the term of norms as well. For instance, Centola et al. (2005) define pluralistic ignorance as "a situation where a majority of group members privately reject a norm, but assume (incorrectly) that most others accept it".

Misperceiving other people's norms or beliefs can occur without it being a case of pluralistic ignorance. Pluralistic ignorance is the case of systematic errors in norm/belief estimation of others. Thus, pluralistic ignorance is a genuine social phenomenon and not just people holding wrong beliefs about other people's norms or beliefs (O'Gorman 1986). This might be the reason why pluralistic ignorance is often portrayed as a fragile phenomenon. Just one public announcement of a private belief or norm will resolve the case of pluralistic ignorance. In "The Emperor's New Clothes" a little boy's outcry is enough to dissolve the pluralistic ignorance. If, in the classroom example, one student dares to ask a question (and thus announces his academic ignorance) the other students will surely follow with questions of their own. In some versions of pluralistic ignorance, the mere awareness of the possibility of pluralistic

ignorance is enough to suspend it. This fragility might not always be the case and, as we shall see, there is nothing in the standard definitions of pluralistic ignorance that forces it to be a fragile phenomenon.

2 Plausibility models: A logical model of belief, knowledge, doubt, and ignorance

We will model knowledge and beliefs using modal logic. More specifically, we will be using the framework of Baltag and Smets (2008). This section is a review of that framework. We will work in a multi-agent setting and thus, assume a finite set of agents $\mathbb{A}$ to be given. Furthermore, we also assume a set of propositional variables PROP to be given. The models of the logic will be special kinds of Kripke models called plausibility models:

Definition 2.1. A plausibility model is a tuple $\mathcal{M} = \langle W, (\leq_a)_{a\in\mathbb{A}}, V\rangle$, where W is a non-empty set of possible worlds/states, $\leq_a$ is a locally connected converse well-founded preorder on W for each $a \in \mathbb{A}$, and V is a valuation that to each $p \in \mathsf{PROP}$ assigns a subset of W.

A relation is a *locally connected converse well-founded preorder* on W if it is locally connected (wherever x and y are related and y and z are related, then x and z are also related), converse well-founded (every non-empty subset of W has a maximal element), and is a preorder (it is reflexive and transitive). In the following we will sometimes refer to the plausibility models simply as models.

The intuition behind plausibility models is that the possible worlds represent different ways the world might be. That $w \leq_a v$, for an agent a, means that agent a thinks that the world v is at least as possible as world w, but a cannot distinguish which of the two is the case. The relation $\leq_a$ will be used to define what an agent a believes. To define what an agent a knows we introduce an equivalence relation $\sim_a$ defined by:

$$w \sim_a v \quad \text{if, and only if} \quad w \leq_a v \text{ or } v \leq_a w$$

The intuition behind $w \sim_a v$ is that for all that agent a knows, she cannot distinguish between which of the worlds w and v is the case. Given an agent a and a world w, the set $|w|_a = \{v \in W \mid v \sim_a w\}$ is the information cell at w of agent a and represents all the worlds that agent a considers possible at the world w. In other words, this set encodes the hard information of agent a at the world w.

Based on the introduced notions, we can now define knowledge and beliefs. Let K_a and B_a be modal operators for all agents $a \in \mathbb{A}$. We read $K_a\varphi$ as "agent a knows that

φ" and $B_a\varphi$ as "agent a believes that φ". We specify the formal language $\mathcal{L}$, which we will be working with, by the following syntax:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid B_a\varphi,$$

where $p \in \mathsf{PROP}$ and $a \in \mathbb{A}$. The logical symbols $\top, \bot, \vee, \rightarrow, \leftrightarrow$ are defined in the usual way. The semantics of the logic is then defined by:

Definition 2.2. Given a plausibility model $\mathcal{M} = \langle W, (\leq_a)_{a\in\mathbb{A}}, V\rangle$ and a world $w \in W$ we define the semantics inductively by:

$$\begin{array}{lll}
\mathcal{M}, w \models p & \text{iff} & w \in V(p) \\
\mathcal{M}, w \models \neg\varphi & \text{iff} & \text{it is not the case that } \mathcal{M}, w \models \varphi \\
\mathcal{M}, w \models \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \models \varphi \text{ and } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models K_a\varphi & \text{iff} & \text{for all } v \in |w|_a,\ \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models B_a\varphi & \text{iff} & \text{for all } v \in max_{\leq_a}(|w|_a),\ \mathcal{M}, v \models \varphi,
\end{array}$$

where $max_{\leq_a}(S)$ is the set of maximal elements of S with respect to the relation $\leq_a$. We say that a formula φ is satisfiable if there is a model $\mathcal{M}$ and a world w in $\mathcal{M}$ such that $\mathcal{M}, w \models \varphi$. A formula φ is valid if for all models $\mathcal{M}$ and all worlds w in $\mathcal{M}$, $\mathcal{M}, w \models \varphi$.

Note, that the semantics makes $K_a\varphi \rightarrow B_a\varphi$ valid. In the framework of Baltag and Smets (2008) other notions of beliefs are also introduced. The first is *conditional beliefs*; $B_a^\varphi\psi$ expresses that agent a believes that ψ was the case, if she learned that φ was the case. The semantic of this modality is:

$$\mathcal{M}, w \models B_a^\varphi\psi \quad \text{iff} \quad \text{for all } v \in max_{\leq_a}(|w|_a \cap [\![\varphi]\!]_\mathcal{M}),\ \mathcal{M}, v \models \psi,$$

where $[\![\varphi]\!]_\mathcal{M}$ is the set of worlds in $\mathcal{M}$ where φ is true. Another notion of belief is *safe belief* for which we use $\Box_a$. The semantics of this modality is:

$$\mathcal{M}, w \models \Box_a\varphi \quad \text{iff} \quad \text{for all } v \in W, \text{ if } w \leq_a v, \text{ then } \mathcal{M}, v \models \varphi.$$

Note, this is the usual modality defined from the relation $\leq_a$. Since $\leq_a$ is reflexive, $\Box_a\varphi \rightarrow \varphi$ is valid and thus, "$\Box_a$-beleifs" are veridical. Hence, safe belief is a very strong notion of belief (or a weak notion of knowledge). Because a central aspect of pluralistic ignorance is people holding *wrong* beliefs, safe belief is not a suitable notion. Yet another notion of belief, that also implies truth, is *weakly safe belief* $\Box_a^{\text{weak}}$

given by the following semantics:

$$\mathcal{M}, w \models \Box_a^{\text{weak}}\varphi \quad \text{iff} \quad \mathcal{M}, w \models \varphi \text{ and for all } v \in W, \text{ if } w <_a v, \text{ then } \mathcal{M}, v \models \varphi,$$

where $<_a$ is defined by; $w <_a v$ if and only if $w \leq_a v$ and $v \not\leq_a w$. Finally, Baltag and Smets (2008) define *strong belief* $S\,b_a$ by

$$S\,b_a\varphi \text{ iff } B_a\varphi \wedge K_a(\varphi \rightarrow \Box_a\varphi).$$

In addition to the several notions of belief, Baltag and Smets (2008) also discuss several ways of updating knowledge and beliefs when new information comes about. These are *update*, *radical upgrade*, and *conservative upgrade* and can be distinguished by the trust that is put in the source of the new information. If the source is known to be infallible, it should be an update. If the source is highly reliable, it should be a radical upgrade and if the source is just barely trusted, it should be a conservative upgrade. In this paper we are interested in what it takes to dissolve pluralistic ignorance and since update is the "strongest" way of updating knowledge and beliefs, we will focus on this. We will also refer to this way of updating as *public announcement*.

We introduce operators $[!\varphi]$, and add to the syntax the clause that for all formulas φ and ψ, $[!\varphi]\psi$ is also a formula. $[!\varphi]\psi$ is read as "after an announcement of φ, ψ is true". Semantically, a public announcement of φ will result in a new plausibility model where all the $\neg\varphi$-worlds have been removed, and the truth of ψ is then checked in this new model. These intuitions are made formal in the following definition:

Definition 2.3. Given a plausibility model $\mathcal{M} = \langle W, (\leq_a)_{a\in\mathbb{A}}, V\rangle$ and a formula φ, we define a new model $\mathcal{M}_{!\varphi} = \langle W', (\leq'_a)_{a\in\mathbb{A}}, V'\rangle$ by,

$$\begin{aligned} W' &= \{w \in W \mid \mathcal{M}, w \models \varphi\} \\ \leq'_a &= \leq_a \cap (W' \times W') \\ V'(p) &= V(p) \cap W' \end{aligned}$$

The semantics of the public announcement formulas are then given by:

$$\mathcal{M}, w \models [!\varphi]\psi \quad \text{iff} \quad \text{if } \mathcal{M}, w \models \varphi \text{ then } \mathcal{M}_{!\varphi}, w \models \psi.$$

Finally, we add to the framework of Baltag and Smets (2008) the two notions of ignorance and doubt. These are notions rarely discussed in the literature on epistemic/doxastic logic. However, ignorance is discussed in (van der Hoek and Lomuscio 2004). On the syntactic level we add two new operators I_a and D_a for each agent $a \in \mathbb{A}$. The formula $I_a\varphi$ is read as "agent a is ignorant about φ" and $D_a\varphi$ is read as "agent a

doubts whether φ". The semantics of these operators are defined from the semantics of the knowledge operator and the belief operator:

Definition 2.4. The operators I_a and D_a are defined be the following equivalences:

$$\begin{aligned} I_a\varphi &:= \neg K_a\varphi \wedge \neg K_a\neg\varphi \\ D_a\varphi &:= \neg B_a\varphi \wedge \neg B_a\neg\varphi. \end{aligned}$$

Note that, since $K_a\varphi \rightarrow B_a\varphi$ is valid, $D_a\varphi \rightarrow I_a\varphi$ is also valid.

3 Modeling pluralistic ignorance

Based on the logic introduced in the previous section, we will now formalize different versions of pluralistic ignorance that are all consistent. Then, we will discuss whether these formalizations make pluralistic ignorance into a fragile phenomenon.

3.1 Formalizations and consistency of pluralistic ignorance

As discussed in Section 1, there are many ways of defining pluralistic ignorance and in this section we attempt to formalize a few of these. We will also discuss whether these formalizations lead to consistent concepts in the sense that the formalizations are by satisfiable formulas.

Firstly, we assume that pluralistic ignorance is a situation where no agent believes φ, but every agent believes that everyone else believes φ. This can easily be formalized as:

$$\bigwedge_{a\in\mathbb{A}}\Big(\neg B_a\varphi \wedge \bigwedge_{b\in\mathbb{A}\setminus\{a\}} B_aB_b\varphi\Big). \tag{1}$$

For boolean formulas φ[4], the formula (1) is satisfiable since a plausibility model can easily be constructed such that it contains a possible world that satisfies it. Such a model is given in Figure 1, where we assume that the set of agents is $\mathbb{A} = \{a_1, a_2, ..., a_n\}$. In the following, when drawing models like this one, an arrow from a state w to a state v labeled by a_i will represent that $w <_{a_i} v$ holds in the model. An arrow from w to v labeled by a set $B \subseteq \mathbb{A}$ represent that $w <_b v$ for all $b \in B$. The full plausibility relations of the model will be the reflexive transitive closures of the relations drawn in the pictures. When a formula φ appears next to a state it means that φ is true at that state.

[4] A formula is boolean if it constructed solely from propositional variables and the logical connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$.

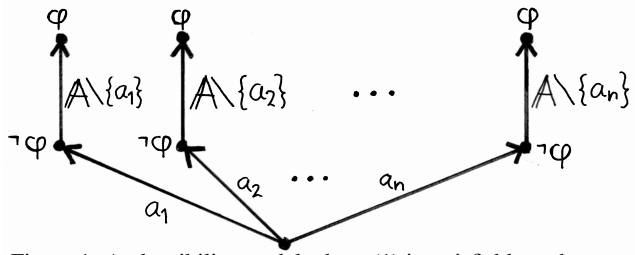


Figure 1: A plausibility model where (1) is satisfiable at the root

There are also formulas φ for which (1) is unsatisfiable, take for instance φ to be $B_b\psi$ or $\neg B_b\psi$ for any agent $b \in \mathbb{A}$ and any formula ψ. It cannot be the case that agent a does not believe that agent b believes that ψ, but at the same time a believes that b believes that b believes that ψ, i.e. (1) is unsatisfiable when φ is $B_b\psi$ or $\neg B_b\psi$ because $\neg B_a B_b\psi \wedge B_a B_b B_b\psi$ and $\neg B_a \neg B_b\psi \wedge B_a B_b \neg B_b\psi$ are unsatisfiable. In the following, when discussing pluralistic ignorance as defined by (1) we will therefore assume that φ is a boolean formula.

If belief is replaced by strong belief, such that (1) becomes

$$\bigwedge_{a\in\mathbb{A}} \Big(\neg S\, b_a\varphi \wedge \bigwedge_{b\in\mathbb{A}\setminus\{a\}} S\, b_a S\, b_b\varphi\Big). \tag{2}$$

pluralistic ignorance remains satisfiable for boolean formulas, which is testified by Figure 1 again. Furthermore, (2) is also not satisfiable if φ is of the form $S\, b_b\psi$ or $\neg S\, b_b\psi$ for a $b \in \mathbb{A}$. However, if we use safe belief and weak safe belief instead of belief in (1), pluralistic ignorance becomes unsatisfiable. This is obvious since both safe belief and weak safe belief implies truth.

In the classroom example of Section 1, a better definition of pluralistic ignorance may be obtained using the ignorance operator. This leads to the following definition of pluralistic ignorance:

$$\bigwedge_{a\in\mathbb{A}} \Big(I_a\varphi \wedge \bigwedge_{b\in\mathbb{A}\setminus\{a\}} B_a \neg I_b\varphi\Big). \tag{3}$$

This formula expresses a case where all the agents are ignorant about φ, but believe that all the other agents are not ignorant about φ. Instead of ignorance, doubt could be

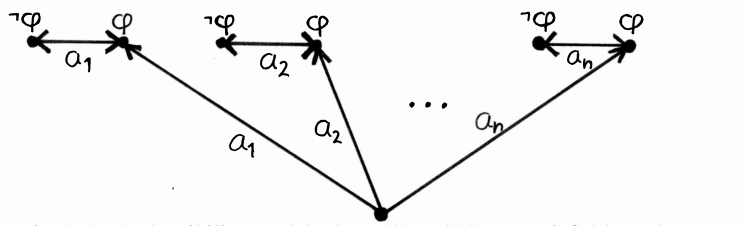


Figure 2: A plausibility model where (3) and (4) are satisfiable at the root

used as well, providing yet another definition of pluralistic ignorance:

$$\bigwedge_{a\in\mathbb{A}}\Big(D_a\varphi \wedge \bigwedge_{b\in\mathbb{A}\setminus\{a\}} B_a\neg D_b\varphi\Big). \tag{4}$$

Note that, since $D_a\varphi \rightarrow I_a\varphi$, (4) implies (3).

The two definitions of pluralistic ignorance (3) and (4) are also easily seen to be satisfiable for boolean formulas φ. This is made apparent by Figure 2. Now, however, formulas of the form $B_b\varphi$, for $b \in \mathbb{A}$, can also be subject to pluralistic ignorance. It is possible that agent a can doubt whether agent b believes φ and at the same time believe that agent b does not doubt whether he himself /agent b believes φ.

In (3) and (4) we can also replace the belief operator by the strong belief operator and obtain the following versions of pluralistic ignorance:

$$\bigwedge_{a\in\mathbb{A}}\Big(I_a\varphi \wedge \bigwedge_{b\in\mathbb{A}\setminus\{a\}} S\,b_a\neg I_b\varphi\Big), \tag{5}$$

$$\bigwedge_{a\in\mathbb{A}}\Big(D_a\varphi \wedge \bigwedge_{b\in\mathbb{A}\setminus\{a\}} S\,b_a\neg D_b\varphi\Big). \tag{6}$$

These new definitions of pluralistic ignorance are consistent as they are satisfiable at the root of the model in Figure 2. We still cannot obtain versions of (3) and (4) with safe belief and weak safe belief for the same reason as before.

It seems obvious that we can formalize even further versions of pluralistic ignorance within this framework. Thus, using the logic introduced in Section 2, we can characterize and distinguish many different versions of pluralistic ignorance. Furthermore, all the definitions (1)-(6) were satisfiable, which seems to entail that the concept of pluralistic ignorance is not inconsistent.

3.2 The fragility of pluralistic ignorance

After having formalized different versions of pluralistic ignorance, we can ask whether any of the definitions entail that pluralistic ignorance is a fragile phenomenon. However, first of all we need to spell out what we mean by a fragile phenomenon. The question of whether pluralistic ignorance is fragile or not reduces to the question of what it takes to dissolve it. We will regard pluralistic ignorance as dissolved only when none of the agents have wrong beliefs about the other agents' beliefs anymore.[5] The way agents can change their beliefs, will in this section be modeled by the $[!\varphi]$ operators of Section 2.

For the time being, we fix pluralistic ignorance to be defined as (1). According to several descriptions of pluralistic ignorance, it should be dissolved if just one agent announces his true belief. If the formula $!\neg B_b\varphi$ is announced, it naturally follows that $\bigwedge_{a\in\mathbb{A}} B_a\neg B_b\varphi$. However, this does not dissolve the pluralistic ignorance since all agents might keep their wrong beliefs about any other agent than b. In other words, a model satisfying (1) can be constructed such that after the announcement of $!\neg B_b\varphi$ it still holds that $\bigwedge_{a\in\mathbb{A}}\left(\bigwedge_{c\in\mathbb{A}\setminus\{a,b\}} B_aB_c\varphi\right)$.

It turns out that there is nothing in the definition (1) that prevents the wrong beliefs of the agents from being quite robust. Even if everybody except an agent c announce that they do not believe φ, all the agents might still believe that c believes φ. Using a formula of $\mathcal{L}$ we can define a notion of robustness in the following way: *agent a robustly believes that the group of agents $B\subseteq\mathbb{A}\setminus\{a\}$ believes φ* if

$$\bigwedge_{C\subseteq B}\Big([!\neg B_c\varphi]_{c\in C}(\bigwedge_{b\in B\setminus C} B_aB_b\varphi)\Big), \tag{7}$$

where $[!\neg B_c\varphi]_{c\in C}$ is an abbreviation for $[!\neg B_{c_1}\varphi][!\neg B_{c_2}\varphi]...[!\neg B_{c_k}\varphi]$, when $C=\{c_1,c_2,...,c_k\}$.[6] An example of a model where agent 1 believes $\neg\varphi$ and robustly believes that the agents $\{2,3,4,5\}$ believe φ is shown in Figure 3.

Another way of looking at the formula (7) is that it describes a situation where agent a believes that all the other agents' beliefs about φ are independent; maybe they

[5] In other words, if pluralistic ignorance in form of $\bigwedge_{a\in\mathbb{A}}\left(\neg B_a\varphi\wedge\bigwedge_{b\in\mathbb{A}\setminus\{a\}}B_aB_b\varphi\right)$ is the case, we will regard it as dissolved only when $\bigwedge_{a\in\mathbb{A},b\in\mathbb{A}\setminus\{a\}}B_a\neg B_b\varphi$ is the case. Thus, from pluralistic ignorance being the case in its "full" form $\bigwedge_{a\in\mathbb{A}}\left(\neg B_a\varphi\wedge\bigwedge_{b\in\mathbb{A}\setminus\{a\}}B_aB_b\varphi\right)$ until it is dissolved in our terminology there might be intermediate situations where pluralistic ignorance is neither properly dissolved or is the case in its full form.

[6] An alternative to (7) is

$$\bigwedge_{C\subseteq B}\Big([!\bigwedge_{c\in C}\neg B_c\varphi](\bigwedge_{b\in B\setminus C} B_aB_b\varphi)\Big),$$

however, the two are not equivalent. We will not go into a discussion of which definition is preferable.

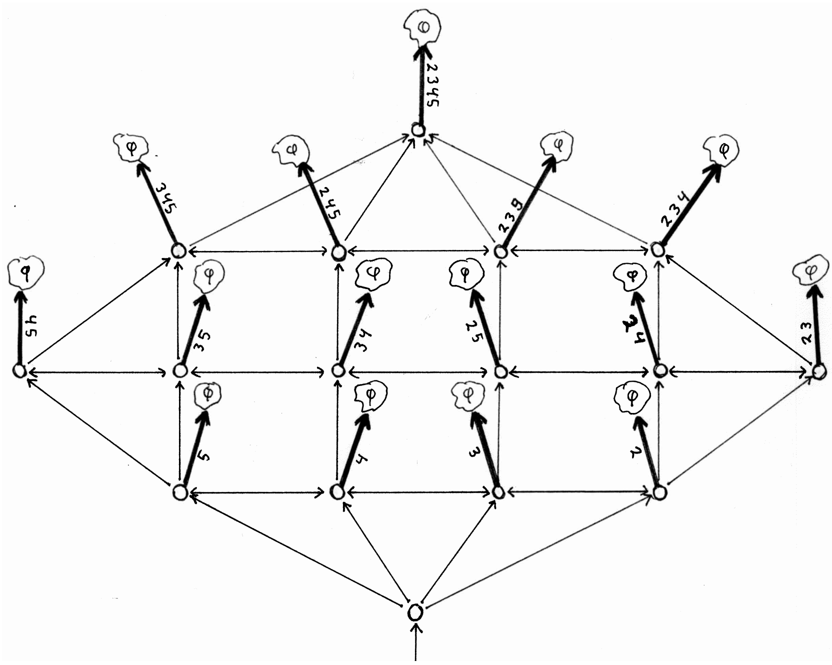


Figure 3: A robust model where agent 1 believes $\neg\varphi$ and has a strong robust belief in that the agents 2, 3, 4, and 5 believe φ. The worlds marked with "$\circ$" are worlds where $\neg\varphi$ is true and the "clouds" marked with φ are collections of worlds where φ is true all over. The arrows not marked with numbers represent the plausibility relation for agent 1 only.

all believe φ for different reasons. Thus, learning about some agents' beliefs about φ tells a nothing about what the other agents believe about φ.

With robustness defined by (7), pluralistic ignorance is consistent with all the agents having wrong robust beliefs about the other agents' beliefs. Taking disjoint copies of the model in Figure 3 for each agent and joining the roots shows that:

Proposition 1. *Pluralistic ignorance in form of* (1) *is consistent with that all the agents* $a \in \mathbb{A}$*, robustly believes that the group of agents* $\mathbb{A} \backslash \{a\}$ *believe* φ*.*

Another way of interpreting this result is that announcements of the true beliefs of some of the involved agents are not enough to dissolve pluralistic ignorance. Either all the agents need to announce their true beliefs or new information has to come from

an outside trusted source. Thus, announcements of the forms $!B_a\varphi$ or $!\neg B_a\varphi$ are not guaranteed to dissolve pluralistic ignorance. However, a public announcement of $!\neg\varphi$ in the model of Figure 3 will remove the pluralistic ignorance. But an announcement of the form $!\neg\varphi$ (or $!\varphi$) is precisely an announcement from an trusted outsider. An agent a can only announce formulas of the form $!B_a\psi$ or $!\neg B_a\psi$.

What turns pluralistic ignorance into a fragile phenomenon in most cases, is the fact that the agents consider the other agents' beliefs not to be independent as is the case if (7) is satisfied. In other words, pluralistic ignorance in the fragile form occurs mainly when the beliefs of the involved agents are correlated. This fits well with the view that pluralistic ignorance is a genuine social phenomenon as claimed by O'Gorman (1986).

Proposition 1 only regards pluralistic ignorance as defined by (1). However, for the definitions (3) and (4) similar results hold. Neither of the definitions (3) and (4) entail that pluralistic ignorance is fragile to public announcements of doubts ($[!D_b\varphi]$) or ignorance ($[!I_b\varphi]$). We can construct a new model, similar to the one i Figure 3, in which an agent a doubts whether φ but has a strong robust belief in that all the agents in $\mathbb{A}\backslash\{a\}$ do not doubt whether φ (and the same goes for ignorance). This new model is shown in Figure 4.

When it comes to the definitions of pluralistic ignorance based on strong beliefs (2), (5), and (6), something interesting happens. In the model of Figure 3 agent 1 does not have a strong belief that the other agents have strong beliefs in φ. For instance, there is a state where $B_1 S b_5\varphi$ and $S b_5\varphi$ are true, but $\Box_1 S b_5\varphi$ is not true. The same issue occurs in the model of Figure 4. It is still unknown whether robust models can be constructed such that they satisfy the strong belief versions of pluralistic ignorance as defined by (2), (5), and (6). Thus, it is left for further research whether there are strong belief versions of pluralistic ignorance that are not fragile. There are several other questions for further research, which we will turn to now.

4 Further research on logic and pluralistic ignorance

We have given several consistent formalizations of pluralistic ignorance, but there still seems to be more possible variations to explore. Furthermore, we have been working within one specific framework, and the question remains of whether there are other natural frameworks in which all formalizations of pluralistic ignorance become inconsistent. This would be highly unexpected though. Another question regarding formalizations of pluralistic ignorance in different frameworks is whether it changes the fragility of the phenomenon. This is still an open question.

Even though pluralistic ignorance needs not be fragile, neither as a "real life" phenomenon nor according to the formalizations given in this paper, it seems that the

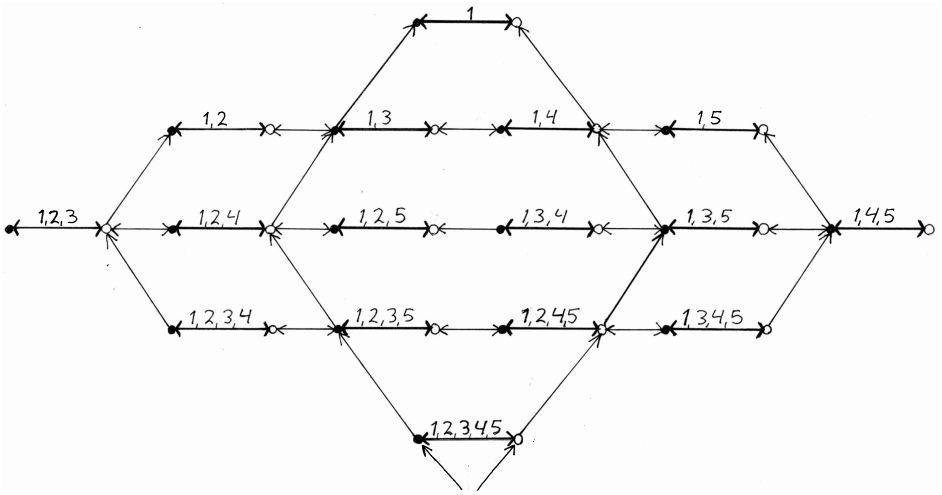


Figure 4: A robust model where agent 1 doubts whether φ, has a strong robust belief in that the agents 2, 3, 4, and 5 do not doubt whether φ. The worlds marked with "$\circ$" are worlds where $\neg\varphi$ is true and the worlds marked with "$\bullet$" are worlds where φ is true. The arrows with no numbers on are arrows for agent 1. Remember that the full plausibility relations of the model are the reflexive transitive closures of the arrows in the pictures

really interesting cases occur when pluralistic ignorance is, in fact, fragile. Whether pluralistic ignorance is fragile appears to be closely related to it being a genuine social phenomenon; the dependence between agents' beliefs is what makes pluralistic ignorance fragile. Thus, the real interesting question for further research is how agents' beliefs are interdependent in the case of pluralistic ignorance and how best to model this in logic. In answering this question, a shift in focus from what it takes to dissolve pluralistic ignorance, to what it takes for pluralistic ignorance to arise, seems natural.

4.1 Informational Cascades: How pluralistic ignorance comes about and how it vanishes

An agent's beliefs may depend on other agents' beliefs in many ways; one way is through testimony of facts by other agents in which the agent trusts. Modeling trust and testimony is for instance done in Holliday (2010). Another way in which agents'

beliefs may depend on each other could be through a common information source (Bikhchandani et al. 1998). Yet another way is through informational cascades.

Informational cascades are a phenomenon widely discussed in the social sciences (Bikhchandani et al. 1998, Lohmann 1994) and was introduced by Bikhchandani et al. (1992). When actions/signals are preformed sequentially and agents start to ignore their private information and instead base their actions/signaling merely on information obtained from the actions/signals of the previous individuals, an *informational cascade* has occurred. If the acts of the first people in the cascade oppose to their private beliefs and the remaining people join in with the same actions (also oppose to their private beliefs) the result might be a case of pluralistic ignorance. However, informational cascades are also fragile (Bikhchandani et al. 1992) and opposite cascades may occur, thus eliminating pluralistic ignorance again.

These kinds of informational cascades, which have been shown to occur in numerous of places, may very well be the cause of pluralistic ignorance. Hence, logical framework that can model informational cascades might also be suited to model pluralistic ignorance. To the knowledge of the author, the only paper on logic-based models of informational cascades is (Holliday 2010), but it may very well be possible to model pluralistic ignorance in that framework. However, further work on the logics of informational cascades is still to come.

4.2 Private versus public beliefs – the need for new notions of group beliefs

The concept of pluralistic ignorance, regardless of which version one adopts, seems to hint at the need for new notions of common knowledge/beliefs. Pluralistic ignorance can be viewed as a social phenomenon where everybody holds a private belief in φ, but publicly display a belief in $\neg\varphi$ and thus contribute to a "common belief" ("public belief" might be a better world) in $\neg\varphi$. Due to the usual definition of common belief (everybody believes φ and everybody believes that everybody believes φ and ...), a common belief in $\neg\varphi$ leads to private belief in $\neg\varphi$ for all agents in the group, but this is exactly the thing that fails in social epistemic scenarios involving pluralistic ignorance. Hence, a new notion of common group belief seems to be needed. In general, there are various ways in which group beliefs can be related to the beliefs of individuals of the group. Thus, a logic that distinguishes between private and public beliefs or contains new notions of common beliefs may help model pluralistic ignorance more adequately. Once again, this is left for further research.

4.3 How agents act

The way agents act in cases of pluralistic ignorance also seems to play an important role. The reason why most students believe that other students are comfortable with drinking might be that they observe the other students drinking heavily. In the classroom example students are also obtaining their wrong beliefs based on the observation of others. Furthermore, focusing on actions might also tell us something about how pluralistic ignorance evolves in the first place.

Therefore, a logic combining beliefs and actions might be the natural tool for modeling pluralistic ignorance. There exist several logics that combine beliefs/knowledge and actions, but which one to chose and the actual modeling, is left for further research to decide.

5 Conclusion

Firstly, we have seen that there are many ways of defining pluralistic ignorance, all of which by satisfiable formulas. Therefore, pluralistic ignorance is (seemingly) not a phenomenon that goes against logic. In other words, wrong logical reasoning is not necessarily involved in pluralistic ignorance.

Secondly, the standard definitions of pluralistic ignorance, for instance as a situation where no one believes, but everyone believes that everyone else believes, do not entail that the phenomenon is fragile. Public announcements of the true beliefs of some of the involved agents are not enough to dissolve pluralistic ignorance. Either all the agents need to announce their true beliefs or new information has to come from an outside, trusted source. However, pluralistic ignorance often seems to occur in cases where the agents' beliefs are correlated and in such cases pluralistic ignorance might be increasingly more fragile.

The paper has hinted at a first logic approach to pluralistic ignorance. Some features and problems have been singled out, but the main aim of the paper was to pave the way for further research into logical modeling of social phenomena such as pluralistic ignorance.

6 Postscript

Since the first appearance of this paper in *Future Directions for Logic - Proceedings of PhDs in Logic III*, College Publications, 2012, several philosophical and logical papers addressing pluralistic ignorance have appeared. In this postscript, we will discuss some of them.

First of all, the issue of how exactly to define pluralistic ignorance is thoroughly addressed by Bjerring et al. (to appear), as well. They cite several versions of pluralistic ignorance, and discuss the relationship between them, before settling on a version of pluralistic ignorance that highlights the key epistemic and social interactive aspects of the phenomenon. In their definition, the behavior of the agents involved in pluralistic ignorance play an essential role, and as such, they take the considerations of Section 4.3 seriously. However, it should be noted that the inclusion of the agents' behavior in the definition of pluralistic ignorance makes their definition of pluralistic ignorance diverge substantially from any of the definitions given in this paper.

It is argued in this paper that pluralistic ignorance is not a phenomenon that goes against logic. However, one may still think that rational agents cannot find themselves in a situation of pluralistic ignorance without it being dissolved immediately. Contrary to this, Bjerring et al. (to appear) also provide an in-depth philosophical argument as to why pluralistic ignorance might arise and persist among rational agents. Thus, it might takes more than just correct logical and rational reasoning to avoid or dissolve pluralistic ignorance.

The question of dissolving pluralistic ignorance and whether the phenomenon is fragile is also addressed by Proietti and Olsson (2013). They build on the framework of this paper, but go on to discuss what happens if agents are equipped with *descriptive norms of assertion* that specify when they announces their true beliefs. Each agent is assigned a percentage threshold such that she will announce her true belief if the percentage of her peers announcing similar beliefs exceeds her threshold. If the agents are arranged in a particular social network structure and do not have to high thresholds, Proietti and Olsson (2013) show that pluralistic ignorance might be dissolved if the right agent starts announcing her true belief. The pluralistic ignorance is dissolved through an cascade just as mentioned in Section 4.1.

In addition to (Holliday 2010) another paper providing logical models of informational cascades has been published since this paper first occurred. Baltag et al. (2013) provide two logical models of informational cascades based on two evidence logics. Whether these logics are useful in modeling pluralistic ignorance is still an open question.

In Section 4.2, the distinction between private and public beliefs was argued to be important for pluralistic ignorance. In a recent paper, Christoff and Hansen (2013) have developed a simple logic that exactly distinguishes between private and public beliefs. Their model of beliefs is significantly simpler than the one used in this paper based on plausibility models. However, Christoff and Hansen (2013) also include machinery to talk about the social network structure of the agents and a notion of social influence, based on the work by Zhen and Seligman (2011), Seligman et al. (2011). With this notion of social influence, Christoff and Hansen (2013) show that pluralistic ignorance

is a "robust" state in the sense that it constitutes a fix-point under the introduced notion of social influence. On the other hand, they also show that pluralistic ignorance is fragile under their notion social influence in the sense that pluralistic ignorance might be dissolve if just one agent announces her true belief. Moreover, the situations where pluralistic ignorance might be dissolved if just one agent announces her true belief is completely characterized by the social network structure the agents are arranged in (Christoff and Hansen 2013). Thus, the work by Proietti and Olsson (2013) and Christoff and Hansen (2013) have contributed to a better understanding of the fragility of pluralistic since this paper was first published.

Finally, Hansen et al. (2013) consider pluralistic ignorance in connection with other information phenomena such as informational cascades, bystander effects and belief polarization, and discuss how such phenomena might be magnified significantly by modern information technologies.

Acknowledgements This paper is a slightly revised and expanded version of of a paper of the same title occurring in Jonas De Vuyst and Lorenz Demey (eds.), *Future Directions for Logic: Proceedings of PhDs in Logic III*, IfColog Proceedings Volume 2, College Publications, London, ISBN 978-1848900790, 2012.

In connection with that paper, the author would like to thank the participants of the workshop PhDs in Logic III (Brussels, February 17-18th, 2011) and the participants of the Copenhagen-Lund workshop in Social Epistemology (Lund, February 25th, 2011). Furthermore, the author would like to thank Jens Christian Bjerring, Nikolaj Jang Lee Linding Pedersen, Carlo Proietti, Vincent F. Hendricks, Eric Pacuit, and Olivier Roy for useful discussions and comments.

In connection with this updated version of the paper, the author would like to thank the organizers and participants of the LIRa seminar 2013 as well as the participants of the ILLC January student project "Social Dynamics of Information and its Distortions".

References

A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In W. van der Hoek and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory, Texts in Logic and Games, Vol 3*, pages 9–58. Amsterdam University Press, 2008.

A. Baltag, Z. Christoff, J. U. Hansen, and S. Smets. Logical models of informational cascades. In J. van Benthem and F. Lui, editors, *Logic across the University: Founda-*

tions and Applications, Studies in Logic, pages 405–432. College Publications, 2013. ISBN 978-1-84890-122-3.

J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.

S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5): 992–1026, 1992.

S. Bikhchandani, D. Hirshleifer, and I. Welch. Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives*, 12 (3):151–170, 1998.

J. C. Bjerring, J. U. Hansen, and N. J. L. L. Pedersen. On the rationality of pluralistic ignorance. *Synthese*, to appear.

D. Centola, R. Willer, and M. Macy. The emperor's dilemma. A computational model of self-enforcing norms. *American Journal of Sociology*, 110(4):1009–1040, 2005.

Z. Christoff and J. U. Hansen. A two-tiered formalization of social influence. In H. Huang, D. Grossi, and O. Roy, editors, *Logic, Rationality and Interaction, Proceedings of the Fourth International Workshop (LORI 2013)*, volume 8196 of *Lecture Notes in Computer Science*, pages 68–81. Springer, 2013.

P. G. Hansen, V. F. Hendricks, and R. K. Rendsvig. Infostorms. *Metaphilosophy*, 44 (3):301–326, 2013.

V. F. Hendricks. Knowledge transmissibility and pluralistic ignorance: A first stab. *Metaphilosophy*, 41(3):279–291, 2010.

W. H. Holliday. Trust and the dynamics of testimony. In D. Grossi, L. Kurzen, and F. Velázquez-Quesada, editors, *Logic and Interactive Rationality – Seminar's Yearbook 2009*, pages 147–178. Institute for Logic, Language and Computation, Universiteit van Amsterdam, 2010.

D. Katz and F. H. Allport. *Student Attitudes*. Syracuse, N. Y.: The Craftsman Press, 1931.

D. Krech and R. S. Crutchfield. *Theories and Problems of Social Psychology*. New York: McGraw-Hill, 1948.

S. Lohmann. The dynamics of informational cascades: The monday demonstrations in leipzig, east germany, 1989-91. *World Politics*, 47(October):42–101, 1994.

H. J. O'Gorman. The discovery of pluralistic ignorance: An ironic lesson. *Journal of the History of the Behavioral Sciences*, 22(October):333–347, 1986.

D. A. Prentice and D. T. Miller. Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, 64(2):243–256, 1993.

C. Proietti and E. J. Olsson. A DDL approach to pluralistic ignorance and collective belief. *Journal of Philosophical Logic*, 2013.

J. Seligman, F. Liu, and P. Girard. Logic in the community. In M. Banerjee and A. Seth, editors, *Logic and Its Applications*, volume 6521 of *Lecture Notes in Computer Science*, pages 178–188. Springer, 2011.

W. van der Hoek and A. Lomuscio. A logic for ignorance. *Electronic Notes in Theoretical Computer Science*, 85(2):117–133, 2004.

L. Zhen and J. Seligman. A logical model of the dynamics of peer pressure. *Electronic Notes in Theoretical Computer Science*, 278(0):275–288, 2011. Proceedings of the 7th Workshop on Methods for Modalities (M4M 2011) and the 4th Workshop on Logical Aspects of Multi-Agent Systems (LAMAS 2011).

Bubbles

Vincent F. Hendricks

University of Copenhagen
vincent@hum.ku.dk

The more you can create that magic bubble, that suspension of disbelief, for a while, the better.

- Edward Norton

The term "bubble" has traditionally been associated with a particular situation occurring on financial markets:

> A bubble is considered to have developed when assets trade at prices that are far in excess of an estimate of the fundamental value of the asset, as determined from discounted expected future cash flows using current interest rates and typical long-run risk premiums associated with asset class. Speculators, in such circumstances, are more interested in profiting from trading the asset than in its use or earnings capacity or true value. (Vogel 2010, p. 16)

Textbook examples of bubbles include the Dutch tulip bulbs frenzy in the 1600s, the South Sea and Mississippi excesses about a century later, the US stock market as of 1929, the Japanese real estate and equity markets of the 1980s, the dot.com period and Internet stock boom of the 1990s, and of course the balloons, frenzies and speculative mania in the world economy leading to the global financial crisis of 2008 of which we are still in the midst of the aftermath.

In wake of the current crisis there have been many suggestions as to why financial bubbles occur, most of them composites in terms of explanatory factors involving

different mixing ratios of bubble-hospitable market configurations and social psychological features of human nature and informational phenomena like the ones discussed here in *Infostorms* (Hendricks and Rasmussen 2012).

One seemingly paradoxical hypothesis suggests that too much liquidity is actually poisonous rather than beneficial for a financial market (Buchanan 2008). Monetary liquidity in excess stimulated by easy access to credit, large disposable incomes and lax lending standards combined with expansionary monetary policies of lowering interests by banks and advantageous tax breaks and bars by the state, flush the market with capital. This extra liquidity leaves financial markets vulnerable to volatile asset price inflation the cause of which is to be found in short-term and possibly leveraged speculation by investors.

The situation becomes that too much money chases too few assets, good as well as bad, both of which in return are elevated well beyond their fundamental value to a level of general unsustainability. Pair up too much liquidity with robustly demonstrated socio-psychological features of human nature like boom-thinking, group-thinking, herding, informational cascades and other aggregated phenomena of social proof, it becomes a matter of time before the bubbles start to burst (Lee 1998) - at least in finance.

However, behind every financial bubble, crash and subsequent crisis "lurks a political bubble - policy biases that foster market behaviors leading to financial instability" (McCarty and Rosenthal 2013) with reference to the 2008 financial crunch. Thus there are political bubbles too ... and other sorts as well.

There are stock, real-estate and other bubbles associated with financial markets but also filter bubbles, opinion bubbles, political bubbles, science bubbles, social bubbles, status bubbles, fashion bubbles, art bubbles ... all pushing collectives of agents in the same (often unfortunate) direction; not only buying the same stock or real estate but also thinking the same thing, holding the same opinions, appreciating the same art, "liking" the same posts on social media, purchasing the same brand names, subscribing to the same research program in science etc.

Internet activist Eli Pariser coined the term "filter bubble" (Pariser 2011) to refer to selective information acquisition by website algorithms (in search-engines, news feeds, flash messages, tweets) personalizing search results for users based on past search history, click behavior and location accordingly filtering away information in conflict, with user interest, viewpoint or opinion. An automated but personalized information selection process in line with polarization mechanics isolating users in their cultural, political, ideological or religious bubbles. Filter bubbles may stimulate individual narrow-mindedness but are also potentially harmful to the general society undermining informed civic or public deliberation, debate and discourse making citizens ever more susceptible to propaganda and manipulation:

> A world constructed from the familiar is a world in which there's nothing to learn ... (since there is) invisible autopropaganda, indoctrinating us with our own ideas. (Pariser 2011, *The Economist*, June 30)

Harvesting or filtering information in a particular way is part of aggregating opinion. One may invest an opinion on the free market place of ideas and a certain idea or stance, whether political, religious or otherwise, may at a certain point gain popularity or prominence and become an asset by the number of people apparently subscribing to it in terms of likes, upvotes, clicks or similar endorsements of minimum personal investment. Public opinion tends to shift depending on a variety of factors ranging from zeitgeist, new facts, current interests to premiums of social imprimatur. Opinion bubbles may accordingly suddenly go bust or gradually deflate depending.

Everyday personal opinions can serve as intellectual liquidity chasing assets of political or cultural ideas. But scientific inquiry may also be geared with too much intellectual liquidity in terms of explanatory expectations and available funding, paired up with boom-thinking in the scientific community. The short-term and possibly leveraged speculation by scientist may exactly occur in the way characterizing a ballooning market - science bubbles emerge (Pedersen and Hendricks 2013). The modern commercialization of science and research has even been compared to downright Ponzi-schemes only surviving as long as you can steal from Peter to pay Paul scientifically so to speak (Mirowski 2013).

Fashion in particular rely on getting everybody, or a selective few, to trend the same way - that's the point of the entire enterprise besides the occasional claim to artistic diligence. But even the art scene is tangibly ridden with bubbles: "The bubble that is Con Art blew up, like the sub-prime mortgage business, in the smoke-and-mirrors world of financial markets, where fortunes have been made on nothing" says Julian Spalding to *The Independent* (March 26, 2012), famous British gallery owner commenting on his recent book *Con Art - Why you ought to sell your Damien Hirst's while you can* (2012).

The concept of bubbles appears in seemingly different spheres. Perhaps it is more than just terminological coincidence - across spheres bubbles share similar structure and dynamics - from science to society. Irrational group behavior fuels bubbles. For instance, individual scientists may have doubts about the merits of bibliometric evaluation or excessive publishing practices much in vogue these days. However a strong public signal aggregated by the previous actions and endorsements of colleagues and institutions suggesting an aggressive publication strategy and abiding to the regulatory rules of evaluation and funding schemes may suppress the personal doubt of the individual scientist. But when personal information gets suppressed in favor of a public signal regulating individual behavior it may in turn initialize a lemming-effect, an informational cascade. Now, informational cascades have proven robust features in

the generation of financial bubbles, where "individuals choose to ignore or downplay their private information and instead jump the bandwagon by mimicking the actions of individuals acting previously" (Vogel 2010, p. 85).

Informational cascades may thus be considered pivotal to building bubbles - in science and elsewhere, and using modern formal logic we have the means for uncovering their logical structure and dynamics independently of their realm of reign - and that's exactly what we are going to do (see Hansen and Rendsvig 2013, Rendsvig and Hendricks 2014, Hansen and Hendricks 2014).

Acknowledgements This paper is based on joint work with Henrik Boensvang, David Budtz Pedersen, Pelle G. Hansen and Rasmus K. Rendsvig.

References

M. Buchanan. Why economic theory is out of whack. *New Scientist*, July 2008.

P. Hansen and V. Hendricks. *Infostorms: How to Take Information Punches and Save Democracy*. New York Copernicus Books, 2014.

P. Hansen, R. Rendsvig, and V. Hendricks. Infostorms. *Metaphilosophy*, 44(3):301–326, 2013.

V. Hendricks and J. Rasmussen. *Nedtur! Finanskrisen forstÂet filosofisk*. Gyldendal Business, 2012.

I. Lee. Market crashes and informational avalanches. *Review of Economic Studies*, 65:741–759, 1998.

P. Mirowski. The modern commercialization of science is a passel of ponzi schemes. *Social Epistemology*, 26(4):285–310, 2013.

K. P. N. McCarty and H. Rosenthal. *Political Bubbles: Financial Crises and the Failure of American Democracy*. Princeton University Press, 2013.

E. Pariser. *The Filter Bubble: What the Internet is Hiding from you*. Penguin Books, 2011.

D. B. Pedersen and V. Hendricks. Science bubbles. *Philosophy and Technology*, in press, 2013.

R. Rendsvig and V. Hendricks. Social proof in extensive games. Submitted for Publication, 2014.

H. Vogel. *Financial Market Bubbles and Crashes*. Cambdridge University Press, 2010.

Don't Plan for the Unexpected: Planning Based on Plausibility Models

Mikkel Birkegaard Andersen, Thomas Bolander, and Martin Holm Jensen

DTU Compute, Technical University of Denmark
mibi@dtu.dk, tobo@dtu.dk, mhje@dtu.dk

Abstract

We present a framework for automated planning based on plausibility models, as well as algorithms for computing plans in this framework. The framework presented extends a previously developed framework based on dynamic epistemic logic (DEL), without plausibilities/beliefs. In the pure epistemic framework, one can distinguish between strong and weak epistemic plans for achieving some, possibly epistemic, goal. A strong plan guarantees that the agent achieves the goal, whereas a weak plan promises only the possibility of leading to the goal. Weak epistemic planning is not satisfactory, as there is no way to qualify which of two weak plans is more likely to lead to the goal. This seriously limits the practical uses of weak planning, as the planning agent might for instance always choose a plan that relies on serendipity. In the present paper we introduce a planning framework with the potential of overcoming this problem. The framework is based on plausibility models, allowing us to define different types of plausibility planning. The simplest type of plausibility plan is one in which the goal will be achieved when all actions in the plan turn out to have the outcomes found most plausible by the agent. This covers many cases of everyday planning by human agents, where we—to limit our computational efforts—only plan for the most plausible outcomes of our actions.

1 Introduction

Whenever an agent deliberates about the future with the purpose of achieving a goal, she is engaging in the act of planning. Automated Planning is a widely studied area of AI dealing with such issues under many different assumptions and restrictions. In this paper we consider *planning under uncertainty* (nondeterminism and partial observability, see Ghallab et al. 2004), where the agent has knowledge and beliefs about the environment and how her actions affect it. We formulate scenarios using plausibility models obtained by merging the frameworks in (Baltag and Smets 2006, van Ditmarsch and Kooi 2008).

Example 1 (The Basement)**.** An agent is standing at the top of an unlit stairwell leading into her basement. If she walks down the steps in the dark, it's likely that she will trip. On the other hand, if the lights are on, she is certain to descend unharmed. There is a light switch just next to her, though she doesn't know whether the bulb is broken.

She wishes to find a plan that gets her safely to the bottom of the stairs. Planning in this scenario is contingent on the situation; e.g. is the bulb broken? Will she trip when attempting her descent? In planning terminology a plan that *might* achieve the goal is a *weak solution*, whereas one that *guarantees* it is a *strong solution*.

In this case, a weak solution is to simply descend the stairs in the dark, risking life and limb for a trip to the basement. On the other hand, there is no strong solution as the bulb might be broken (assuming it cannot be replaced). Intuitively, the *best* plan is to flick the switch (expecting the bulb to work) and then descend unharmed, something neither weak nor strong planning captures.

Extending the approach in (Andersen et al. 2012) to a logical framework incorporating beliefs via a plausibility ordering, we formalise plans which an agent considers most likely to achieve her goals. This notion is incorporated into algorithms developed for the framework in (Andersen et al. 2012), allowing us to synthesise plans like the *best* one in Example 1.

In the following section we present the logical framework we consider throughout the paper. Section 3 formalises planning in this framework, and introduces the novel concept of plausibility solutions to planning problems. As planning is concerned with representing possible ways in which the future can unfold, it turns out we need a belief modality corresponding to a globally connected plausibility ordering, raising some technical challenges. Section 4 introduces an algorithm for plan synthesis (i.e., generation of plans). Further we show that the algorithm is terminating, sound and complete. To prove termination, we must define bisimulations and bisimulation contractions.

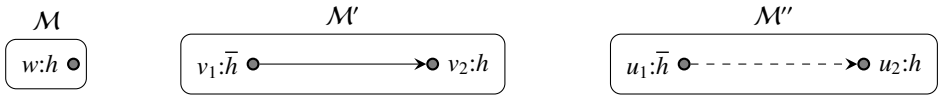


Figure 1: Three plausibility models

2 Dynamic Logic of Doxastic Ontic Actions

The framework we need for planning is based on a dynamic logic of doxastic ontic actions. Actions can be epistemic (changing knowledge), doxastic (changing beliefs), ontic (changing facts) or any combination. The following formalisation builds on the *dynamic logic of doxastic actions* (Baltag and Smets 2006), adding postconditions to event models as in (van Ditmarsch and Kooi 2008). We consider only the single-agent case. Before the formal definitions are given, we present some intuition behind the framework in the following example, which requires some familiarity with epistemic logic.

Example 2. Consider an agent and a coin biased towards heads, with the coin lying on a table showing heads (h). She contemplates tossing the coin and realizes that it can land either face up, but (due to nature of the coin) believes it will land heads up. In either case, after the toss she knows exactly which face is showing.

The initial situation is represented by the *plausibility model* (defined later) $\mathcal{M}$ and the contemplation by $\mathcal{M}''$ (see Figure 1). The two worlds u_1, u_2 are epistemically distinguishable ($u_1 \not\sim u_2$) and represent the observable non-deterministic outcome of the toss. The *dashed* directed edge signifies a *(global) plausibility relation*, where the direction indicates that she finds u_2 more plausible than u_1 (we overline proposition symbols that are false).

Example 3. Consider again the agent and biased coin. She now reasons about shuffling the coin under a dice cup, leaving the dice cup on top to conceal the coin. She cannot observe which face is up, but due to the bias of the coin believes it to be heads. She then reasons further about lifting the dice cup in this situation, and realises that she will observe which face is showing. Due to her beliefs about the shuffle she finds it most plausible that heads is observed.

The initial situation is again $\mathcal{M}$. Consider the model $\mathcal{M}'$, where the *solid* directed edge indicates a *local plausibility relation*, and the direction that v_2 is believed over v_1. By *local* we mean that the two worlds v_1, v_2 are *(epistemically) indistinguishable*

($v_1 \sim v_2$), implying that she is ignorant about whether h or $\neg h$ is the case.[1] Together this represents the concealed, biased coin. Her contemplations on lifting the cup is represented by the model $\mathcal{M}''$ as in the previous example.

In Example 2 the agent reasons about a non-deterministic action whose outcomes are *distinguishable* but not equally plausible, which is different from the initial contemplation in Example 3 where the outcomes are *not* distinguishable (due to the dice cup). In Example 3 she subsequently reasons about the observations made after a sensing action. In both examples she reasons about the future, and in both cases the final result is the model $\mathcal{M}''$. In Example 4 we formally elaborate on the actions used here.

It is the nature of the agent's ignorance that make $\mathcal{M}'$ and $\mathcal{M}''$ two inherently different situations. Whereas in the former she is ignorant about h due to the coin being concealed, her ignorance in the latter stems from not having lifted the cup yet. In general we can model ignorance either as a consequence of epistemic indistinguishability, or as a result of not yet having acted. Neither type subsumes the other and both are necessary for reasoning about actions. We capture this distinction by defining both local and global plausibility relations. The end result is that local plausibility talks about belief in a particular epistemic equivalence class, and global plausibility talks about belief in the entire model. We now remedy the informality we allowed ourselves so far by introducing the necessary definitions for a more formal treatment.

Definition 2.1 (Dynamic Language)**.** Let a countable set of propositional symbols P be given. The language $L(P)$ is given by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid B^{\varphi}\varphi \mid X\varphi \mid [\mathcal{E}, e]\,\varphi$$

where $p \in P$, $\mathcal{E}$ is an *event model* on $L(P)$ as (simultaneously) defined below, and $e \in D(\mathcal{E})$. K is the *local* knowledge modality, B^{φ} the *global* conditional belief modality, X is a (non-standard) *localisation* modality (explained later) and $[\mathcal{E}, e]$ the dynamic modality.

We use the usual abbreviations for the other boolean connectives, as well as for the dual dynamic modality $\langle \mathcal{E}, e\rangle\,\varphi := \neg\,[\mathcal{E}, e]\,\neg\varphi$ and unconditional (or absolute) global belief $B\varphi := B^{\top}\varphi$. The duals of K and B^{φ} are denoted $\widehat{K}$ and $\widehat{B}^{\varphi}$.

$K\varphi$ reads as "the (planning) agent knows φ", $B^{\psi}\varphi$ as "conditional on ψ, the (planning) agent believes φ", and $[\mathcal{E}, e]\,\varphi$ as "after all possible executions of $(\mathcal{E}, e)$, φ holds". $X\varphi$ reads as "locally φ".

Definition 2.2 (Plausibility Models)**.** A *plausibility model* on a set of propositions P is a tuple $\mathcal{M} = (W, \sim, \leq, V)$, where

[1] In the remainder, we use (in)distinguishability without qualification to refer to epistemic (in)distinguishability.

- W is a set of *worlds*,
- $\sim \subseteq W \times W$ is an equivalence relation called the *epistemic relation*,
- $\leq \subseteq W \times W$ is a connected well-preorder called the *plausibility relation*,[2]
- $V : P \to 2^W$ is a *valuation*.

$D(\mathcal{M}) = W$ denotes the *domain* of $\mathcal{M}$. For $w \in W$ we name $(\mathcal{M}, w)$ a *pointed plausibility model*, and refer to w as the *actual world* of $(\mathcal{M}, w)$. $<$ denotes the *strict plausibility relation*, that is $w < w'$ iff $w \leq w'$ and $w' \not\leq w$. $\simeq$ denotes *equiplausibility*, that is $w \simeq w'$ iff $w \leq w'$ and $w' \leq w$.

In our model illustrations a directed edge from w to w' indicates $w' \leq w$. By extension, strict plausibility is implied by unidirected edges and equiplausibility by bidirected edges. For the models in Figure 1, we have $v_1 \sim v_2$, $v_2 < v_1$ in $\mathcal{M}'$ and $u_1 \not\sim u_2$, $u_2 < u_1$ in $\mathcal{M}''$. The difference between these two models is in the epistemic relation, and is what gives rise to local (solid edges) and global (dashed edges) plausibility. In (Baltag and Smets 2006) the local plausibility relation is defined as $\trianglelefteq := \sim \cap \leq$; i.e., $w \trianglelefteq w'$ iff $w \sim w'$ and $w \leq w'$. $\trianglelefteq$ is a *locally well-preordered relation*, meaning that it is a union of *mutually disjoint well-preorders*. Given a plausibility model, the domain of each element in this union corresponds to an $\sim$-equivalence class.

Our distinction between local and global is not unprecedented in the literature, but it can be a source of confusion. In (Baltag and Smets 2006), $\leq$ was indeed connected (i.e. global), but in later versions of the framework (Baltag and Smets 2008) this was no longer required. The iterative development in (van Ditmarsch 2005) also discuss the distinction between local and global plausibility (named *preference* by the author). Relating the notions to the wording in (Baltag and Smets 2006), $\leq$ captures *a priori* beliefs about *virtual* situations, *before* obtaining any direct information about the actual situation. On the other hand, $\trianglelefteq$ captures *a posteriori* beliefs about an *actual* situation, that is, the agent's beliefs *after* she obtains (or assumes) information about the actual world.

$\mathcal{M}''$ represents two distinguishable situations (v_1 and v_2) that are a result of reasoning about the future, with v_2 being considered more plausible than v_1. These situations are identified by restricting $\mathcal{M}''$ to its $\sim$-equivalence classes; i.e., $\mathcal{M}'' \restriction \{v_1\}$ and $\mathcal{M}'' \restriction \{v_2\}$. Formally, given an epistemic model $\mathcal{M}$, the *information cells* in $\mathcal{M}$ are the submodels of the form $\mathcal{M} \restriction [w]_\sim$ where $w \in D(\mathcal{M})$. We overload the term and name any $\sim$-connected plausibility model on P an *information cell*. This use is slightly different from the notion in (Baltag and Smets 2008), where an information cell is an

[2] A well-preorder is a reflexive, transitive binary relation s.t. every non-empty subset has minimal elements (Baltag and Smets 2008).

$\sim$-equivalence class rather than a restricted model. An immediate property of information cells is that $\leq = \trianglelefteq$; i.e., the local and global plausibility relations are identical. A partition of a plausibility model into its information cells corresponds to a *localisation* of the plausibility model, where each information cell represents a local situation. The (later defined) semantics of X enables reasoning about such localisations using formulas in the dynamic language.

Definition 2.3 (Event Models)**.** An *event model* on the language $L(P)$ is a tuple $\mathcal{E} = (E, \sim, \leq, pre, post)$, where

- E is a finite set of *(basic) events*,
- $\sim \subseteq E \times E$ is an equivalence relation called the *epistemic relation*,
- $\leq \subseteq E \times E$ is a connected well-preorder called the *plausibility relation*,
- $pre : E \to L(P)$ assigns to each event a *precondition*,
- $post : E \to (P \to L(P))$ assigns to each event a *postcondition* for each proposition. Each $post(e)$ is required to be only finitely different from the identity.

$D(\mathcal{E}) = E$ denotes the *domain* of $\mathcal{E}$. For $e \in E$ we name $(\mathcal{E}, e)$ a *pointed event model*, and refer to e as the *actual event* of $(\mathcal{E}, e)$. We use the same conventions for accessibility relations as in the case of plausibility models.

Definition 2.4 (Product Update)**.** Let $\mathcal{M} = (W, \sim, \leq, V)$ and $\mathcal{E} = (E, \sim', \leq', pre, post)$ be a plausibility model on P resp. event model on $L(P)$. The *product update* of $\mathcal{M}$ with $\mathcal{E}$ is the plausibility model denoted $\mathcal{M} \otimes \mathcal{E} = (W', \sim'', \leq'', V')$, where

- $W' = \{(w, e) \in W \times E \mid \mathcal{M}, w \models pre(e)\}$,
- $\sim'' = \{((w, e), (v, f)) \in W' \times W' \mid w \sim v \text{ and } e \sim' f\}$,
- $\leq'' = \{((w, e), (v, f)) \in W' \times W' \mid e <' f \text{ or } (e \simeq' f \text{ and } w \leq v)\}$,
- $V'(p) = \{(w, e) \in W' \mid \mathcal{M}, w \models post(e)(p)\}$ for each $p \in P$.

The reader may consult (Baltag and Moss 2004, Baltag and Smets 2006; 2008, van Ditmarsch and Kooi 2008) for thorough motivations and explanations of the product update. Note that the event model's plausibilities take priority over those of the plausibility model (action-priority update).

Example 4. Consider Figure 2, where the event model $\mathcal{E}$ represents the biased nondeterministic coin toss of Example 2, $\mathcal{E}'$ shuffling the coin under a dice cup, and $\mathcal{E}''$ lifting the dice cup of Example 3. We indicate $\sim$ and $\leq$ with edges as in our illustrations of plausibility models. Further we use the convention of labelling basic events e by

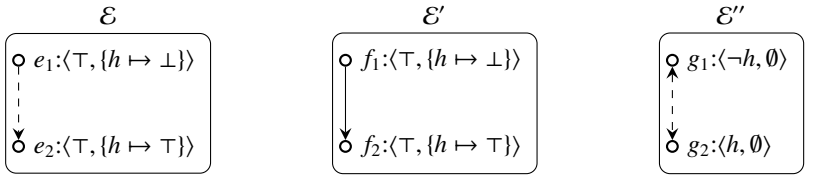


Figure 2: Three event models

$\langle pre(e), post(e)\rangle$. We write $post(e)$ on the form $\{p_1 \mapsto \varphi_1, \ldots, p_n \mapsto \varphi_n\}$, meaning that $post(e)(p_i) = \varphi_i$ for all i, and $post(e)(q) = q$ for $q \notin \{p_1, \ldots, p_n\}$.

Returning to Example 2 we see that $\mathcal{M} \otimes \mathcal{E} = \mathcal{M}''$ where $u_1 = (w, e_1), u_2 = (w, e_2)$. In $\mathcal{E}$ we have that $e_2 < e_1$, which encodes the bias of the coin, and $e_1 \not\sim e_2$ encoding the observability, which leads to u_1 and u_2 being distinguishable.

Regarding Example 3 we have that $\mathcal{M} \otimes \mathcal{E}' = \mathcal{M}'$ (modulo renaming). In contrast to $\mathcal{E}$, we have that $f_1 \sim f_2$, representing the inability to see the face of the coin due to the dice cup. For the sensing action $\mathcal{E}''$, we have $\mathcal{M} \otimes \mathcal{E}' \otimes \mathcal{E}'' = \mathcal{M}''$, illustrating how, when events are equiplausible ($g_1 \simeq g_2$), the plausibilities of $\mathcal{M}'$ carry over to $\mathcal{M}''$.

We have shown examples of how the interplay between plausibility model and event model can encode changes in belief, and further how to model both ontic change and sensing. In (Andersen and Bolander 2011) there is a more general treatment of action types, but here such a classification is not our objective. Instead we simply encode actions as required for our exposition and leave these considerations as future work.

Among the possible worlds, $\leq$ gives an ordering defining what is believed. Given a plausibility model $\mathcal{M} = (W, \sim, \leq, V)$, any non-empty subset of W will have one or more minimal worlds with respect to $\leq$, since $\leq$ is a well-preorder. For $S \subseteq W$, the set of $\leq$-minimal worlds, denoted $Min_{\leq}S$, is defined as:

$$Min_{\leq}S = \{s \in S \mid \forall s' \in S : s \leq s'\}.$$

The worlds in $Min_{\leq}S$ are called the *most plausible* worlds in S. The worlds of $Min_{\leq}D(\mathcal{M})$ are referred to as the *most plausible* of $\mathcal{M}$. With belief defined via minimal worlds (see the definition below), the agent has the same beliefs for any $w \in D(\mathcal{M})$. Analogous to most plausible worlds, an information cell $\mathcal{M}'$ of $\mathcal{M}$ is called *most plausible* if $D(\mathcal{M}') \cap Min_{\leq}D(\mathcal{M}) \neq \emptyset$ ($\mathcal{M}'$ contains at least one of the most plausible worlds of $\mathcal{M}$).

Definition 2.5 (Satisfaction Relation)**.** Let a plausibility model $\mathcal{M} = (W, \sim, \leq, V)$ on P be given. The satisfaction relation is given by, for all $w \in W$:

$$
\begin{array}{ll}
\mathcal{M}, w \models p & \text{iff } w \in V(p) \\
\mathcal{M}, w \models \neg\varphi & \text{iff } \textit{not } \mathcal{M}, w \models \varphi \\
\mathcal{M}, w \models \varphi \wedge \psi & \text{iff } \mathcal{M}, w \models \varphi \text{ and } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models K\varphi & \text{iff } \mathcal{M}, v \models \varphi \text{ for all } w \sim v \\
\mathcal{M}, w \models B^{\psi}\varphi & \text{iff } \mathcal{M}, v \models \varphi \text{ for all } v \in \mathit{Min}_{\leq}\{u \in W \mid \mathcal{M}, u \models \psi\} \\
\mathcal{M}, w \models X\varphi & \text{iff } \mathcal{M} \upharpoonright [w]_{\sim}, w \models \varphi \\
\mathcal{M}, w \models [\mathcal{E}, e]\,\varphi & \text{iff } \mathcal{M}, w \models \mathit{pre}(e) \text{ implies } \mathcal{M} \otimes \mathcal{E}, (w, e) \models \varphi
\end{array}
$$

where $\varphi, \psi \in L(P)$ and $(\mathcal{E}, e)$ is a pointed event model. We write $\mathcal{M} \models \varphi$ to mean $\mathcal{M}, w \models \varphi$ for all $w \in D(\mathcal{M})$. Satisfaction of the dynamic modality for non-pointed event models $\mathcal{E}$ is introduced by abbreviation, viz. $[\mathcal{E}]\,\varphi := \bigwedge_{e \in \mathcal{D}(\mathcal{E})} [\mathcal{E}, e]\,\varphi$. Furthermore, $\langle \mathcal{E} \rangle\,\varphi := \neg\,[\mathcal{E}]\,\neg\varphi$.[3]

The reader may notice that the semantic clause for $\mathcal{M}, w \models X\varphi$ is equivalent to the clause for $\mathcal{M}, w \models [\mathcal{E}, e]\varphi$ when $[\mathcal{E}, e]$ is a public announcement of a *characteristic formula* (van Benthem 1998) being true exactly at the worlds in $[w]_{\sim}$ (and any other world modally equivalent to one of these). In this sense, the X operator can be thought of as a public announcement operator, but a special one that always announces the current information cell. In the special case where $\mathcal{M}$ is an information cell, we have for all $w \in D(\mathcal{M})$ that $\mathcal{M}, w \models X\varphi$ iff $\mathcal{M}, w \models \varphi$.

3 Plausibility planning

The previous covered a framework for dealing with knowledge and belief in a dynamic setting. In the following, we will detail how a rational agent would adapt these concepts to model her own reasoning about how her actions affect the future. Specifically, we will show how an agent can predict whether or not a particular plan leads to a desired goal. This requires reasoning about the *conceivable* consequences of actions without *actually* performing them.

Two main concepts are required for our formulation of planning, both of which build on notions from the logic introduced in the previous section. One is that of states, a representation of the planning agent's view of the world at a particular time. Our states are plausibility models. The other concept is that of actions. These represent the agent's view of everything that can happen when she does something. Actions are event models, changing states into other states via product update.

[3] Hence, $\mathcal{M}, w \models \langle \mathcal{E} \rangle\,\varphi \Leftrightarrow \mathcal{M}, w \models \neg\,[\mathcal{E}]\,\neg\varphi \Leftrightarrow \mathcal{M}, w \models \neg(\bigwedge_{e \in \mathcal{D}(\mathcal{E})} [\mathcal{E}, e]\,\neg\varphi) \Leftrightarrow \mathcal{M}, w \models \bigvee_{e \in \mathcal{D}(\mathcal{E})} \neg\,[\mathcal{E}, e]\,\neg\varphi \Leftrightarrow \mathcal{M}, w \models \bigvee_{e \in D(\mathcal{E})} \langle \mathcal{E}, e \rangle\,\varphi$.

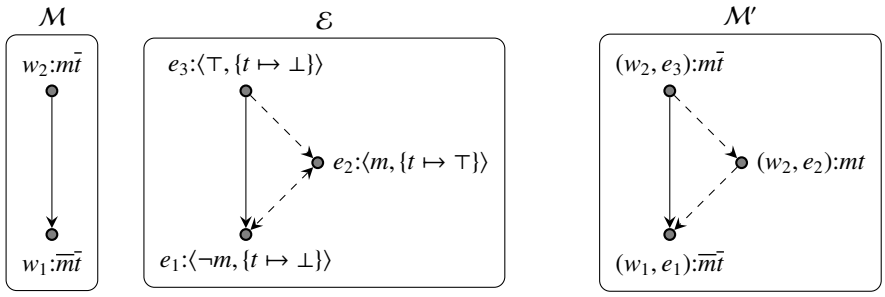


Figure 3: The situation before and after attempting to pay with a debit card, plus the event model depicting the attempt. This illustrates that the most plausible information cell can contain the least plausible world

In our case, the agent has knowledge and beliefs about the initial situation, knowledge and beliefs about actions, and therefore also knowledge and beliefs about the result of actions.

3.1 Reasoning about actions

Example 5 (Friday Beer)**.** Nearing the end of the month, an agent is going to have an end-of-week beer with her coworkers. Wanting to save the cash she has on hand for the bus fare, she would like to buy the beer using her debit card. Though she isn't certain, she believes that there's no money ($\overline{m}$) on the associated account. Figure 3 shows this initial situation as $\mathcal{M}$, where $\bar{t}$ signifies that the transaction hasn't been completed. In this small example her goal is to make t true.

When attempting to complete the transaction (using a normal debit card reader), a number of different things can happen, captured by $\mathcal{E}$ in Figure 3. If there is money on the account, the transaction will go through (e_2), and if there isn't, it won't (e_1). This is how the card reader operates most of the time and why e_1 and e_2 are the most plausible events. Less plausible, but still possible, is that the reader malfunctions for some other reason (e_3). The only feedback the agent will receive is whether the transaction was completed, not the reasons why it did or didn't ($e_1 \sim e_3 \not\sim e_2$). That the agent finds out whether the transaction was successful is why we do not collapse e_1 and e_2 to one event e' with $pre(e') = \top$ and $post(e')(t) = m$.

$\mathcal{M}\otimes\mathcal{E}$ expresses the agent's view on the possible outcomes of attempting the transaction. The model $\mathcal{M}'$ is the bisimulation contraction of $\mathcal{M} \otimes \mathcal{E}$, according to the

definition in Section 4.1 (the world (w_1, e_3) having been removed, as it is bisimilar to (w_1, e_1)).

$\mathcal{M}'$ consists of two information cells, corresponding to whether or not the transaction was successful. What she believes will happen is given by the global plausibility relation. When actually attempting the transaction the result will be one of the information cells of $\mathcal{M}'$, namely $\mathcal{M}_{\bar{t}} = \mathcal{M}' \restriction \{(w_1, e_1), (w_2, e_3)\}$ or $\mathcal{M}_t = \mathcal{M}' \restriction \{(w_2, e_2)\}$, in which she will know $\neg t$ and t respectively. As (w_1, e_1) is the most plausible, we can say that she *expects* to end up in (w_1, e_1), and, by extension, in the information cell $\mathcal{M}_{\bar{t}}$: She expects to end up in a situation where she knows $\neg t$, but is ignorant concerning m. If, unexpectedly, the transaction *is* successful, she will know that the balance is sufficient (m). The most plausible information cell(s) in a model are those the agent expects. That (w_2, e_3) is in the expected information cell, when the globally more plausible world (w_2, e_2) is not, might seem odd. It isn't. The partitioning of $\mathcal{M}$ into the information cells $\mathcal{M}_{\bar{t}}$ and $\mathcal{M}_t$ suggests that she will sense the value of t ($\neg t$ holds everywhere in the former, t everywhere in the latter). As she expects to find out that t does not to hold, she expects to be able to rule out all the worlds in which t *does* hold. Therefore, she expects to be able to rule out (w_2, e_2) and *not* (w_2, e_3) (or w_1, e_1). This gives $\mathcal{M}' \models BX(K\neg t \wedge B\neg m \wedge \widehat{K}m)$: She expects to come to know that the transaction has failed and that she will believe there's no money on the account (though she does consider it possible that there is).

Under the definition of planning that is to follow in Section 3.2, an agent has a number of actions available to construct plans. She needs a notion of which actions can be considered at different stages of the planning process. As in the planning literature, we call this notion *applicability*.

Definition 3.1 (Applicability)**.** An event model $\mathcal{E}$ is said to be *applicable* in a plausibility model $\mathcal{M}$ if $\mathcal{M} \models \langle\mathcal{E}\rangle \top$.

Unfolding the definition of $\langle\mathcal{E}\rangle$, we see what applicability means:

$$
\begin{aligned}
&\mathcal{M} \models \langle\mathcal{E}\rangle \top \Leftrightarrow \forall w \in D(\mathcal{M}) : \mathcal{M}, w \models \langle\mathcal{E}\rangle \top \Leftrightarrow \\
&\forall w \in D(\mathcal{M}) : \mathcal{M}, w \models \vee_{e \in D(\mathcal{E})} \langle\mathcal{E}, e\rangle \top \Leftrightarrow \\
&\forall w \in D(\mathcal{M}), \exists e \in D(\mathcal{E}) : \mathcal{M}, w \models \langle\mathcal{E}, e\rangle \top \Leftrightarrow \\
&\forall w \in D(\mathcal{M}), \exists e \in D(\mathcal{E}) : \mathcal{M}, w \models pre(e) \text{ and } \mathcal{M} \otimes \mathcal{E}, (w, e) \models \top \Leftrightarrow \\
&\forall w \in D(\mathcal{M}), \exists e \in D(\mathcal{E}) : \mathcal{M}, w \models pre(e).
\end{aligned}
$$

This says that no matter which is the actual world (it must be one of those considered possible), the action defines an outcome. This concept of applicability is equivalent to

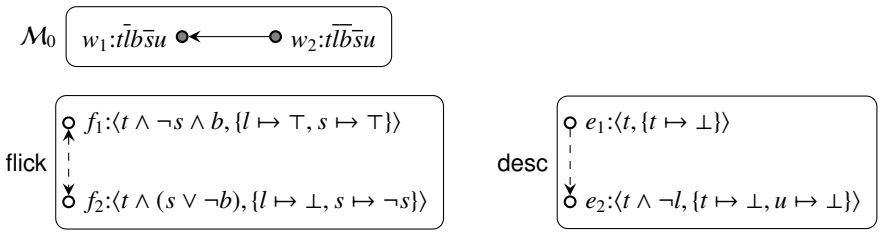


Figure 4: An information cell, $\mathcal{M}_0$, and two event models, flick and desc

the one in (Andersen and Bolander 2011). The discussion in (Section 6.6 of de Lima 2007) also notes this aspect, insisting that actions must be *meaningful*. The same sentiment is expressed by our notion of applicability.

Proposition 1. *Given a plausibility model $\mathcal{M}$ and an applicable event model $\mathcal{E}$, we have $D(\mathcal{M} \otimes \mathcal{E}) \neq \emptyset$.*

The product update $\mathcal{M} \otimes \mathcal{E}$ expresses the outcome(s) of doing $\mathcal{E}$ in the situation $\mathcal{M}$, in the planning literature called *applying* $\mathcal{E}$ in $\mathcal{M}$. The dynamic modality $[\mathcal{E}]$ expresses reasoning about what holds after applying $\mathcal{E}$.

Lemma 1. *Let $\mathcal{M}$ be a plausibility model and $\mathcal{E}$ an event model. Then $\mathcal{M} \models [\mathcal{E}]\varphi$ iff $\mathcal{M} \otimes \mathcal{E} \models \varphi$.*

Proof. $\mathcal{M} \models [\mathcal{E}]\varphi \Leftrightarrow \forall w \in \mathcal{D}(\mathcal{M}) : \mathcal{M}, w \models [\mathcal{E}]\varphi \Leftrightarrow$
$\forall w \in \mathcal{D}(\mathcal{M}) : \mathcal{M}, w \models \bigwedge_{e \in \mathcal{D}(\mathcal{E})} [\mathcal{E}, e]\varphi \Leftrightarrow$
$\forall (w, e) \in \mathcal{D}(\mathcal{M}) \times \mathcal{D}(\mathcal{E}) : \mathcal{M}, w \models [\mathcal{E}, e]\varphi \Leftrightarrow$
$\forall (w, e) \in \mathcal{D}(\mathcal{M}) \times \mathcal{D}(\mathcal{E}) : \mathcal{M}, w \models pre(e)$ implies $\mathcal{M} \otimes \mathcal{E}, (w, e) \models \varphi \Leftrightarrow$
$\forall (w, e) \in \mathcal{D}(\mathcal{M} \otimes \mathcal{E}) : \mathcal{M} \otimes \mathcal{E}, (w, e) \models \varphi \Leftrightarrow \mathcal{M} \otimes \mathcal{E} \models \varphi$. □

Here we are looking at *global satisfaction*, by evaluating $[\mathcal{E}]\varphi$ in all of $\mathcal{M}$, rather than a specific world. The reason is that evaluation in planning must happen from the perspective of the planning agent and its "information state". Though one of the worlds of $\mathcal{M}$ is the actual world, the planning agent is ignorant about which it is. Whatever plan it comes up with, it must work in all of the worlds which are indistinguishable to the agent, that is, in the entire model. A similar point, and a similar solution, is found in (Jamroga and Ågotnes 2007).

Example 6. We now return to the agent from Example 1. Her view of the initial situation ($\mathcal{M}_0$) and her available actions (flick and desc) are seen in Figure 4. The propositional letters mean t: "top of stairs", l: "light on", b: "bulb working", s : "switch

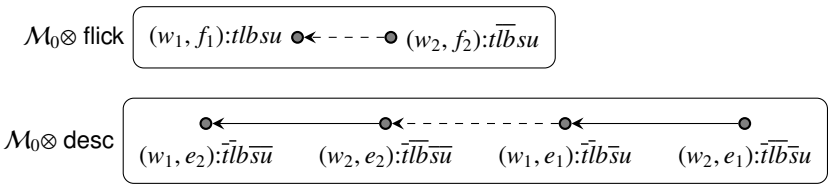


Figure 5: The models resulting from applying the actions flick and desc in $\mathcal{M}_0$. Reflexive edges are not shown and the transitive closure is left implicit.

on" and u: "unharmed". Initially, in $\mathcal{M}_0$, she believes that the bulb is working, and knows that she is at the top of the stairs, unharmed and that the switch and light is off: $\mathcal{M}_0 \models Bb \wedge K(t \wedge u \wedge \neg l \wedge \neg s)$.

flick and desc represent flicking the light switch and trying to descend the stairs, respectively. Both require being at the top of the stairs (t). f_1 of flick expresses that if the bulb is working, turning on the switch will turn on the light, and f_2 that if the bulb is broken or the switch is currently on, the light will be off. The events are epistemically distinguishable, as the agent will be able to tell whether the light is on or off. desc describes descending the stairs, with or without the light on. e_1 covers the agent descending the stairs unharmed, and can happen regardless of there being light or not. The more plausible event e_2 represents the agent stumbling, though this can only happen in the dark. If the light is on, she will descend safely. Definition 3.1 and Lemma 1 let us express the action sequences possible in this scenario.

- $\mathcal{M}_0 \models \langle \mathsf{flick} \rangle \top \wedge \langle \mathsf{desc} \rangle \top$. The agent can initially do either flick or desc.
- $\mathcal{M}_0 \models [\mathsf{flick}] \langle \mathsf{desc} \rangle \top$. After doing flick, she can do desc.
- $\mathcal{M}_0 \models [\mathsf{desc}] (\neg \langle \mathsf{flick} \rangle \top \wedge \neg \langle \mathsf{desc} \rangle \top)$. Nothing can be done after desc.

Figure 5 shows the plausibility models arising from doing flick and desc in $\mathcal{M}_0$. Via Lemma 1 she can now conclude:

- $\mathcal{M}_0 \models [\mathsf{flick}] (Kb \vee K\neg b)$: Flicking the light switch gives knowledge of whether the bulb works or not.
- $\mathcal{M}_0 \models [\mathsf{flick}] BKb$. She expects to come to know that it works.
- $\mathcal{M}_0 \models [\mathsf{desc}] (K\neg t \wedge B\neg u)$. Descending the stairs in the dark will definitely get her to the bottom, though she believes she will end up hurting herself.

3.2 Planning

We now turn to formalising planning and then proceed to answer two questions of particular interest: How do we verify that a given plan achieves a goal? And can we compute such plans? This section deals with the first question, plan verification, while the second, plan synthesis, is detailed in Section 4.

Definition 3.2 (Plan Language)**.** Given a finite set A of event models on $L(P)$, the *plan language* $\mathcal{L}(P, \mathsf{A})$ is given by:

$$\pi ::= \mathcal{E} \mid \mathsf{skip} \mid \mathsf{if}\ \varphi\ \mathsf{then}\ \pi\ \mathsf{else}\ \pi \mid \pi;\pi$$

where $\mathcal{E} \in \mathsf{A}$ and $\varphi \in L(P)$. We name members π of this language *plans*, and use $\mathsf{if}\ \varphi\ \mathsf{then}\ \pi$ as shorthand for $\mathsf{if}\ \varphi\ \mathsf{then}\ \pi\ \mathsf{else}\ \mathsf{skip}$.

The reading of the plan constructs are "do $\mathcal{E}$", "do nothing", "if φ then π, else π'", and "first π then π'" respectively. In the translations provided in Definition 3.3, the condition of the if-then-else construct becomes a K-formula, ensuring that branching depends only on worlds which are distinguishable to the agent. The idea is similar to the *meaningful plans* of (de Lima 2007), where branching is allowed on *epistemically interpretable formulas* only.

Definition 3.3 (Translation)**.** Let α be one of s, w, sp or wp. We define an *α-translation* as a function $[\cdot]_\alpha : \mathcal{L}(P, \mathsf{A}) \to (L(P) \to L(P))$:

$$[\mathcal{E}]_\alpha \varphi := \langle \mathcal{E} \rangle \top \wedge \begin{cases} [\mathcal{E}]\, XK\varphi & \text{if } \alpha = s \\ \widehat{K}\, \langle \mathcal{E} \rangle\, XK\varphi & \text{if } \alpha = w \\ [\mathcal{E}]\, BXK\varphi & \text{if } \alpha = sp \\ [\mathcal{E}]\, \widehat{B}XK\varphi & \text{if } \alpha = wp \end{cases}$$

$$[\mathsf{skip}]_\alpha \varphi := \varphi$$

$$[\mathsf{if}\ \varphi'\ \mathsf{then}\ \pi\ \mathsf{else}\ \pi']_\alpha \varphi := (K\varphi' \to [\pi]_\alpha \varphi) \wedge (\neg K\varphi' \to [\pi']_\alpha \varphi)$$

$$[\pi;\pi']_\alpha \varphi := [\pi]_\alpha ([\pi']_\alpha \varphi)$$

We call $[\cdot]_s$ the *strong translation*, $[\cdot]_w$ the *weak translation*, $[\cdot]_{sp}$ the *strong plausibility* translation and $[\cdot]_{wp}$ the *weak plausibility* translation.

The translations are constructed specifically to make the following lemma hold, providing a semantic interpretation of plans (leaving out skip and $\pi_1;\pi_2$).

Lemma 2. *Let $\mathcal{M}$ be an information cell, $\mathcal{E}$ an event model and φ a formula of $L(P)$. Then:*

(1) $\mathcal{M} \models [\mathcal{E}]_s\varphi$ iff $\mathcal{M} \models \langle\mathcal{E}\rangle\top$ *and for each information cell* $\mathcal{M}'$ *of* $\mathcal{M}\otimes\mathcal{E}$: $\mathcal{M}' \models \varphi$.

(2) $\mathcal{M} \models [\mathcal{E}]_w\varphi$ iff $\mathcal{M} \models \langle\mathcal{E}\rangle\top$ *and for some information cell* $\mathcal{M}'$ *of* $\mathcal{M}\otimes\mathcal{E}$: $\mathcal{M}' \models \varphi$.

(3) $\mathcal{M} \models [\mathcal{E}]_{sp}\varphi$ iff $\mathcal{M} \models \langle\mathcal{E}\rangle\top$ *and for each most plausible information cell* $\mathcal{M}'$ *of* $\mathcal{M}\otimes\mathcal{E}$: $\mathcal{M}' \models \varphi$.

(4) $\mathcal{M} \models [\mathcal{E}]_{wp}\varphi$ iff $\mathcal{M} \models \langle\mathcal{E}\rangle\top$ *and for some most plausible information cell* $\mathcal{M}'$ *of* $\mathcal{M}\otimes\mathcal{E}$: $\mathcal{M}' \models \varphi$.

(5) $\mathcal{M} \models [\text{if } \varphi' \text{ then } \pi \text{ else } \pi']_\alpha\varphi$ iff
($\mathcal{M} \models \varphi'$ *implies* $\mathcal{M} \models [\pi]_\alpha\varphi$) *and* ($\mathcal{M} \not\models \varphi'$ *implies* $\mathcal{M} \models [\pi']_\alpha\varphi$).

Proof. We only prove 4 and 5, as 1–4 are very similar. For 4 we have:

$$\begin{aligned}
&\mathcal{M} \models [\mathcal{E}]_{wp}\,\varphi \;\Leftrightarrow\; \mathcal{M} \models \langle\mathcal{E}\rangle\top \wedge [\mathcal{E}]\,\widehat{B}XK\varphi \;\Leftrightarrow^{\text{Lemma 1}}\\
&\mathcal{M} \models \langle\mathcal{E}\rangle\top \text{ and } \mathcal{M}\otimes\mathcal{E} \models \widehat{B}XK\varphi \;\Leftrightarrow\\
&\mathcal{M} \models \langle\mathcal{E}\rangle\top \text{ and } \forall(w,e) \in D(\mathcal{M}\otimes\mathcal{E}) : \mathcal{M}\otimes\mathcal{E}, (w,e) \models \widehat{B}XK\varphi \;\Leftrightarrow^{\text{Prop. 1}}\\
&\mathcal{M} \models \langle\mathcal{E}\rangle\top \text{ and } \exists(w,e) \in Min_{\leq}D(\mathcal{M}\otimes\mathcal{E}) : \mathcal{M}\otimes\mathcal{E}, (w,e) \models XK\varphi \;\Leftrightarrow\\
&\mathcal{M} \models \langle\mathcal{E}\rangle\top \text{ and } \exists(w,e) \in Min_{\leq}D(\mathcal{M}\otimes\mathcal{E}) : \mathcal{M}\otimes\mathcal{E} \upharpoonright [(w,e)]_\sim, (w,e) \models K\varphi \;\Leftrightarrow\\
&\mathcal{M} \models \langle\mathcal{E}\rangle\top \text{ and } \exists(w,e) \in Min_{\leq}D(\mathcal{M}\otimes\mathcal{E}) : \mathcal{M}\otimes\mathcal{E} \upharpoonright [(w,e)]_\sim \models \varphi \;\Leftrightarrow\\
&\mathcal{M} \models \langle\mathcal{E}\rangle\top \text{ and in some most plausible information cell } \mathcal{M}' \text{ of } \mathcal{M}\otimes\mathcal{E},\ \mathcal{M}' \models \varphi.
\end{aligned}$$

For if-then-else, first note that:

$$\begin{aligned}
&\mathcal{M} \models \neg K\varphi' \to [\pi]_\alpha\varphi \;\Leftrightarrow\; \forall w \in D(\mathcal{M})\colon \mathcal{M}, w \models \neg K\varphi' \to [\pi]_\alpha\varphi \;\Leftrightarrow\\
&\forall w \in D(\mathcal{M})\colon \mathcal{M}, w \models \neg K\varphi' \text{ implies } \mathcal{M}, w \models [\pi]_\alpha\varphi \;\Leftrightarrow^{\mathcal{M} \text{ is an info. cell}}\\
&\forall w \in D(\mathcal{M})\colon \text{if } \mathcal{M}, v \models \neg\varphi' \text{ for some } v \in D(\mathcal{M}) \text{ then } \mathcal{M}, w \models [\pi]_\alpha\varphi \;\Leftrightarrow\\
&\text{if } \mathcal{M}, v \models \neg\varphi' \text{ for some } v \in D(\mathcal{M}) \text{ then } \forall w \in D(\mathcal{M})\colon \mathcal{M}, w \models [\pi]_\alpha\varphi \;\Leftrightarrow\\
&\mathcal{M} \not\models \varphi' \text{ implies } \mathcal{M} \models [\pi']_\alpha\varphi.
\end{aligned}$$

Similarly, we can prove:

$$\mathcal{M} \models K\varphi' \to [\pi]_\alpha\varphi \;\Leftrightarrow\; \mathcal{M} \models K\varphi' \text{ implies } \mathcal{M} \models [\pi']_\alpha\varphi.$$

Using these facts, we get:

$$\mathcal{M} \models [\text{if } \varphi' \text{ then } \pi \text{ else } \pi']_\alpha\varphi \;\Leftrightarrow\; \mathcal{M} \models (K\varphi' \to [\pi]_\alpha\varphi) \wedge (\neg K\varphi' \to [\pi']_\alpha\varphi) \;\Leftrightarrow$$

$\mathcal{M} \models K\varphi' \rightarrow [\pi]_\alpha \varphi$ and $\mathcal{M} \models \neg K\varphi' \rightarrow [\pi']_\alpha \varphi \Leftrightarrow$
($\mathcal{M} \models \varphi'$ implies $\mathcal{M} \models [\pi]_\alpha \varphi$) and ($\mathcal{M} \not\models \varphi'$ implies $\mathcal{M} \models [\pi']_\alpha \varphi$).

□

Using XK (as is done in all translations) means that reasoning after an action is relative to a particular information cell (as $\mathcal{M}, w \models XK\varphi \Leftrightarrow \mathcal{M} \upharpoonright [w]_\sim, w \models K\varphi \Leftrightarrow \mathcal{M} \upharpoonright [w]_\sim \models \varphi$).

Definition 3.4 (Planning Problems and Solutions)**.** Let P be a finite set of propositional symbols. A planning problem on P is a triple $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$ where

- $\mathcal{M}_0$ is a finite information cell on P called the *initial state*.
- A is a finite set of event models on $L(P)$ called the *action library*.
- $\varphi_g \in L(P)$ is the *goal (formula)*.

A plan $\pi \in \mathcal{L}(P, \mathsf{A})$ is an α*-solution* to $\mathcal{P}$ if $\mathcal{M}_0 \models [\pi]_\alpha \varphi_g$. For a specific choice of $\alpha = s/w/sp/wp$, we will call π a *strong/weak/strong plausibility/weak plausibility*-solution respectively.

Given a π, we wish to check whether π is an α-solution (for some particular α) to $\mathcal{P}$. This can be done via model checking the dynamic formula given by the translation $[\pi]_\alpha \varphi_g$ in the initial state of $\mathcal{P}$.

A strong solution π is one that guarantees that φ_g will hold after executing it ("π achieves φ_g"). If π is a weak solution, it achieves φ_g for at least one particular sequence of outcomes. Strong and weak plausibility-solutions are as strong- and weak-solutions, except that they need only achieve φ_g for *all of/some of* the most plausible outcomes.

Example 7. The basement scenario (Example 1) can be formalised as the planning problem $\mathcal{P}_B = (\mathcal{M}_0, \{\mathsf{flick}, \mathsf{desc}\}, \varphi_g)$ with $\mathcal{M}_0$, flick and desc being defined in Figure 4 and $\varphi_g = \neg t \wedge u$. Let $\pi_1 = \mathsf{desc}$. We then have that:

$$\mathcal{M}_0 \models [\mathsf{desc}]_w (\neg t \wedge u) \Leftrightarrow \mathcal{M}_0 \models \langle \mathsf{desc} \rangle \top \wedge \widehat{K} \langle \mathsf{desc} \rangle XK(\neg t \wedge u) \Leftrightarrow^{\mathsf{desc}\text{ is applicable}}$$
$$\mathcal{M}_0 \models \widehat{K} \langle \mathsf{desc} \rangle XK(\neg t \wedge u) \Leftrightarrow \exists w \in D(\mathcal{M}_0) : \mathcal{M}_0, w \models \langle \mathsf{desc} \rangle XK(\neg t \wedge u).$$

Picking w_1, we have

$$\mathcal{M}_0, w_1 \models \langle \mathsf{desc} \rangle XK(\neg t \wedge u) \Leftrightarrow \mathcal{M}_0 \otimes \mathsf{desc}, (w_1, e_1) \models XK(\neg t \wedge u) \Leftrightarrow$$
$$\mathcal{M}_0 \otimes \mathsf{desc} \upharpoonright [(w_1, e_1)]_\sim \models (\neg t \wedge u)$$

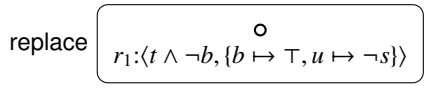


Figure 6: Event model for replacing a broken bulb

which holds as seen in Figure 5. Thus, π_1 is a weak solution. Further, Lemma 2 tells us that π_1 is not a $s/wp/sp$ solution, as u does not hold in the (most plausible) information cell $\mathcal{M} \otimes \mathsf{desc} \upharpoonright \{(w_1, e_2), (w_2, e_2)\}$.

The plan $\pi_2 = \mathsf{flick}; \mathsf{desc}$ is a strong plausibility solution, as can be verified by $\mathcal{M}_0 \models [\pi_2]_{sp}(\neg t \wedge u)$. Without an action for replacing the lightbulb, there are no strong solutions. Let replace be the action in Figure 6, where $post(r_1)(u) = \neg s$ signifies that if the power is on, the agent will hurt herself, and define a new problem $\mathcal{P}'_B = \{\mathcal{M}_0, \{\mathsf{flick}, \mathsf{desc}, \mathsf{replace}\}, \varphi_g\}$. Then $\pi_3 = \mathsf{flick}$; (if $\neg l$ then flick ; replace ; flick) ; desc is a strong solution (we leave verification to the reader): If the light comes on after flicking the switch (as expected) she can safely walk down the stairs. If it does not, she turns off the power, replaces the broken bulb, turns the power on again (this time knowing that the light will come on), and then proceeds as before.

Besides being an sp-solution, π_2 is also a w- and a wp-solution, indicating a hierarchy of strengths of solutions. This should come as no surprise, given both the formal and intuitive meaning of planning and actions presented so far. In fact, this hierarchy exists for any planning problem, as shown by the following result which is a consequence of Lemma 2 (stated without proof).

Lemma 3. *Let $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$ be a planning problem. Then:*

- *Any strong solution to $\mathcal{P}$ is also a strong plausibility solution:* $\mathcal{M}_0 \models [\pi]_s \varphi_g \Rightarrow \mathcal{M}_0 \models [\pi]_{sp} \varphi_g$.
- *Any strong plausibility solution to $\mathcal{P}$ is also a weak plausibility solution:* $\mathcal{M}_0 \models [\pi]_{sp} \varphi_g \Rightarrow \mathcal{M}_0 \models [\pi]_{wp} \varphi_g$.
- *Any weak plausibility solution to $\mathcal{P}$ is also a weak solution:* $\mathcal{M}_0 \models [\pi]_{wp} \varphi_g \Rightarrow \mathcal{M}_0 \models [\pi]_w \varphi_g$.

4 Plan synthesis

In this section we show how to synthesise conditional plans for solving planning problems. Before we can give the concrete algorithms, we establish some technical results which are stepping stones to proving termination of our planning algorithm, and hence decidability of plan existence in our framework.

4.1 Bisimulations, contractions and modal equivalence

We now define bisimulations on plausibility models. For our purpose it is sufficient to define bisimulations on $\sim$-connected models, that is, on information cells.[4] First we define a *normal plausibility relation* which will form the basis of our bisimulation definition.

Definition 4.1 (Normality)**.** Given is an information cell $\mathcal{M} = (W, \sim, \leq, V)$ on P. By slight abuse of language, two worlds $w, w' \in W$ are said to *have the same valuation* if for all $p \in P$: $w \in V(p) \Leftrightarrow w' \in V(p)$. Define an equivalence relation on W: $w \approx w'$ iff w and w' has the same valuation. Now define $w \preceq w'$ iff $Min_{\leq}([w]_{\approx}) \leq Min_{\leq}([w']_{\approx})$. This defines the *normal plausibility relation*. $\mathcal{M}$ is called *normal* if $\preceq \; = \; \leq$. The *normalisation* of $\mathcal{M} = (W, \sim, \leq, V)$ is $\mathcal{M}' = (W, \sim, \preceq, V)$.

Definition 4.2 (Bisimulation)**.** Let $\mathcal{M} = (W, \sim, \leq, V)$ and $\mathcal{M}' = (W', \sim', \leq', V')$ be information cells on P. A non-empty relation $\mathcal{R} \subseteq W \times W'$ is a *bisimulation* between $\mathcal{M}$ and $\mathcal{M}'$ (and $\mathcal{M}, \mathcal{M}'$ are called *bisimilar*) if for all $(w, w') \in \mathcal{R}$:

[atom] For all $p \in P$: $w \in V(p)$ iff $w' \in V'(p)$.

[forth] If $v \in W$ and $v \leq w$ then there is a $v' \in W'$ s.t. $v' \leq' w'$ and $(v, v') \in \mathcal{R}$.

[back] If $v' \in W'$ and $v' \leq w'$ then there is a $v \in W$ s.t. $v \leq w$ and $(v, v') \in \mathcal{R}$.

If $\mathcal{R}$ has domain W and codomain W', it is called *total*. If $\mathcal{M} = \mathcal{M}'$, it is called an *autobisimulation* (on $\mathcal{M}$). Worlds w and w' of an information cell $\mathcal{M} = (W, \sim, \leq, V)$ are called *bisimilar* if there exists an autobisimulation $\mathcal{R}$ on $\mathcal{M}$ with $(w, w') \in \mathcal{R}$.

We are here only interested in total bisimulations, so, unless otherwise stated, we assume this in the following. Note that our definition of bisimulation immediately implies that there exists a (total) bisimulation between any information cell and its normalisation. Note also that for normal models, the bisimulation definition becomes the standard modal logic one.[5]

Lemma 4. *If two worlds of an information cell have the same valuation they are bisimilar.*

[4] The proper notion of bisimulation for plausibility structures is explored in more detail by Andersen, Bolander, van Ditmarsch and Jensen in ongoing research. A similar notion for slightly different types of plausibility structures is given in (van Ditmarsch 2013, to appear). Surprisingly, Demey does not consider our notion of bisimulation in his thorough survey (Demey 2011) on different notions of bisimulation for plausibility structures.

[5] We didn't include a condition for the epistemic relation, $\sim$, in [back] and [forth], simply because we are here only concerned with $\sim$-connected models.

Proof. Assume worlds w and w' of an information cell $\mathcal{M} = (W, \sim, \leq, V)$ have the same valuation. Let $\mathcal{R}$ be the relation that relates each world of $\mathcal{M}$ to itself and additionally relates w to w'. We want to show that $\mathcal{R}$ is a bisimulation. This amounts to showing [atom], [forth] and [back] for the pair $(w, w') \in \mathcal{R}$. [atom] holds trivially since $w \approx w'$. For [forth], assume $v \in W$ and $v \leq w$. We need to find a $v' \in W$ s.t. $v' \leq w'$ and $(v, v') \in \mathcal{R}$. Letting $v' = v$, it suffices to prove $v \leq w'$. Since $w \approx w'$ this is immediate: $v \leq w \Leftrightarrow Min_{\leq}([v]_{\approx}) \leq Min_{\leq}([w]_{\approx}) \overset{w \approx w'}{\Leftrightarrow} Min_{\leq}([v]_{\approx}) \leq Min_{\leq}([w']_{\approx}) \Leftrightarrow v \leq w'$. [back] is proved similarly. □

Unions of autobisimulations are autobisimulations. We can then in the standard way define the *(bisimulation) contraction* of a normal information cell as its quotient with respect to the union of all autobisimulations (Blackburn and van Benthem 2006).[6] The contraction of a non-normal model is taken to be the contraction of its normalisation. In a contracted model, no two worlds are bisimilar, by construction. Hence, by Lemma 4, no two worlds have the same valuation. Thus, the contraction of an information cell on a finite set of proposition symbols P contains at most $2^{|P|}$ worlds. Since any information cell is bisimilar to its contraction (Blackburn and van Benthem 2006), this shows that there can only exist finitely many non-bisimilar information cells on any given finite set P.

Two information cells $\mathcal{M}$ and $\mathcal{M}'$ are called *modally equivalent*, written $\mathcal{M} \equiv \mathcal{M}'$, if for all formulas φ in $L(P)$: $\mathcal{M} \models \varphi \Leftrightarrow \mathcal{M}' \models \varphi$. Otherwise, they are called *modally inequivalent*. We now have the following standard result (the result is standard for standard modal languages and bisimulations, but it is not trivial that it also holds here).

Theorem 1. *If two information cells are (totally) bisimilar they are modally equivalent.*

Proof. We need to show that if $\mathcal{R}$ is a total bisimulation between information cells $\mathcal{M}$ and $\mathcal{M}'$, then for all formulas φ of $L(P)$: $\mathcal{M} \models \varphi \Leftrightarrow \mathcal{M}' \models \varphi$. First we show that we only have to consider formulas φ of the static sublanguage of $L(P)$, that is, the language without the $[\mathcal{E}, e]$ modalities. In (Baltag and Smets 2006), reduction axioms from the dynamic to the static language are given for a language similar to $L(P)$. The differences in language are our addition of postconditions and the fact that our belief modality is defined from the global plausibility relation rather than being localised to epistemic equivalence classes. The latter difference is irrelevant when only considering information cells as we do here. The former difference of course means that the reduction axioms presented in (Baltag and Smets 2006) will not suffice for

[6] More precisely, let $\mathcal{M}$ be a normal information cell and let $\mathcal{R}$ be the union of all autobisimulations on $\mathcal{M}$. Then the contraction $\mathcal{M}' = (W', \sim', \leq', V')$ of $\mathcal{M}$ has as worlds the equivalence classes $[w]_{\mathcal{R}} = \{w' \mid (w, w') \in \mathcal{R}\}$ and has $[w]_{\mathcal{R}} \leq' [w']_{\mathcal{R}}$ iff $v \leq v'$ for some $v \in [w]_{\mathcal{R}}$ and $v' \in [w']_{\mathcal{R}}$.

our purpose. van Ditmarsch and Kooi (2008) shows that adding postconditions to the language without the doxastic modalities only requires changing the reduction axiom for $[\mathcal{E}, e]\, p$, where p is a propositional symbol. Thus, if we take the reduction axioms of Baltag and Smets (2006) and replace the reduction axiom for $[\mathcal{E}, e]\, p$ by the one in (van Ditmarsch and Kooi 2008), we get reduction axioms for our framework. We leave out the details.

We now need to show that if $\mathcal{R}$ is a total bisimulation between information cells $\mathcal{M}$ and $\mathcal{M}'$, then for all $[\mathcal{E}, e]$-free formulas φ of $L(P)$: $\mathcal{M} \models \varphi \Leftrightarrow \mathcal{M}' \models \varphi$. Since $\mathcal{R}$ is total, it is sufficient to prove that for all $[\mathcal{E}, e]$-free formulas φ of $L(P)$ and all $(w, w') \in \mathcal{R}$: $\mathcal{M}, w \models \varphi \Leftrightarrow \mathcal{M}', w' \models \varphi$. The proof is by induction on φ. In the induction step we are going to need the induction hypothesis for several different choices of $\mathcal{R}, w$ and w', so what we will actually prove by induction on φ is this: For all formulas φ of $L(P)$, if $\mathcal{R}$ is a total bisimulation between information cells $\mathcal{M}$ and $\mathcal{M}'$ on P and $(w, w') \in \mathcal{R}$, then $\mathcal{M}, w \models \varphi \Leftrightarrow \mathcal{M}', w' \models \varphi$.

The base case is when φ is propositional. Then the required follows immediately from [atom], using that $(w, w') \in \mathcal{R}$. For the induction step, we have the following cases of φ: $\neg\psi, \psi \wedge \gamma, X\psi, K\psi, B^{\gamma}\psi$. The first two cases are trivial. So is $X\psi$, as $X\psi \leftrightarrow \psi$ holds on any information cell. For $K\psi$ we reason as follows. Let $\mathcal{R}$ be a total bisimulation between information cells $\mathcal{M}$ and $\mathcal{M}'$ with $(w, w') \in \mathcal{R}$. Using that $\mathcal{R}$ is total and that $\mathcal{M}$ and $\mathcal{M}'$ are both $\sim$-connected we get: $\mathcal{M}, w \models K\psi \Leftrightarrow \forall v \in W$: $\mathcal{M}, v \models \psi \overset{\text{i.h.}}{\Leftrightarrow} \forall v' \in W'$: $\mathcal{M}', v \models \psi \Leftrightarrow \mathcal{M}', w' \models K\psi$.

The case of $B^{\gamma}\psi$ is more involved. Let $\mathcal{M}, \mathcal{M}', \mathcal{R}, w$ and w' be as above. By symmetry, it suffices to prove $\mathcal{M}, w \models B^{\gamma}\psi \Rightarrow \mathcal{M}', w' \models B^{\gamma}\psi$. So assume $\mathcal{M}, w \models B^{\gamma}\psi$, that is, $\mathcal{M}, v \models \psi$ for all $v \in Min_{\leq}\{u \in W \mid \mathcal{M}, u \models \gamma\}$. We need to prove $\mathcal{M}', v' \models \psi$ for all $v' \in Min_{\leq'}\{u' \in W' \mid \mathcal{M}', u' \models \gamma\}$. So let $v' \in Min_{\leq'}\{u' \in W' \mid \mathcal{M}', u' \models \gamma\}$. By definition of $Min_{\leq'}$ this means that:

$$\text{for all } u' \in W', \text{ if } \mathcal{M}', u' \models \gamma \text{ then } v' \leq' u'. \tag{1}$$

Choose an $x \in Min_{\leq}\{u \in W \mid u \approx u' \text{ and } (u', v') \in \mathcal{R}\}$. We want to use (1) to show that the following holds:

$$\text{for all } u \in W, \text{ if } \mathcal{M}, u \models \gamma \text{ then } x \leq u. \tag{2}$$

To prove (2), let $u \in W$ with $\mathcal{M}, u \models \gamma$. Choose u' with $(u, u') \in \mathcal{R}$. The induction hypothesis implies $\mathcal{M}', u' \models \gamma$. We now prove that $v' \leq' Min_{\leq'}([u']_{\approx})$. To this end, let $u'' \in [u']_{\approx}$. We need to prove $v' \leq' u''$. Since $u'' \approx u'$, Lemma 4 implies that

u' and u'' are bisimilar. By induction hypothesis we then get $\mathcal{M}', u'' \models \gamma$.[7] Using (1) we now get $v' \leq' u''$, as required. This show $v' \leq' Min_{\leq'}([u']_{\approx})$. We now have $Min_{\leq'}([v']_{\approx}) \leq' v' \leq' Min_{\leq'}([u']_{\approx})$, and hence $v' \preceq u'$. By [back] there is then a v s.t. $(v, v') \in \mathcal{R}$ and $v \preceq u$. By choice of x, $x \leq Min_{\leq}([v]_{\approx})$. Using $v \preceq u$, we now finally get: $x \leq Min_{\leq}([v]_{\approx}) \leq Min_{\leq}([u]_{\approx}) \leq u$. This shows that (2) holds.

From (2) we can now conclude $x \in Min_{\leq}\{u \in W \mid \mathcal{M}, u \models \gamma\}$ and hence, by original assumption, $\mathcal{M}, x \models \psi$. By choice of x there is an $x' \approx x$ with $(x', v') \in \mathcal{R}$. Since $\mathcal{M}, x \models \psi$ and $x' \approx x$, we can again use Lemma 4 and the induction hypothesis to conclude $\mathcal{M}, x' \models \psi$. Since $(x', v') \in \mathcal{R}$, another instance of the induction hypothesis gives us $\mathcal{M}', v' \models \psi$, and we are done. □

Previously we proved that there can only be finitely many non-bisimilar information cells on any finite set P. Since we have now shown that bisimilarity implies modal equivalence, we immediately get the following result, which will be essential to our proof of termination of our planning algorithms.

Corollary 1. *Given any finite set P, there are only finitely many modally inequivalent information cells on P.*

4.2 Planning trees

When synthesising plans, we explicitly construct the search space of the problem as a labelled AND-OR tree, a familiar model for planning under uncertainty (Ghallab et al. 2004). Our AND-OR trees are called *planning trees*.

Definition 4.3 (Planning Tree)**.** A *planning tree* is a finite, labelled AND-OR tree in which each node n is labelled by a plausibility model $\mathcal{M}(n)$, and each edge (n, m) leaving an OR-node is labelled by an event model $\mathcal{E}(n, m)$.

Planning trees for planning problems $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$ are constructed as follows: Let the initial planning tree T_0 consist of just one OR-node $root(T_0)$ with $\mathcal{M}(root(T_0)) = \mathcal{M}_0$ (the root labels the initial state). A planning tree for $\mathcal{P}$ is then any tree that can be constructed from T_0 by repeated applications of the following non-deterministic tree expansion rule.

Definition 4.4 (Tree Expansion Rule)**.** Let T be a planning tree for a planning problem $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$. The tree expansion rule is defined as follows. Pick an OR-node n in T and an event model $\mathcal{E} \in \mathsf{A}$ applicable in $\mathcal{M}(n)$ with the proviso that $\mathcal{E}$ does not label any existing outgoing edges from n. Then:

[7] Note that we here use the induction hypothesis for the autobisimulation on $\mathcal{M}'$ linking u' and u'', not the bisimulation $\mathcal{R}$ between $\mathcal{M}$ and $\mathcal{M}'$.

(1) Add a new AND-node m to T with $\mathcal{M}(m) = \mathcal{M}(n) \otimes \mathcal{E}$, and add an edge (n, m) with $\mathcal{E}(n, m) = \mathcal{E}$.

(2) For each information cell $\mathcal{M}'$ in $\mathcal{M}(m)$, add an OR-node m' with $\mathcal{M}(m') = \mathcal{M}'$ and add the edge (m, m').

The tree expansion rule is similar in structure to—and inspired by—the expansion rules used in tableau calculi, e.g. for modal and description logics (Horrocks et al. 2006). Note that the expansion rule applies only to OR-nodes, and that an applicable event model can only be used once at each node.

Considering single-agent planning a two-player game, a useful analogy for planning trees are game trees. At an OR-node n, the agent gets to pick any applicable action $\mathcal{E}$ it pleases, winning if it ever reaches an information model in which the goal formula holds (see the definition of solved nodes further below). At an AND-node m, the environment responds by picking one of the information cells of $\mathcal{M}(m)$—which of the distinguishable outcomes is realised when performing the action.

Without restrictions on the tree expansion rule, even very simple planning problems might be infinitely expanded (e.g. by repeatedly choosing a no-op action). Finiteness of trees (and therefore termination) is ensured by the following blocking condition.

$\mathcal{B}$ The tree expansion rule may not be applied to an OR-node n for which there exists an ancestor OR-node m with $\mathcal{M}(m) \equiv \mathcal{M}(n)$.[8]

Lemma 5 (Termination)**.** *Any planning tree built by repeated application of the tree expansion rule under condition $\mathcal{B}$ is finite.*

Proof. Planning trees built by repeated application of the tree expansion rule are finitely branching: the action library is finite, and every plausibility model has only finitely many information cells (the initial state and all event models in the action library are assumed to be finite, and taking the product update of a finite information cell with a finite event model always produces a finite result). Furthermore, condition $\mathcal{B}$ ensures that no branch has infinite length: there only exists finitely many modally inequivalent information cells over any language $L(P)$ with finite P (Corollary 1). König's Lemma now implies finiteness of the planning tree. □

Example 8. Let's consider a planning tree in relation to our basement scenario (cf. Example 7). Here the planning problem is $\mathcal{P}_B = (\mathcal{M}_0, \{\mathsf{flick}, \mathsf{desc}\}, \varphi_g)$ with $\mathcal{M}_0$, flick and desc being defined in Figure 4 and $\varphi_g = \neg t \wedge u$. We have illustrated the planning

[8] Modal equivalence between information cells can be decided by taking their respective bisimulation contractions and then compare for isomorphism, cf. Section 4.1.

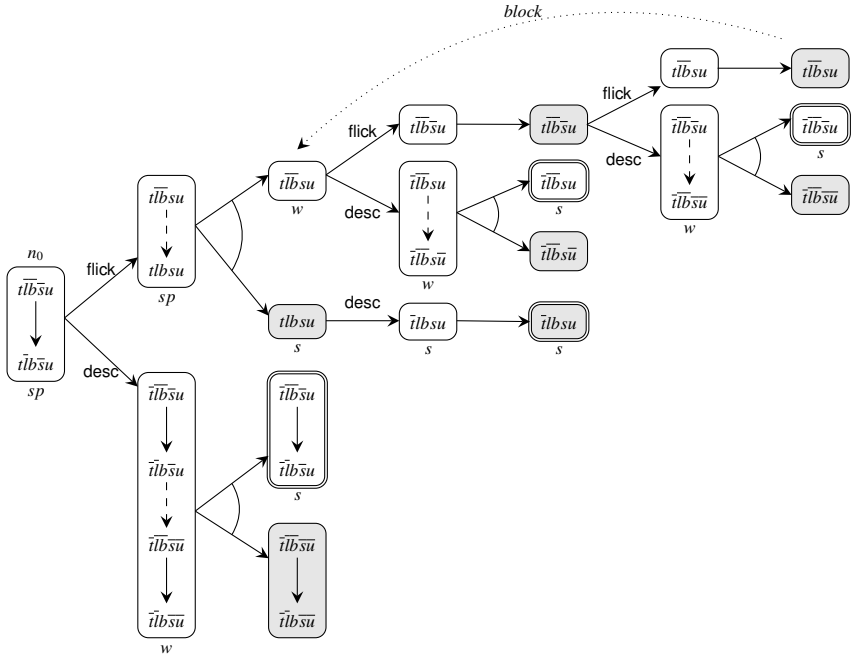


Figure 7: A planning tree T for $\mathcal{P}_B$. Each node contains a (visually compacted) plausibility model. Most plausible children of AND-nodes are gray, doubly drawn OR-nodes satisfy the goal formula, and below solved nodes we've indicated their strength.

tree T in Figure 7, page 274. The root n_0 is an OR-node (representing the initial state $\mathcal{M}_0$), to which the tree expansion rule of Definition 4.4 has been applied twice, once with action $\mathcal{E} = \mathsf{flick}$ and once with $\mathcal{E} = \mathsf{desc}$.

The result of the two tree expansions on n_0 is two AND-nodes (children of n_0) and four OR-nodes (grandchildren of n_0). We end our exposition of the tree expansion rule here, and note that the tree has been fully expanded under the blocking condition $\mathcal{B}$, the dotted edge indicating a leaf having a modally equivalent ancestor. Without the blocking condition, this branch could have been expanded ad infinitum.

Let T denote a planning tree containing an AND-node n with a child m. The node m is called a *most plausible child* of n if $\mathcal{M}(m)$ is among the most plausible information cells of $\mathcal{M}(n)$.

Definition 4.5 (Solved Nodes)**.** Let T be any planning tree for a planning problem $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$. Let α be one of s, w, sp or wp. By recursive definition, a node n in T is called *α-solved* if one of the following holds:

- $\mathcal{M}(n) \models \varphi_g$ (the node satisfies the goal formula).
- n is an OR-node having at least one α-solved child.
- n is an AND-node and:
  - If $\alpha = s$ then all children of n are α-solved.
  - If $\alpha = w$ then at least one child of n is α-solved.
  - If $\alpha = sp$ then all most plausible children of n are α-solved.
  - If $\alpha = wp$ then at least one of the most plausible children of n is α-solved.

Let T denote any planning tree for a planning problem $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$. Below we show that when an OR-node n of T is α-solved, it is possible to construct an α-solution to the planning problem $(\mathcal{M}(n), \mathsf{A}, \varphi_g)$. In particular, if the root node is α-solved, an α-solution to $\mathcal{P}$ can be constructed. As it is never necessary to expand an α-solved node, nor any of its descendants, we can augment the blocking condition $\mathcal{B}$ in the following way (parameterised by α where α is one of s, w, sp or wp).

$\mathcal{B}_\alpha$ The tree expansion rule may not be applied to an OR-node n if one of the following holds: 1) n is α-solved; 2) n has an α-solved ancestor; 3) n has an ancestor OR-node m with $\mathcal{M}(m) \equiv \mathcal{M}(n)$.

A planning tree that has been built according to $\mathcal{B}_\alpha$ is called an *α-planning tree*. Since $\mathcal{B}_\alpha$ is more strict than $\mathcal{B}$, Lemma 5 immediately gives finiteness of α-planning trees—and hence termination of any algorithm building such trees by repeated application of the tree expansion rule. Note that a consequence of $\mathcal{B}_\alpha$ is that in any α-planning tree an α-solved OR-node is either a leaf or has exactly one α-solved child. We make use of this in the following definition.

Definition 4.6 (Plans for Solved Nodes)**.** Let T be any α-planning tree for $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$. For each α-solved node n in T, a plan $\pi(n)$ is defined recursively by:

- if $\mathcal{M}(n) \models \varphi_g$, then $\pi(n) = \mathsf{skip}$.
- if n is an OR-node and m its α-solved child, then $\pi(n) = \mathcal{E}(n, m); \pi(m)$.
- if n is an AND-node and $m_1, \ldots, m_k$ its α-solved children, then
  - If $k = 1$ then $\pi(n) = \pi(m_1)$.

– If $k > 1$ then for all $i = 1, \ldots, k$ let δ_{m_i} denote a formula true in $\mathcal{M}(m_i)$ but not in any of the $\mathcal{M}(m_j) \not\equiv \mathcal{M}(m_i)$ and let $\pi(n) =$
$\mathsf{if}\ \delta_{m_1}\ \mathsf{then}\ \pi(m_1)\ \mathsf{else}\ \mathsf{if}\ \delta_{m_2}\ \mathsf{then}\ \pi(m_2)\ \mathsf{else}\ \cdots\ \mathsf{if}\ \delta_{m_k}\ \mathsf{then}\ \pi(m_k)$.

Note that the plan $\pi(n)$ of a α-solved node n is only uniquely defined up to the choice of δ-formulas in the if-then-else construct. This ambiguity in the definition of $\pi(n)$ will not cause any troubles in what follows, as it only depends on formulas satisfying the stated property. We need, however, to be sure that such formulas always exist and can be computed. To prove this, assume n is an AND-node and $m_1, \ldots, m_k$ its α-solved children. Choose $i \in \{1, \ldots, k\}$, and let $m_{n_1}, \ldots, m_{n_l}$ denote the subsequence of $m_1, \ldots, m_k$ for which $\mathcal{M}(m_{n_j}) \not\equiv \mathcal{M}(m_i)$. We need to prove the existence of a formula δ_{m_i} such that $\mathcal{M}(m_i) \models \delta_{m_i}$ but $\mathcal{M}(m_{n_j}) \not\models \delta_{m_i}$ for all $j = 1, \ldots, l$. Since $\mathcal{M}(m_{n_j}) \not\equiv \mathcal{M}(m_i)$ for all $j = 1, \ldots, l$, there exists formulas δ_j such that $\mathcal{M}(m_i) \models \delta_j$ but $\mathcal{M}(m_{n_j}) \not\models \delta_j$. We then get that $\delta_1 \wedge \delta_2 \wedge \cdots \wedge \delta_l$ is true in $\mathcal{M}(m_i)$ but none of the $\mathcal{M}(m_{n_j})$. Such formulas can definitely be computed, either by brute force search through all formulas ordered by length or more efficiently and systematically by using characterising formulas as in (Andersen et al. 2012) (however, characterising formulas for the present formalism are considerably more complex than in the purely epistemic framework of the cited paper).

Let n be a node of a planning tree T. We say that n is *solved* if it is α-solved for some α. If n is s-solved then it is also sp-solved, if sp-solved then wp-solved, and if wp-solved then w-solved. This gives a natural ordering $s > sp > wp > w$. Note the relation to Lemma 3. We say that a solved node n has *strength* α, if it is α-solved but not β-solved for any $\beta > \alpha$, using the aforementioned ordering.

Example 9. Consider again the planning tree T in Figure 7, page 274, for the planning problem $\mathcal{P}_B = (\mathcal{M}_0, \{\mathsf{flick}, \mathsf{desc}\}, \varphi_g)$ with $\varphi_g = \neg t \wedge u$. Each solved node has been labelled by its strength. The reader is encouraged to check that each node has been labelled correctly according to Definition 4.5. The leafs satisfying the goal formula φ_g have strength s, by definition. The strength of the root node is sp, as its uppermost child has strength sp. The reason this child has strength sp is that *its* most plausible child has strength s.

We see that T is an sp-planning tree, as it is possible to achieve T from n_0 by applying tree expansions in an order that respects $\mathcal{B}_{sp}$. However, it is not the smallest sp-planning tree for the problem, as e.g. the lower subtree is not required for n_0 to be sp-solved. Moreover, T is *not* a w-planning tree, as $\mathcal{B}_w$ would have blocked further expansion once either of the three solved leafs were expanded.

In our soundness result below, we show that plans of α-solved roots are always α-solutions to their corresponding planning problems. Applying Definition 4.6 to

the *sp*-planning tree T gives an *sp*-solution to the basement planning problem, viz. $\pi(n_0) = \mathsf{flick};\ \mathsf{desc};\ \mathsf{skip}$. This is the solution we referred to as the *best* in Example 1: Assuming all actions result in their most plausible outcomes, the best plan is to flick the switch and then descend. After having executed the first action of the plan, flick, the agent will know whether the bulb is broken or not. This is signified by the two distinct information cells resulting from the flick action, see Figure 7, page 274. An agent capable of replanning could thus choose to revise her plan and/or goal if the bulb turns out to be broken.

Theorem 2 (Soundness)**.** *Let α be one of s, w, sp or wp. Let T be an α-planning tree for a problem $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$ such that root(T) is α-solved. Then $\pi(root(T))$ is an α-solution to $\mathcal{P}$.*

Proof. We need to prove that $\pi(root(T))$ is an α-solution to $\mathcal{P}$, that is, $\mathcal{M}_0 \models [\pi(root(T))]_\alpha\, \varphi_g$. Since $\mathcal{M}_0$ is the label of the root, this can be restated as $\mathcal{M}(root(T)) \models [\pi(root(T))]_\alpha\, \varphi_g$. To prove this fact, we will prove the following stronger claim:

$$\text{For each } \alpha\text{-solved } \textsc{or}\text{-node } n \text{ in } T,\ \mathcal{M}(n) \models [\pi(n)]_\alpha\, \varphi_g.$$

We prove this by induction on the height of n. The base case is when n is a leaf (height 0). Since n is α-solved, we must have $\mathcal{M}(n) \models \varphi_g$. In this case $\pi(n) = \mathsf{skip}$. From $\mathcal{M}(n) \models \varphi_g$ we can conclude $\mathcal{M}(n) \models [\mathsf{skip}]_\alpha\, \varphi_g$, that is, $\mathcal{M}(n) \models [\pi(n)]_\alpha\, \varphi_g$. This covers the base case. For the induction step, let n be an arbitrary α-solved $\textsc{or}$-node n of height $h > 0$. Let m denote the α-solved child of n, and $m_1, \ldots, m_l$ denote the children of m. Let $m_{n_1}, \ldots, m_{n_k}$ denote the subsequence of $m_1, \ldots, m_l$ consisting of the α-solved children of m. Then, by Definition 4.6,

- If $k = 1$ then $\pi(n) = \mathcal{E}(n, m); \pi(m_{n_1})$.
- If $k > 1$ then $\pi(n) = \mathcal{E}(n, m); \pi(m)$ where $\pi(m) =$ $\mathsf{if}\ \delta_{m_{n_1}}\ \mathsf{then}\ \pi(m_{n_1})\ \mathsf{else\ if}\ \delta_{m_{n_2}}\ \mathsf{then}\ \pi(m_{n_2})\ \mathsf{else}\ \cdots\ \mathsf{if}\ \delta_{m_{n_k}}\ \mathsf{then}\ \pi(m_{n_k})$.

We here consider only the (more complex) case $k > 1$. Our goal is to prove $\mathcal{M}(n) \models [\pi(n)]_\alpha\, \varphi_g$, that is, $\mathcal{M}(n) \models [\mathcal{E}(n, m); \pi(m)]_\alpha\, \varphi_g$. By the induction hypothesis we have $\mathcal{M}(m_{n_i}) \models [\pi(m_{n_i})]_\alpha\, \varphi_g$ for all $i = 1, \ldots, k$ (the m_{n_i} are of lower height than n).

Claim 1. *$\mathcal{M}(m_{n_i}) \models [\pi(m)]_\alpha\, \varphi_g$ for all $i = 1, \ldots, k$.*

Proof of claim. Let i be given. We need to prove

$$\mathcal{M}(m_{n_i}) \models \left[\mathsf{if}\ \delta_{m_{n_1}}\ \mathsf{then}\ \pi(m_{n_1})\ \mathsf{else}\ \cdots\ \mathsf{if}\ \delta_{m_{n_k}}\ \mathsf{then}\ \pi(m_{n_k})\right]_\alpha \varphi_g.$$

Note that by using item 5 of Lemma 2 it suffices to prove that for all $j = 1, \ldots, k$,

$$\mathcal{M}(m_{n_i}) \models \delta_{m_{n_j}} \text{ implies } \mathcal{M}(m_{n_i}) \models \left[\pi(m_{n_j})\right]_\alpha \varphi_g. \tag{3}$$

Let $j \in \{1, \ldots, k\}$ be chosen arbitrarily. Assume first $j = i$. By induction hypothesis we have $\mathcal{M}(m_{n_j}) \models \left[\pi(m_{n_j})\right]_\alpha \varphi_g$, and hence $\mathcal{M}(m_{n_i}) \models \left[\pi(m_{n_j})\right]_\alpha \varphi_g$. From this (3) immediately follows. Assume now $j \neq i$. By the construction of the δ-formulas, either $\mathcal{M}(m_{n_j}) \equiv \mathcal{M}(m_{n_i})$ or $\mathcal{M}(m_{n_i}) \not\models \delta_{m_{n_j}}$. In the latter case, (3) holds trivially. In case of $\mathcal{M}(m_{n_j}) \equiv \mathcal{M}(m_{n_i})$ we immediately get $\mathcal{M}(m_{n_i}) \models \left[\pi(m_{n_j})\right]_\alpha \varphi_g$, since by induction hypothesis we have $\mathcal{M}(m_{n_j}) \models \left[\pi(m_{n_j})\right]_\alpha \varphi_g$. This concludes the proof of the claim.

Note that by definition of the tree expansion rule (Definition 4.4), $\mathcal{M}(m_1)$, …, $\mathcal{M}(m_l)$ are the information cells in $\mathcal{M}(m)$.

Claim 2. *The following holds:*

- *If $\alpha = s\,(w)$, then for every (some) information cell $\mathcal{M}'$ in $\mathcal{M}(m)$: $\mathcal{M}' \models [\pi(m)]_\alpha \varphi_g$.*
- *If $\alpha = sp\,(wp)$, then for every (some) most plausible information cell $\mathcal{M}'$ in $\mathcal{M}(m)$: $\mathcal{M}' \models [\pi(m)]_\alpha \varphi_g$.*

Proof of claim. We only consider the most complex cases, $\alpha = sp$ and $\alpha = wp$. First consider $\alpha = sp$. Let $\mathcal{M}'$ be a most plausible information cell in $\mathcal{M}(m)$. We need to prove $\mathcal{M}' \models [\pi(m)]_\alpha \varphi_g$. Since, as noted above, $\mathcal{M}(m_1), \ldots, \mathcal{M}(m_l)$ are the information cells in $\mathcal{M}(m)$, we must have $\mathcal{M}' = \mathcal{M}(m_i)$ for some $i \in \{1, \ldots, l\}$. Furthermore, as $\mathcal{M}'$ is among the most plausible information cells in $\mathcal{M}(m)$, m_i must by definition be a most plausible child of m. Definition 4.5 then gives us that m_i is α-solved. Thus $m_i = m_{n_j}$ for some $j \in \{1, \ldots, k\}$. By Claim 1 we have $\mathcal{M}(m_{n_j}) \models [\pi(m)]_\alpha \varphi_g$, and since $\mathcal{M}' = \mathcal{M}(m_i) = \mathcal{M}(m_{n_j})$ this gives the desired conclusion. Now consider the case $\alpha = wp$. Definition 4.5 gives us that at least one of the most plausible children of m are α-solved. By definition, this must be one of the m_{n_i}, $i \in \{1, \ldots, k\}$. Claim 1 gives $\mathcal{M}(m_{n_i}) \models [\pi(m)]_\alpha \varphi_g$. Since m_{n_i} is a most plausible child of m, we must have that $\mathcal{M}(m_{n_i})$ is among the most plausible information cells in $\mathcal{M}(m)$. Hence we have proven that $[\pi(m)]_\alpha \varphi_g$ holds in a most plausible information cell of $\mathcal{M}(m)$.

By definition of the tree expansion rule (Definition 4.4), $\mathcal{M}(m) = \mathcal{M}(n) \otimes \mathcal{E}(n, m)$. Thus we can replace $\mathcal{M}(m)$ by $\mathcal{M}(n) \otimes \mathcal{E}(n, m)$ in Claim 2 above. Using items 1–4 of Lemma 2, we immediately get from Claim 2 that independently of α the following holds: $\mathcal{M}(n) \models [\mathcal{E}(n, m)]_\alpha [\pi(m)]_\alpha \varphi_g$ (the condition $\mathcal{M}(n) \models \langle \mathcal{E}(n, m) \rangle \top$

holds trivially by the tree expansion rule). From this we can then finally conclude $\mathcal{M}(n) \models [\mathcal{E}(n,m); \pi(m)]_\alpha \varphi_g$, as required. □

Theorem 3 (Completeness)**.** *Let α be one of s, w, sp or wp. If there is an α-solution to the planning problem $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$, then an α-planning tree T for $\mathcal{P}$ can be constructed, such that root(T) is α-solved.*

Proof. First note that we have $[\mathsf{skip}; \pi]_\alpha \varphi_g = [\mathsf{skip}]_\alpha ([\pi]_\alpha \varphi_g) = [\pi]_\alpha \varphi_g$. Thus, we can without loss of generality assume that no plan contains a subexpression of the form $\mathsf{skip}; \pi$. The length of a plan π, denoted $|\pi|$, is defined recursively by: $|\mathsf{skip}| = 1$; $|\mathcal{E}| = 1$; $|\mathsf{if}\ \varphi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2| = |\pi_1| + |\pi_2|$; $|\pi_1; \pi_2| = |\pi_1| + |\pi_2|$.

Claim 1. *Let π be an α-solution to $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$ with $|\pi| \geq 2$. Then there exists an α-solution of the form $\mathcal{E}; \pi'$ with $|\mathcal{E}; \pi'| \leq |\pi|$.*

Proof of claim. Proof by induction on $|\pi|$. The base case is $|\pi| = 2$. We have two cases, $\pi = \mathsf{if}\ \varphi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2$ and $\pi = \pi_1; \pi_2$, both with $|\pi_1| = |\pi_2| = 1$. If π is the latter, it already has desired the form. If $\pi = \mathsf{if}\ \varphi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2$ then, by assumption on π, $\mathcal{M}_0 \models [\mathsf{if}\ \varphi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2]_\alpha \varphi_g$. Item 5 of Lemma 2 now gives that $\mathcal{M}_0 \models \varphi$ implies $\mathcal{M}_0 \models [\pi_1]_\alpha \varphi_g$ and $\mathcal{M}_0 \not\models \varphi$ implies $\mathcal{M}_0 \models [\pi_2]_\alpha \varphi_g$. Thus we must either have $\mathcal{M}_0 \models [\pi_1]_\alpha \varphi_g$ or $\mathcal{M}_0 \models [\pi_2]_\alpha \varphi_g$, that is, either π_1 or π_2 is an α-solution to $\mathcal{P}$. Thus either $\pi_1; \mathsf{skip}$ or $\pi_2; \mathsf{skip}$ is an α-solution to $\mathcal{P}$, and both of these have length $|\pi|$. This completes the base case. For the induction step, consider a plan π of length $l > 2$ which is an α-solution to $\mathcal{P}$. We again have two cases to consider, $\pi = \mathsf{if}\ \varphi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2$ and $\pi = \pi_1; \pi_2$. If $\pi = \pi_1; \pi_2$ is an α-solution to $\mathcal{P}$, then π_1 is an α-solution to the planning problem $\mathcal{P}' = (\mathcal{M}_0, \mathsf{A}, [\pi_2]_\alpha \varphi_g)$, as $\mathcal{M}_0 \models [\pi_1; \pi_2]_\alpha \varphi_g \Leftrightarrow \mathcal{M}_0 \models [\pi_1]_\alpha [\pi_2]_\alpha \varphi_g$. Clearly $|\pi_1| < l$, so the induction hypothesis gives that there is an α-solution $(\mathcal{E}; \pi_1')$ to $\mathcal{P}'$, with $|\mathcal{E}; \pi_1'| \leq |\pi_1|$. Then, $\mathcal{E}; \pi_1'; \pi_2$ is an α-solution to $\mathcal{P}$ and we have $|\mathcal{E}; \pi_1'; \pi_2| = |\mathcal{E}; \pi_1'| + |\pi_2| \leq |\pi_1| + |\pi_2| = |\pi|$. If $\pi = \mathsf{if}\ \varphi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2$ is an α-solution to $\mathcal{P}$, then we can as above conclude that either π_1 or π_2 is an α-solution to $\mathcal{P}$. With both $|\pi_1| < l$ and $|\pi_2| < l$, the induction hypothesis gives the existence an α-solution $\mathcal{E}; \pi'$, with $|\mathcal{E}; \pi'| \leq |\pi|$. This completes the proof of the claim.

We now prove the theorem by induction on $|\pi|$, where π is an α-solution to $\mathcal{P} = (\mathcal{M}_0, \mathsf{A}, \varphi_g)$. We need to prove that there exists an α-planning tree for $\mathcal{P}$ in which the root is α-solved. Let T_0 denote the planning tree for $\mathcal{P}$ only consisting of its root node with label $\mathcal{M}_0$. The base case is when $|\pi| = 1$. Here, we have two cases, $\pi = \mathsf{skip}$ and $\pi = \mathcal{E}$. In the first case, the planning tree T_0 already has its root α-solved, since $\mathcal{M}_0 \models [\mathsf{skip}]_\alpha \varphi_g \Leftrightarrow \mathcal{M}_0 \models \varphi_g$. In the second case, $\pi = \mathcal{E}$, we have $\mathcal{M}_0 \models [\mathcal{E}]_\alpha \varphi_g$ as $\pi = \mathcal{E}$ is an α-solution to $\mathcal{P}$. By definition, this means that $\mathcal{E}$ is applicable in $\mathcal{M}_0$, and we can apply the tree expansion rule to T_0, which will produce:

(1) A child m of the root node with $\mathcal{M}(m) = \mathcal{M}_0 \otimes \mathcal{E}$.

(2) Children $m_1, \ldots, m_l$ of m, where $\mathcal{M}(m_1), \ldots, \mathcal{M}(m_l)$ are the information cells of $\mathcal{M}(m)$.

Call the expanded tree T_1. Since $\mathcal{M}_0 \models [\mathcal{E}]_\alpha \varphi_g$, Lemma 2 implies that for every/some/every most plausible/some most plausible information cell $\mathcal{M}'$ in $\mathcal{M}_0 \otimes \mathcal{E}$, $\mathcal{M}' \models \varphi_g$ (where $\alpha = s/w/sp/wp$). Since $\mathcal{M}(m_1), \ldots, \mathcal{M}(m_l)$ are the information cells of $\mathcal{M}_0 \otimes \mathcal{E}$, we can conclude that every/some/every most plausible/some most plausible child of m is α-solved. Hence also m and thus n are α-solved. The base is hereby completed.

For the induction step, let π be an α-solution to $\mathcal{P}$ with length $l > 1$. Let T_0 denote the planning tree for $\mathcal{P}$ consisting only of its root node with label $\mathcal{M}_0$. By Claim 1, there exists an α-solution to $\mathcal{P}$ of the form $\mathcal{E}; \pi'$ with $|\mathcal{E}; \pi'| \leq |\pi|$. As $\mathcal{M}_0 \models [\mathcal{E}; \pi']_\alpha \varphi_g \Leftrightarrow \mathcal{M}_0 \models [\mathcal{E}]_\alpha [\pi']_\alpha \varphi_g$, $\mathcal{E}$ is applicable in $\mathcal{M}_0$. Thus, as in the base case, we can apply the tree expansion rule to T_0 which will produce nodes as in 1 and 2 above. Call the expanded tree T_1. Since $\mathcal{M}_0 \models [\mathcal{E}]_\alpha [\pi']_\alpha \varphi_g$, items 1–4 of Lemma 2 implies that for every/some/every most plausible/some most plausible information cell in $\mathcal{M}_0 \otimes \mathcal{E}$, $[\pi']_\alpha \varphi_g$ holds. Hence, for every/some/every most plausible/some most plausible child m_i of m, $\mathcal{M}(m_i) \models [\pi']_\alpha \varphi_g$. Let $m_{n_1}, \ldots, m_{n_k}$ denote the subsequence of $m_1, \ldots, m_l$ consisting of the children of m for which $\mathcal{M}(m_{n_i}) \models [\pi']_\alpha \varphi_g$. Then, by definition, π' is an α-solution to each of the planning problem $\mathcal{P}_i = (\mathcal{M}(m_{n_i}), \mathsf{A}, \varphi_g)$, $i = 1, \ldots, k$. As $|\pi'| < |\mathcal{E}; \pi'| \leq l$, the induction hypothesis gives that α-planning trees T_i' with α-solved roots can be constructed for each $\mathcal{P}_i$. Let T_2 denote T_1 expanded by adding each planning tree T_i' as the subtree rooted at $\mathcal{M}_{n_i}$. Then each of the nodes m_{n_i} are α-solved in T, and in turn both m and $root(T_2)$ are α-solved. The final thing we need to check is that T_2 has been correctly constructed according to the tree expansion rule, more precisely, that condition $\mathcal{B}_\alpha$ has not been violated. Since each T_i' has in itself been correctly constructed in accordance with $\mathcal{B}_\alpha$, the condition can only have been violated if for one of the non-leaf OR-nodes m' in one of the T_i's, $\mathcal{M}(m') \equiv \mathcal{M}(root(T_2))$. We can then replace the entire planning tree T_2 by a (node-wise modally equivalent) copy of the subtree rooted at m', and we would again have an α-planning tree with an α-solved root. □

4.3 Planning algorithm

In the following, let $\mathcal{P}$ denote any planning problem, and α be one of s, w, sp or wp. With all the previous in place, we now have an algorithm for synthesising an α-solution to $\mathcal{P}$, given as follows.

PLAN($\alpha, \mathcal{P}$)

1 Let T be the α-planning tree only consisting of $root(T)$ labelled by the initial state of $\mathcal{P}$.
2 Repeatedly apply the tree expansion rule of $\mathcal{P}$ to T until no more rules apply satisfying condition $\mathcal{B}_\alpha$.
3 If $root(T)$ is α-solved, return $\pi(root(T))$, otherwise return FAIL.

Theorem 4. PLAN($\alpha, \mathcal{P}$) *is a terminating, sound and complete algorithm for producing α-solutions to planning problems $\mathcal{P}$. Soundness means that if* PLAN($\alpha, \mathcal{P}$) *returns a plan, it is an α-solution to $\mathcal{P}$. Completeness means that if $\mathcal{P}$ has an α-solution,* PLAN($\alpha, \mathcal{P}$) *will return one.*

Proof. Termination comes from Lemma 5 (with $\mathcal{B}$ replaced by the stronger condition $\mathcal{B}_\alpha$), soundness from Theorem 2 and completeness from Theorem 3 (given any two $\mathcal{B}_\alpha$-saturated α-planning trees T_1 and T_2 for the same planning problem, the root node of T_1 is α-solved iff the root node of T_2 is). □

With PLAN($\alpha, \mathcal{P}$) we have given an algorithm for solving α-parametrised planning problems. The α parameter determines the strength of the synthesised plan π, cf. Lemma 3. Whereas the cases of weak ($\alpha = w$) and strong ($\alpha = s$) plans have been the subject of much research, the generation of weak plausibility ($\alpha = wp$) and strong plausibility ($\alpha = sp$) plans based on pre-encoded beliefs is a novelty of this paper. Plans taking plausibility into consideration have several advantages. Conceptually, the basement scenario as formalised by $\mathcal{P}_B$ (cf. Example 7) allowed for several weak solutions (with the shortest one being hazardous to the agent) and *no* strong solutions. In this case, the synthesised strong plausibility solution corresponds to the course of action a rational agent (mindful of her beliefs) should take. There are also computational advantages. An invocation of PLAN($sp, \mathcal{P}$) will expand at most as many nodes as an invocation of PLAN($s, \mathcal{P}$) before returning a result (assuming the same order of tree expansions). As plausibility plans only consider the most plausible information cells, we can prune non-minimal information cells during plan search.

We also envision using this technique in the context of an agent framework where planning, acting and execution monitoring are interleaved.[9] Let us consider the case of strong plausibility planning ($\alpha = sp$). From some initial situation an sp-plan is synthesised which the agent starts executing. If reaching a situation that is not covered by the plan, she restarts the process from this point; i.e. she *replans*. Note that the information cell to replan from is present in the tree as a sibling of the most plausible information

[9] Covering even more mechanisms of agency is *situated planning* (Ghallab et al. 2004).

cell(s) expected from executing the last action. Such replanning mechanisms allow for the *repetition* of actions necessary in some planning problems with cyclic solutions.

We return one last time to the basement problem and consider a modified replace action such that the replacement light bulb might, though it is unlikely, be broken. This means that there is no strong solution. Executing the sp-solution flick; desc, she would replan after flick if that action didn't have the effect of turning on the light. A strong plausibility solution from this point would then be flick; replace; flick; desc.

5 Related and future work

In this paper we have presented α-solutions to planning problems incorporating ontic, epistemic and doxastic notions. The cases of $\alpha = sp/sw$ are, insofar as we are aware, novel concepts not found elsewhere in the literature. Our previous paper (Andersen et al. 2012) concerns the cases $\alpha = s/w$, so that framework deals only with *epistemic* planning problems without a doxastic component. Whereas we characterise solutions as formulas, Andersen and Bolander (2011) take a semantic approach to strong solutions for epistemic planning problems. In their work plans are sequences of actions, requiring conditional choice of actions at different states to be encoded in the action structure itself. By using the $\mathcal{L}(P, \mathsf{A})$ we represent this choice explicitly.

The meaningful plans of de Lima (2007, chap. 2) are reminiscent of the work in this paper. Therein, plan verification is cast as validity of an EDL-consequence in a given system description. Like us, they consider single-agent scenarios, conditional plans, applicability and incomplete knowledge in the initial state. Unlike us, they consider only deterministic epistemic actions (without plausibility). In the multi-agent treatment (de Lima 2007, chap. 4), action laws are translated to a fragment of DEL with only public announcements and public assignments, making actions singleton event models. This means foregoing nondeterminism and therefore sensing actions.

Epistemic planning problems in (Löwe et al. 2011) are solved by producing a sequence of pointed epistemic event models where an external variant of applicability (called *possible at*) is used. Using such a formulation means outcomes of actions are fully determined, making conditional plans and weak solutions superfluous. As noted by the authors, and unlike our framework, their approach does not consider factual change. We stress that Andersen and Bolander (2011), Löwe et al. (2011), de Lima (2007) all consider the multi-agent setting which we have not treated here.

In our work so far, we haven't treated the problem of where domain formulations come from, assuming just that they are given. Standardised description languages are vital if modal logic-based planning is to gain wide acceptance in the planning commu-

nity. Recent work worth noting in this area includes (Baral et al. 2012), which presents a specification language for the multi-agent belief case.

As suggested by our construction of planning trees, there are several connections between our approach for $\alpha = s$ and two-player imperfect information games. First, product updates imply perfect recall (van Benthem 2001). Second, when the game is at a node belonging to an information set, the agent knows a proposition only if it holds throughout the information set. Finally, the strong solutions we synthesise are very similar to mixed strategies. A strong solution caters to any information cell (contingency) it may bring about, by selecting exactly one sub-plan for each (Aumann and Hart 1992).

Our work relates to (Ghallab et al. 2004), where the notions of strong and weak solutions are found, but without plausibilites. Their belief states are sets of states which may be partioned by observation variables. The framework in (Rintanen 2004) describes strong conditional planning (prompted by nondeterministic actions) with partial observability modelled using a fixed set of observable state variables. Our partition of plausibility models into information cells follows straight from the definition of product update. A clear advantage in our approach is that *actions* readily encode both nondetermism and partial observability. Jensen (2013) shows that the *strong plan existence problem* for the framework in (Andersen et al. 2012) is 2-EXP-complete. In our formulation, $\textsc{Plan}(s, \mathcal{P})$ answers the same question for $\mathcal{P}$ (it gives a strong solution if one exists), though with a richer modal language.

We would like to do plan verification and synthesis in the multi-agent setting. We believe that generalising the notions introduced in this paper to multi-pointed plausibility and event models are key. Plan synthesis in the multi-agent setting is undecidable (Andersen and Bolander 2011), but considering restricted classes of actions as is done in (Löwe et al. 2011) seems a viable route for achieving decidable multi-agent planning. Other ideas for future work include replanning algorithms and learning algorithms where plausibilities of actions can be updated when these turn out to have different outcomes than expected.

Acknowledgements

For valuable comments at various stages of the work presented in this article, we would like to extend our gratitude to the following persons: Patrick Allo, Alexandru Baltag, Johan van Benthem, Hans van Ditmarsch, Jens Ulrik Hansen, Sonja Smets and the anonymous reviewers.

References

M. B. Andersen and T. Bolander. Epistemic planning for single- and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34, 2011.

M. B. Andersen, T. Bolander, and M. H. Jensen. Conditional epistemic planning. In L. F. del Cerro, A. Herzig, and J. Mengin, editors, *JELIA*, Lecture Notes in Computer Science. Springer, 2012.

R. J. Aumann and S. Hart, editors. *Handbook of Game Theory with Economic Applications*. Elsevier, 1992.

A. Baltag and L. S. Moss. Logics for Epistemic Programs. *Synthese*, 139:165–224, 2004.

A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. In *G. Bonanno, W. van der Hoek, M. Woolridge (eds.), Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision (LOFT 2006)*, pages 11–24, 2006.

A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundation of Game and Decision Theory (LOFT7)*, volume 3 of *Texts in Logic and Games*, pages 13–60. Amsterdam University Press, 2008.

C. Baral, G. Gelfond, E. Pontelli, and T. C. Son. An action language for reasoning about beliefs in multi-agent domains. In *Proceedings of the 14th International Workshop on Non-Monotonic Reasoning*, 2012.

J. van Benthem. Dynamic odds and ends. Technical Report ML-1998-08, University of Amsterdam, 1998.

J. van Benthem. Games in dynamic-epistemic logic. *Bulletin of Economic Research*, 53(4):219–48, 2001.

P. Blackburn and J. van Benthem. Modal logic: A semantic perspective. In *Handbook of Modal Logic*. Elsevier, 2006.

T. de Lima. *Optimal Methods for Reasoning about Actions and Plans in Multi-Agents Systems*. PhD thesis, IRIT, University of Toulouse 3, France, 2007.

L. Demey. Some remarks on the model theory of epistemic plausibility models. *Journal of Applied Non-Classical Logics*, 21(3-4):375–395, 2011.

H. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147: 229–275, 2005.

H. van Ditmarsch. Revocable belief revision. *Studia Logica*, 2013, to appear.

H. van Ditmarsch and B. Kooi. Semantic results for ontic and epistemic change. In *LOFT 7*, pages 87–117. Amsterdam University Press, 2008.

M. Ghallab, D. S. Nau, and P. Traverso. *Automated Planning: Theory and Practice*. Morgan Kaufmann, 2004.

I. Horrocks, U. Hustadt, U. Sattler, and R. Schmidt. Computational modal logic. In *Handbook of Modal Logic*. Elsevier, 2006.

W. Jamroga and T. Ågotnes. Constructive knowledge: what agents can achieve under imperfect information. *Journal of Applied Non-Classical Logics*, 17(4):423–475, 2007.

M. H. Jensen. The computational complexity of single agent epistemic planning (manuscript). 2013.

B. Löwe, E. Pacuit, and A. Witzel. DEL planning and some tractable cases. *Lecture Notes in Computer Science*, 6953, 2011.

J. Rintanen. Complexity of planning with partial observability. In S. Zilberstein, J. Koehler, and S. Koenig, editors, *ICAPS*, pages 345–354. AAAI, 2004.

The Logic of Joint Ability Under Almost Complete Information

Peter Hawke

Stanford University, Department of Philosophy
phawke@stanford.edu

Abstract

Logics of joint strategic ability have recently received sustained attention in the logic literature, with the most influential being Coalition Logic (CL) and Alternating-time Temporal Logic (ATL). However, the semantical treatment of joint ability claims in these two logics avoids addressing certain epistemic issues related to *coordination* amongst rational agents, by apparently relying on an implicit meta-level assumption of (perfectly reliable) communication between cooperating agents. Yet such epistemic issues arise naturally in settings relevant to ATL and CL: these logics are interpreted on structures that model agents as moving *simultaneously*, and in such scenarios cooperating agents can be subject to uncertainty concerning the *concurrent* actions of other agents in the coalition. In this paper we present a precise syntax and semantics for a variant of CL which we call *Strategic Coordination Logic Mark I* (SCL-I). A key feature of this logic is an operator that aims to capture coalitional ability *without* the assumption of perfect information-sharing between cooperating agents. That is, we use this logic to study a notion of joint ability that is stricter than that in CL. We compare the expressive power and validities of SCL-I to that of CL with some technical results.

1 Introduction: information and joint ability

In recent times, a multitude of *logics for joint strategic ability* have been studied in the literature, chiefly drawing on a tradition emanating from the closely related *Coalition Logic* (CL, Pauly 2001) and *Alternating-time Temporal Logic* (ATL, Alur et al. 2002).

In this paper, we introduce a new logic of joint ability, a novel extension of CL. Our intention in doing so is to explore a quite specific aspect of the interaction between the joint ability of a coalition and the *epistemic status* of agents within the coalition. We locate the general nature of our study with two remarks. First (as is typical in the logical tradition), we are interested in discussing "ability" in a strong sense: the ability to *guarantee* an outcome, a notion closely aligned with that of a "winning strategy" and, possibly, sincere promise-making (Mele 2003)). Secondly, we are interested in the semantics of such joint ability claims under situations where agents face a very *specific* type of uncertainty: uncertainty generated by the *simultaneous moves of other agents*. Following a suggestion in the literature, we call such situations games of *almost perfect information* (van der Hoek and Pauly 2007).

In the present section, we present an informal discussion of the conceptual considerations that motivate a logic of the type that interests us. From section 2 onwards, the discussion will take a more technical turn, as we develop and study our proposed logic in a systematic fashion.

CL enriches classical propositional logic with a family of *coalitional modalities*, with an expression of the form $\langle\langle A\rangle\rangle\mathcal{X}\psi$ intended to mean that the coalition A has the joint ability to secure the outcome ψ in the next move. Attempts to establish precise semantics for the coalitional modalities have highlighted a certain ambiguity in the intuitive notion of (joint) ability, however. While ability *may* be thought of as completely determined by physical and conventional constraints on the actions of the players, such readings do not fully capture the subtle *epistemic* aspects of having a winning strategy. As has long been recognized in game theory, the *information* at an agent's disposal is an enabling (or disabling) factor with respect to ability (witness the discussion of information and "uniform strategy" in (Osborne and Rubinstein 1994)). To illustrate, consider the following variations of the well-known *coordinated attack problem* (Fagin et al. 1995).

Example 1 (Almost Perfect Information)**.** Two armies, respectively commanded by generals **a** and **b**, are positioned in separate locations in the hills overlooking a valley. Below lies their mutual enemy. Individually, each army is not strong enough to defeat the enemy, but if they attack at the same time, it is guaranteed that they will overcome. Suppose further that it is common knowledge amongst the generals that (i) the other army is stationed nearby, (ii) they share the joint goal of defeating the enemy and (iii) that this can only be accomplished through a coordinated effort. However, suppose that the generals have no means for communication and have no predetermined agreement to attack at a certain time. Conclusion: the generals cannot enforce a winning outcome (although they could each roll the dice and win *as a matter of luck*).

Example 2 (Imperfect Information)**.** Imagine a similar scenario to Example 1, with two variations: suppose first that the generals have a perfectly reliable communication channel (short-wave radio, let's say) and can therefore coordinate on a plan of action. Second, suppose that the enemy's strength is such that the two armies, even if working together, can only defeat the enemy with the element of surprise. In this case, victory is only assured if they can attack the enemy at the precise time that the changing of the enemy guard occurs. Neither general knows what times the enemy has chosen as times to change the guard, however. Conclusion: the generals cannot enforce a winning outcome.

Example 3 (Incomplete Information)**.** Suppose that the generals are able to reliably communicate, and can therefore coordinate. However, general **a** has some mistaken beliefs about the possible outcomes of the strategic interaction: arrogantly, she believes (falsely) that general **b**'s army is incompetent and will only hinder an attack. She also believes (falsely) that her army is strong enough to defeat the enemy single-handedly. Conclusion: the generals cannot enforce a winning outcome[1].

These examples showcase different ways in which lack of information disables joint ability. To utilize the terminology of the game theorist, Example 2 is a game of *imperfect information*, where such a game includes situations where players are not sure of the current state of the game, due to limitations on memory or observational powers. Example 3 is an example of a game of *incomplete information*, where some of the players are mistaken or uncertain about the *structural features* of the game, including possibly the nature of the other players. Finally, of special concern to us here, Example 1 may be referred to as a game of *almost perfect information*, following (van der Hoek and Pauly 2007). In such games, players move simultaneously, and may thus be unsure "about the actions the other players are simultaneously taking" (van der Hoek and Pauly 2007, p.1085). The notion of almost perfect information is therefore closely tied to those of communication and binding agreement.

Such epistemic considerations have inspired the development of logics that marry considerations of strategic ability and information, such as Alternating-time Temporal Epistemic Logic (ATEL, van der Hoek and Wooldridge 2003, van der Hoek and Jamroga 2004). The most natural way to interpret these logics, however, is as providing tools for discussing ability in the context of games of imperfect information. Incomplete information and almost perfect information are not addressed directly, ex-

[1] Is it more accurate to say that the generals do not have the ability to enforce the outcome, or to say that they have the ability, but they don't know *how to use* that ability? To our mind, it makes little difference. The latter amounts to saying that the generals are not able to make use of their ability to win. However, denying that one has the ability to use an ability to guarantee an outcome seems essentially equivalent to merely denying that one has the ability to guarantee that outcome.

cept insofar as there may be cases where an imperfect information game may be used to *simulate* these other modes of uncertainty. The result has been that some important epistemic issues related to *coordination* between cooperating agents have received little attention in the logic literature on strategic ability, as pointed out by Ghaderi et al. (2007). This highlights a tension in CL/ATL. The models upon which these logics are interpreted represent agents as moving simultaneously, with the accompanying possibility that in the situations so modeled limits on information-exchange between agents could impede coordination. Yet, in the CL/ATL setting, $\langle\langle A\rangle\rangle X\psi$ is true if there exists an action for each agent in the coalition such that simultaneous performance of those actions will guarantee ψ. As Example 1 illustrates, situations may arise where, for a group of agents to be able to jointly achieve a goal, they must be able to coordinate their actions, and the achievement of such coordination is not (always) simply a matter of there existing an action for each agent such that the coalition can ensure their mutually desired outcome by simultaneously performing those actions. Indeed, it may happen that there are several *incompatible* winning joint strategies for the agents, with the consequence that the agents may need to each select their contributing action with a view to matching their choice to the choices of the other agents in the coalition. So information again assumes importance: is an agent in a position such that he can observe or else predict the choices of the other members of the coalition? The notion of ability upon which CL/ATL (and close variants such as ATEL) operates indicates a meta-level commitment to the strong assumption that agents will always have access to information about the intended moves of all (and only) the members of the coalition to which they belong. This induces the question: can we provide semantics for a coalitional ability operator that drops this assumption?

In this paper, we take up this challenge. To this end, we introduce a new variant of CL which we call, generically, *Strategic Coordination Logic* (SCL). The general idea behind such a logic is to think of the set of joint actions available at a game state as itself a universe of possibilities and then use ideas from (dynamic) epistemic logic to explicitly consider what information an agent will have access to in this universe when selecting their own individual action. In particular, we will precisely define a logic of this type we call Strategic Coordination Logic Mark I (SCL-I)[2]. This logic is interpreted on precisely the same class of structures as CL, making direct comparison to this logic viable.

The key feature of SCL-I is that it contains two types of coalitional modality, $\langle\langle A\rangle\rangle$ and $((A))$. A formula $\langle\langle A\rangle\rangle\psi$ informally means that A can jointly achieve ψ under the assumption of perfectly reliable communication. The strategic ability operator $\langle\langle A\rangle\rangle$

[2]The qualification of "Mark I" is itself a strategic move: we think there is scope for future variations and extensions of the logic we discuss in this paper. In order to pre-emptively avoid introducing increasingly arcane names for such logics, we establish a numbering scheme from the outset.

is intended to be equivalent to the strategic ability operator $\langle\langle A\rangle\rangle\mathcal{X}$ in CL. We provide a precise result to this effect in the paper. More significantly, a formula of the form $((A))\psi$ informally means that A can jointly achieve ψ even when the members of A *cannot* communicate. The most significant contribution of the present paper is to supply precise semantics for this operator. According to these semantics, one case in which $((A))\psi$ holds is if there is an agent in A that can achieve ψ entirely independently of what the other agents in A do. More subtly, however, $((A))\psi$ holds when an agent in A has an individual action she knows can guarantee ψ only because the other agents in the coalition can be expected *not* to choose certain individual actions - namely, no agent in the coalition will choose to play an individual action she knows would guarantee $\neg\psi$. This second case, we believe, is an example of significant *coordinated* action in which there is no communication. In order for these semantics to be plausible, we commit to the meta-level assumption that agents in a coalition have *common knowledge of solidarity*: even if out of communication, it is common knowledge amongst agents in the coalition which agents are in the coalition, what the common goals are of the coalition and that all the agents in the coalition will reliably act so as to fulfill those goals when possible.[3]

Here is our plan for the rest of the paper. In the next section, we provide, as necessary background, the syntax and semantics of two existing logics pivotal to this paper, CL and Public Announcement Logic (PAL), where the latter is a type of dynamic epistemic logic that will be crucial in our definitions of the strategic ability operators in SCL-I. In section 3, we provide a series of simple examples of game-like scenarios intended to more precisely test our intuitions concerning joint ability (and coordination) against the notion of ability at work in standard CL and directly motivate the semantics of a new ability operator. We then develop the syntax and semantics of SCL-I and discuss the success of this logic in dealing with the examples that open section 3. In section 4, we present technical results: we compare the expressive power of SCL-I and CL; and compare the validities of CL and SCL-I (for instance, we demonstrate that the $((A))$-operator, unlike the $\langle\langle A\rangle\rangle$-operator, is not closed under logical implication).[4]

[3] We will also make the assumption of what is sometimes called "complete information" in the game theory literature: there is no confusion or uncertainty amongst the agents as to the structure of the game, including the number of states, agents or actions, or the precise outcomes of actions.

[4] We briefly mention related work. Firstly, we recognize that the study of *signals* and *intentions* in game-like settings is obviously closely related to the issues discussed in this paper. However, we intend to abstract away from both sorts of complexity. More closely related to the project in this paper is (Ghaderi et al. 2007), where the epistemic issues of coordination are identified, discussed and a logical theory for coping with these issues introduced. There are significant differences between this approach and ours, however, both in emphasis and choice of formal techniques. Notably, the logical theory of Ghaderi et al. (2007) is based on the extremely expressive formalism known as the *situation calculus*, while ours stays close to that of CL/ATL tradition. Another work that is related to our own is (van Benthem 2007), where the machinery of

2 Technical background: concurrent game structures, coalition logic and public announcement logic

2.1 Concurrent game structures

We will interpret the formulae of $\mathcal{L}_{CL}$ (as we shall present this logic) and $\mathcal{L}_{SCL-I}$ on *concurrent game structures*. This type of structure is a generalization of labelled transition systems, aimed at providing sufficient flexibility to represent various structures of interest to the game theorist.

Definition 2.1 (Concurrent Game Frame)**.** A *concurrent game frame* (CGF) is a tuple $\langle k, Q, d, \delta \rangle$, where:

- k is a number of agents. We may thus take the set of agents to be $\mathcal{A} = \{1, 2, \ldots, k\}$.
- Q is a non-empty set of states.
- d is a function $d : Q \times \mathcal{A} \longrightarrow \mathbb{N}^+$, where $d(q, a)$ (written more conveniently as $d_a(q)$) returns a positive integer, representing the number of actions available to agent a at state q. We identify the actions available to a at q with the set $D_a(q) = \{1, 2, \ldots, d_a(q)\}$. A *joint action* at state q (which we will denote by σ_q or sometimes simply by σ, where the context is clear) is a tuple $\langle j_1, j_2, \ldots, j_k \rangle$, where $j_i \leq d_i(q)$ for every $i \leq k$. In other words, a joint action at q is simply a collection of actions, one for each agent, that may be performed at state q. Given a joint action $\sigma = \langle j_1, j_2, \ldots, j_k \rangle$, we sometimes write σ^i to indicate j_i. We write $D(q)$ for the set $\{1, \ldots, d_1(q)\} \times \ldots \times \{1, \ldots, d_k(q)\}$ of joint actions at q.
- δ, the *transition function*, is a function that maps a state q and a joint action at q to a state in Q.

Definition 2.2 (Concurrent Game Structure)**.** A *concurrent game structure* (CGS) is a tuple $\langle k, Q, d, \delta, \Pi, \pi \rangle$, namely a concurrent game frame that also includes a countable set of atomic propositions Π and a labeling function $\pi : \Pi \longrightarrow \mathcal{P}(Q)$ that assigns a set of states in Q to each atomic proposition.

---

dynamic epistemic logic is similarly used to study strategic interactions between agents. The most important difference in the focus of the current work versus (van Benthem 2007) is that the former is less concerned with sophisticated solution concepts from non-cooperative game theory, and more with *coalitional ability*, as abstracted away from agent preference.

2.2 Syntax and semantics of Coalition Logic

CL will serve as the benchmark logic against which we will compare SCL-I. We now outline syntax and semantics for CL. The semantics we present differs from the original semantics presented by (Pauly 2001), which is based on coalition effectivity models. Instead, we follow semantics based on concurrent game structures, essentially that of (Alur et al. 2002). Technically, nothing is lost in making this move: it has been shown that the semantics based on effectivity functions is equivalent to that based on concurrent game structures (Goranko and Jamroga 2004).

Definition 2.3 (Syntax of CL)**.** We denote the language of CL by $\mathcal{L}_{CL}$. Let a finite set of agents $\mathcal{A} = \{1, 2, \ldots, k\}$ and a countable set of atomic propositions Π be given. Then the recursive definition of the formulae of this language is as follows:

$$\varphi \ ::= \ p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle\langle A\rangle\rangle \mathcal{X}\varphi$$

where $A \subseteq \mathcal{A}$ and $p \in \Pi$.

The intended informal interpretation of the $\langle\langle A\rangle\rangle \mathcal{X}\psi$ is "coalition A has the joint ability to guarantee the outcome ψ in the next move". Note that we will illustrate in due course that it is appropriate to add the qualification "on the assumption of perfectly reliable communication within the coalition" to each of these informal interpretations.[5]

The language of CL is interpreted on the class of CGSs in which the sets of agents and atomic propositions overlap with those of the language. Before we present the semantics, we require a number of auxiliary notions.

Definition 2.4 (Successor)**.** Given a CGS $\mathcal{S}$, we say that state q' is a *successor* of state q if there is a joint action σ such that $\delta(q, \sigma) = q'$. We denote the set of successors to state q by $succ(q)$.

Definition 2.5 (Computation)**.** Given a CGS $\mathcal{S}$, a *computation* or *run* or *play* on $\mathcal{S}$ is an infinite sequence $\lambda = q_0 q_1 q_2 \ldots$ of states such that $q_{i+1} \in succ(q_i)$ for all $i \geq 0$. A *q-computation* (q-run) is a computation/run where $q_0 = q$. For a computation λ, we use $\lambda[i]$ to denote the ith state of λ.

Definition 2.6 (Strategy)**.** Given a CGS $\mathcal{S}$ and an agent $a \in \mathcal{A}$, a *strategy for a on* $\mathcal{S}$ is a function $f^a : Q \longrightarrow \mathbb{N}^+$ with the restriction that $f^a(q) \leq d_a(q)$. Given a coalition

[5]The reader familiar with CL will note another liberty we have taken in our presentation of the logic: we use $\langle\langle A\rangle\rangle \mathcal{X}$ for the coalitional modality in the language, as opposed to $\langle A\rangle$, in the style of (Pauly 2001). Our choice of notation is that of the language of ATL, where the "next-time" operator - equivalent to the coalitional operator in CL - is of the form $\langle\langle A\rangle\rangle \mathcal{X}$. This is a cosmetic choice, but it serves a function: readability is enhanced by clearly opposing this operator to the operator $\langle\langle A\rangle\rangle$ in the language of SCL-I.

of agents A, an *A-strategy on $\mathcal{S}$* is a set of strategies $F^A = \{f^a | a \in A\}$. Every A-strategy that the coalition A follows from state q onwards induces a set of q-computations on Q, (the *outcomes* of following that strategy), which we denote by $out(q, F^A)$ (or sometimes $out(\mathcal{S}, q, F^A)$ if there is possibility of confusion).

Definition 2.7 (Semantics of CL)**.** Let $\mathcal{S}$ denote a CGS and q any state in $\mathcal{S}$. Then we interpret $\mathcal{L}_{CL}$ formulae in the following way:

- $\mathcal{S}, q \vDash p$, where $p \in \Pi$, iff $p \in \pi(q)$.
- $\mathcal{S}, q \vDash \neg\psi$ iff $\mathcal{S}, q \nvDash \psi$.
- $\mathcal{S}, q \vDash \psi_1 \vee \psi_2$ iff $\mathcal{S}, q \vDash \psi_1$ or $\mathcal{S}, q \vDash \psi_2$.
- $\mathcal{S}, q \vDash \langle\langle A \rangle\rangle \mathcal{X}\psi$ iff there exists an A-strategy F^A such that if $\lambda \in out(q, F^A)$ then $\mathcal{S}, \lambda[1] \vDash \psi$.

We note again that CL is just the next-time fragment of ATL. We will not hesitate to exploit this connection when it is useful to import results or definitions originally framed in the context of ATL.

2.3 Syntax and semantics of PAL

The logic PAL will find utility in our discussion of SCL-I. Roughly, the idea behind Public Announcement Logic (PAL) is to provide logical tools for reasoning about how *public announcements* influence the epistemics of a group of agents: the public announcement of (true) proposition φ *updates* the epistemic situation for the agents, by eliminating from their consideration all possible states in which φ is not true. See (van Ditmarsch et al. 2008) for a full discussion.

"Announcement" need not be interpreted as something literally emanating from a loudspeaker. Generally, an announcement may be understood as an event in which certain information becomes publicly available, by whatever source. In particular, for the applications in this paper, a PAL announcement is best interpreted as a *common inference* that is made by a group of agents on the strength of common knowledge of the structure of the game, nature of the players, and so forth, *not* as an act of communication.

Definition 2.8 (Syntax of PAL)**.** Let a finite set of agents $\mathcal{A}$ and a countable set of atomic propositions Π be given. We inductively define the language of PAL (with distributed knowledge operators), which we refer to as $\mathcal{L}_{PAL}$, by way of the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid K_a\varphi \mid D_A\varphi \mid \langle\varphi\rangle\varphi$$

where $a \in \mathcal{A}$, $A \subseteq \mathcal{A}$ and $p \in \Pi$. Given operator K_a, we define the dual operator $\hat{K}_a$ by $\hat{K}_a\varphi := \neg K_a\neg\varphi$. The connectives $\rightarrow$ and $\wedge$ may be defined in the usual way.

Informally, the intended interpretation of an expression of the form $K_a\varphi$ is "agent a knows that φ". The intended interpretation of $D_A\varphi$ is "it is distributed knowledge amongst the members of A that φ holds". Finally, the intended interpretation of $\langle\varphi\rangle\ \psi$ is "after announcement of φ, it is true that ψ".

Formulae of this language are interpreted on structures we shall call "multi-agent epistemic structures".

Definition 2.9 (Multi-agent Epistemic Structure)**.** Given a finite set of agents $\mathcal{A} = \{1, \ldots, k\}$, a *multi-agent epistemic structure* is a tuple $\mathcal{M} = \langle S, \{\sim_a\}_{a\in\mathcal{A}}, \Pi, V\rangle$, where

- S is a set of possible states.
- For each agent $a \in \mathcal{A}$ there is an equivalence relation $\sim_a \subseteq S \times S$ (that is, each relation is reflexive, transitive and symmetric).
- Π is a countable set of atomic propositions
- $V : \Pi \longrightarrow \mathcal{P}(S)$ is a valuation function that associates each atomic proposition with a set of worlds at which that atomic proposition is true.

Definition 2.10 (Semantics of PAL)**.** Let a finite set of agents $\mathcal{A}$, a countable set of atomic propositions Π and an accompanying multi-agent epistemic structure $\mathcal{M}$ be given. Let $s \in S$. Given any $A \in \mathcal{A}$, let $\sim_A$ denote the relation $(\bigcap_{a\in A} \sim_a)$, the intersection of the equivalence relations associated with the members of A.

- $\mathcal{M}, s \vDash p$, for $p \in \Pi$, iff $s \in V(p)$
- $\mathcal{M}, s \vDash \neg\varphi$ iff $\mathcal{M}, s \nvDash p$
- $\mathcal{M}, s \vDash \varphi \vee \psi$ iff $\mathcal{M}, s \vDash \varphi$ or $\mathcal{M}, s \vDash \psi$
- $\mathcal{M}, s \vDash K_a\varphi$ iff for all $t \in S : s \sim_a t$ implies $\mathcal{M}, t \vDash \varphi$
- $\mathcal{M}, s \vDash D_A\varphi$ iff for all $t \in S : s \sim_A t$ implies $\mathcal{M}, t \vDash \varphi$
- $\mathcal{M}, s \vDash \langle\varphi\rangle\psi$ iff $\mathcal{M}, s \vDash \varphi$ and $\mathcal{M}|_\varphi, s \vDash \psi$

where $\mathcal{M}|_\varphi$, called the *update of* $\mathcal{M}$ *with respect to* φ, is the multi-agent epistemic structure $\left\langle S', \{\sim'_a\}_{a\in\mathcal{A}}, V'\right\rangle$ with

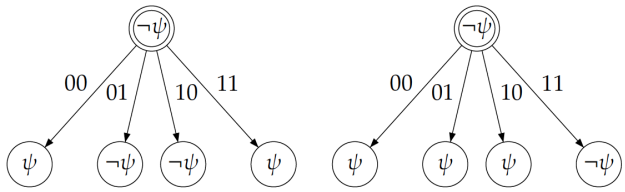


Figure 1: On the left, the concurrent game structure $\mathcal{S}_1$, as discussed in Example 4; on the right, the concurrent game structure $\mathcal{S}_2$, as discussed in Example 5

- $S' = [[\varphi]]_{\mathcal{M}}$, where $[[\varphi]]_{\mathcal{M}}$ is the extension of φ in $\mathcal{M}$
- $\sim'_a = \sim_a \cap \left([[\varphi]]_{\mathcal{M}} \times [[\varphi]]_{\mathcal{M}}\right)$
- $V(p)' = V(p) \cap [[\varphi]]_{\mathcal{M}}$

Note that where confusion is possible, we shall denote the satisfiability relation of PAL by $\models_{PAL}$, in order to distinguish it from another satisfiability relation.

3 Introducing Strategic Coordination Logic

3.1 Motivating examples

What difference can almost perfect information make to coalitional strategic ability? We gather data for the semantics of a new ability operator from the following five examples, some of which will also provide useful counterexamples in due course.

Each example describes a CGS involving two agents **a** and **b**. We are interested in the coalitional ability of these agents at the start state in each structure (represented by a node with a double border in the accompanying figures). Each agent can play action 0 or action 1. For each example, we appeal to intuition to make a judgement about the joint ability of the agents, first under the assumption of reliable communication between the agents, then when supposing this assumption is false.

Example 4. We formalize Example 1 as a simple coordination game $\mathcal{S}_1$, depicted on the left of Figure 1. Suppose that **a** and **b** share the same goal, which we denote by ψ. Now, if both choose action 0 or both choose 1, then they will achieve this goal. On the other hand, if the agents do not choose matching actions, then they will not achieve ψ.

Can our coalition guarantee the achievement of ψ? A common sense answer starts by pointing out an ambiguity in our representation: are the two agents able to share

information with each other? Or are they, like the players in the prisoner's dilemma, in a situation where they cannot communicate? Or are they perhaps forced to make a decision immediately and simultaneously, with there being *no time* for agreement amongst themselves, even if they can communicate in principle? With this in mind, the intuitive assessment of the situation is as follows: if the agents can share information, then they are able to guarantee the achievement of outcome ψ. If it is assumed that the agents are not in a position to share information, then it is clear that the agents *cannot guarantee* the achievement of outcome ψ.

What assessment follows from the standard semantics of CL? Denote the state at which the agents choose an action by q. Then, it is true with respect to the semantics of CL that $\mathcal{S}, q \vDash \langle\langle \mathbf{a}, \mathbf{b} \rangle\rangle \mathcal{X}\psi$. CL gives us a satisfactory answer *only* on the assumption that the agents can both share information and form binding agreements.

Example 5. Consider the game scenario $\mathcal{S}_2$ depicted on right of Figure 1. Here, the agents only fail to achieve ψ if they both select action 1.

Intuitively, our (rational) agents will be able to coordinate in the above game, whether or not they are able to communicate. It is instructive to note the intuitive reasoning required to reach this conclusion. If the agents cannot communicate (but share solidarity), it seems clear that both will select action 0, since, for either agent, if that agent plays action 0, she will guarantee the accomplishment of the coalition's goal ψ, whatever it is the other agent does. A non-communicating rational agent, under the assumption of solidarity, will always select an individual action that they know guarantees the success of the coalition, if such an action exists.

Example 6. Consider a game-scenario $\mathcal{S}_3$ depicted on the left in Figure 2. Here, there is only way in which the agents can achieve ψ: by both selecting action 0.

Intuitively, are the agents able to guarantee ψ? The answer, we propose, is "yes", no matter whether the agents can communicate or not (though, again, presuming solidarity and rationality). In this case, the semantics of CL provide a satisfactory assessment. However, it is instructive to recognize that the reasoning behind this intuitive answer varies depending on whether or not communication is assumed. If it is assumed that the agents can communicate, then the agents will settle on the joint action $(0, 0)$. Suppose that the agents cannot communicate. Supposing solidarity, it seems the agents can rely on each other to not choose an action (namely, action 1 in both cases) that will obviously jeopardize their success[6].

[6]There is an undeniable connection to game-theoretic *dominance* reasoning here. The choice of action 0 is in some sense *dominated* for each agent.

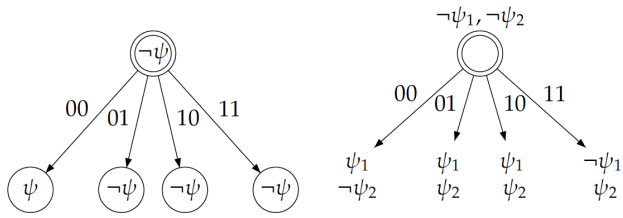


Figure 2: On the left, the concurrent game structure $\mathcal{S}_3$, as discussed in Example 6; on the right, the concurrent game structure $\mathcal{S}_4$, as discussed in Example 7

Example 7. Consider the game $\mathcal{S}_4$ depicted on the right in Figure 2. There are now *two* formulae of interest, ψ_1 and ψ_2, which has the effect of *blending* some of the considerations from the previous examples. Assume that the agents **a** and **b** cannot communicate. Are **a** and **b** able to enforce outcome ψ_1? The answer, intuitively, is 'yes', for similar considerations to Example 5. Similarly, they can achieve ψ_2. However, are the agents able to achieve *both* ψ_1 and ψ_2 *simultaneously*? That is, can they enforce the goal $\psi_1 \wedge \psi_2$? We suggest that with no communication it is intuitively clear that they cannot achieve this goal. The simplest way to see this is to replace the labeling in Figure 2 so as to label the states in the model with the formula $\psi_1 \wedge \psi_2$ and $\neg(\psi_1 \wedge \psi_2)$ where appropriate. It should then strike one that this essentially recreates the situation in Example 4, where the agents are unable to coordinate. This example illustrates is that just because disjoint sub-coalitions in a coalition are able to enforce certain goals, it does not follow, under the assumption of no communication, that the coalition as a whole is then able to enforce those sub-goals *simultaneously*. This is in contrast to the ability operator at work in CL.

Example 8. Consider the game depicted in Figure 3 as the CGS $\mathcal{S}_5$. We can immediately note, drawing on our discussions in earlier examples that the coalition can guarantee the outcome ψ_1 but cannot guarantee the outcome ψ_2 under the assumption of a lack of reliable communication.

Now consider the goal $\psi_1 \vee \psi_2$. That is, imagine that the coalition are indifferent as to which of ψ_1 or ψ_2 is achieved. According again to our earlier assessments, the coalition cannot guarantee the outcome $\psi_1 \vee \psi_2$ without communication (we are dealing with a similar situation to Example 4). Of course, the formula $\psi_1 \rightarrow (\psi_1 \vee \psi_2)$ is a propositional validity. We may conclude that despite the fact that, first, it is a logical truth that θ implies φ and, second, that the coalition can guarantee θ, it does not

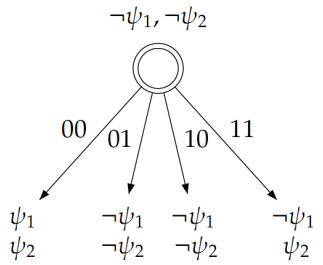


Figure 3: The concurrent game structure $\mathcal{S}_5$ as discussed in Example 8

follow that the coalition can guarantee φ - at least under the assumption of a lack of communication. The ability-with-no-communication operator is not *monontonic*.

3.2 Syntax and semantics of SCL-I

Informed by the motivating examples from the last section, we now present a logic for the $((A))$ operator.

Definition 3.1 (Syntax of SCL-I)**.** Given a set of agents $\mathcal{A}$ and a set of atomic propositions Π, the language of SCL-I with respect to this set of agents and propositions, which we denote by $\mathcal{L}_{SCL\text{-}I}(\mathcal{A}, \Pi)$ (or simply $\mathcal{L}_{SCL\text{-}I}$, where the context is clear), is given inductively by:

$$\varphi \quad ::= \quad p \;\mid\; \neg\varphi \;\mid\; \varphi \vee \varphi \;\mid\; \langle\langle A\rangle\rangle\varphi \;\mid\; ((A))\,\varphi$$

where $p \in \Pi$ and $A \subseteq \mathcal{A}$. A formula in the language is a *propositional formula* if it is a boolean combination of atoms. A formula in the language is a *coalitional formula* if it is of the form $((A))\,\varphi$ or $\langle\langle A\rangle\rangle\varphi$, where φ is any formula in the language.

Connectives such as $\wedge$ and $\rightarrow$ may be defined as usual. The intended interpretation of an expression of the form $\langle\langle A\rangle\rangle\varphi$ is "coalition A can guarantee the outcome φ after the next move, on the assumption that all members of A can reliably communicate with one another". Again, this operator is meant to be essentially equivalent to coalitional operator of CL. The intended interpretation of $((A))\,\varphi$ is "coalition A can guarantee the outcome φ after the next move, even if the members of A cannot communicate with one another".

In order to give semantics for this language, we first introduce the important auxiliary notion of a *generated epistemic structure*. Notice, given some CGS $\mathcal{S}$ and state q in $\mathcal{S}$, that the set $D(q)$ of all joint actions available at q, forms a space of *possibilities* - a space of *possible joint actions* (an *action model* in the terminology of (van Ditmarsch et al. 2008)). We may think of the agents as *jointly choosing* which possibility to actualize, based on each agent's choice of *individual* action. What is needed, then, in the context of the issues under discussion, is some way to represent what *information* any given agent will have when making this choice. To this end, we follow (van Benthem 2007) in noticing that the space of possible joint actions can be naturally endowed with agent-relative indistinguishability relations, thereby generating a multi-agent epistemic structure. Associate with each possible joint action the propositions which result from executing that action. Then, on the supposition that each agent can select his own individual actions but not those of any other agents, we may, for each agent, place an indistinguishability relation between two joint possible actions just in case that agent performs the same individual action in those two joint actions. An agent may then be said to know (that is, accurately predict) that an individual action of theirs will bring about a certain outcome just in case that outcome is brought about by every joint action in which the agent chooses that individual action. This provides for knowledge of outcomes for individual agents under the supposition of no communication. On the other hand, the effects of communication between agents can be understood as captured by *distributed* knowledge.

For technical reasons, a precise definition of generated epistemic structure can only be offered simultaneously as the semantics of SCL-I, as the definitions are mutually recursive. However, for ease of exposition, we present the definition of generated epistemic structure first. Assume for the next definition, then, that the relation $\models_{SCL-I}$ has already been defined.

Definition 3.2 (Generated Epistemic Structure)**.** Given a CGS $\mathcal{S} = \langle k, Q, d, \delta, \Pi, \pi\rangle$ and a state q in $\mathcal{S}$, the *epistemic structure generated by* $\langle \mathcal{S}, q\rangle$, which we denote by $\mathcal{M}_{\mathcal{S}}(q)$ (or just $\mathcal{M}(q)$ or $\mathcal{M}$ where the context is clear), is the epistemic structure $\langle S, \{\sim_a\}_{a\in\mathcal{A}}, \Pi^+, V\rangle$ where

- $S = D(q)$;
- for each $a \in \mathcal{A}$, the relation $\sim_a$ is defined by: for any $\sigma_1, \sigma_2 \in D(q)$, it is the case that $\sigma_1 \sim_a \sigma_2$ iff $\sigma_1^a = \sigma_2^a$;
- $\Pi^+ = \Pi \cup \{\varphi \in \mathcal{L}_{SCL-I} \mid \varphi \text{ is a coalitional formula}\}$;
- for $P \in \Pi^+$,
$$V(P) = \{\sigma \in D(q) \mid \mathcal{S}, \delta(q, \sigma) \models_{SCL-I} P\}\,.$$

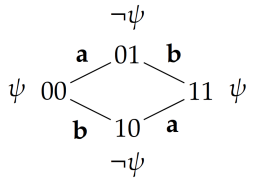


Figure 4: The generated epistemic structure $\mathcal{M}_{S_1}(q)$

Notice that if $P \in \Pi$, then this amounts to:

$$V(P) = \{\sigma \in D(q) \mid \delta(q, \sigma) \in \pi(P)\}.$$

As an example, see Figure 4 for the epistemic structure generated by the coordination game in Example 4.

Given $\mathcal{M}(q)$, an epistemic structure generated by CGS $\mathcal{S}$ at point q, the formulae of $\mathcal{L}_{PAL}$ can be interpreted on such a structure using the standard precise semantics. We intend the informal interpretation of such formulae, in this case, to be slightly non-standard, however. If φ is a propositional formula, we may read $\mathcal{M}(q), \sigma \vDash \varphi$ as "if σ is jointly chosen to be played, then the joint choice of action to be played results in φ". The statement $\mathcal{M}(q), \sigma \vDash K_a\varphi$ may be read as "if σ is jointly chosen to be played, then a has enough information to accurately predict (ie. has "knowledge") that (the individual action they contribute to) the joint choice of action to be played will result in φ". With this in mind, the statement $\mathcal{M}(q), \sigma \vDash D_A\varphi$ may be read as one would expect. The statement $\mathcal{M}(q), \sigma \vDash \langle\psi\rangle K_a\varphi$ may be read as "if σ is jointly chosen to be played, then after the announcement that ψ (ie. after the elimination of all joint actions at which $\neg\psi$ holds), agent a knows that (the individual action they contribute to) the joint choice of action to be played will result in φ".

Now for the semantics of SCL-I. Like CL, the language is interpreted on the set of concurrent game structures.

Definition 3.3 (Semantics of SCL-I)**.** Let $\mathcal{S}$ denote a CGS and q any state in $\mathcal{S}$. Then, in the case of atoms and boolean compositions, the interpretation of $\mathcal{L}_{SCL-I}$ formulae is as in the semantics of CL. For coalitional formulae, on the other hand, we have:

- $\mathcal{S}, q \vDash \langle\langle A\rangle\rangle\varphi$ iff there exists a joint move $\sigma \in D(q)$ such that

$$\mathcal{M}(q), \sigma \vDash_{PAL} D_A\varphi.$$

- $\mathcal{S}, q \vDash ((A))\,\varphi$ iff there exists a joint move $\sigma \in D(q)$ such that

$$\mathcal{M}(q), \sigma \vDash_{PAL} \langle \bigwedge_{a \in A} \hat{K}_a \varphi \rangle \bigvee_{a \in A} K_a \varphi.$$

The rationale for the semantics of $\langle\langle A \rangle\rangle$ is fairly transparent: a coalition can (coordinate to) guarantee an outcome, under the assumption of communication, just in case there is a joint action such that it is distributed knowledge amongst the coalition that their combined contribution to that joint action guarantees their desired outcome, no matter what the agents outside the coalition do. The interpretation of operators of the form $((A))$ requires some more explanation, however. Assuming solidarity, it may be commonly inferred by members of the coalition A (ie. "announced") that no member of the coalition will choose an individual action that that member knows will guarantee that the coalition will *not* achieve its aim. More precisely, if we are considering whether A can jointly achieve ψ, the generated epistemic structure may be updated by eliminating every state in which the sentence $\bigvee_{a \in A} K_a \neg \psi$ holds - or, equivalently, it may be "announced" that $\bigwedge_{a \in A} \hat{K}_a \psi$ holds. After this update, if any agent in the coalition now has an individual action that she knows will guarantee success, then this will be enough to ensure that the coalition's goal will be achieved. More precisely, if there exists a possible joint action, in the updated model, of which it is true that $\bigvee_{a \in A} K_a \psi$, then the agents are able to achieve their goal.[7]

3.3 Application to examples

Consider $\mathcal{S}_1$ as discussed in Example 4, with q denoting the state at which the agents choose an action, $k = 2$ and $A = \{\mathbf{a}, \mathbf{b}\}$. The generated epistemic structure $\mathcal{M}_{\mathcal{S}_1}(q)$ is, again, represented in Figure 4. It is clear that this epistemic structure is unchanged after the announcement of $\bigwedge_{a \in A} \hat{K}_a \psi$, since there is no possible action in the structure at which either $K_{\mathbf{a}} \neg \psi$ or $K_{\mathbf{b}} \neg \psi$. Further, there is no possible action at which either $K_{\mathbf{a}} \psi$ or $K_{\mathbf{b}} \psi$. In total, we have that for any possible action $\sigma \in D(q)$, it is false that $\mathcal{M}_{\mathcal{S}_1}(q), \sigma \vDash_{PAL} \langle \bigwedge_{a \in A} \hat{K}_a \psi \rangle \bigvee_{a \in A} K_a \psi$. Thus, it is false that $\mathcal{S}_1, q \vDash ((A))\,\psi$, as desired. Nevertheless, it is clearly true that $\mathcal{S}_1, q \vDash \langle\langle A \rangle\rangle \psi$.

[7] It is unusual to define semantics for a formal logical language using another formal logical language (except perhaps when the latter is first-order logic). It may be noted, however, that there is nothing essential about using the language of PAL in the semantics for the coalitional operators in SCL-I: the clauses above can easily be expressed using only statements about the concurrent game structure, with no application of PAL sentences or even generated epistemic structures. However, we favour our current approach because a) using the language of PAL allows us to express cumbersome definitions succinctly, and b) we wish to emphasize the interesting point that there is utility in using PAL to express notions of strategic ability in situations of almost perfect information.

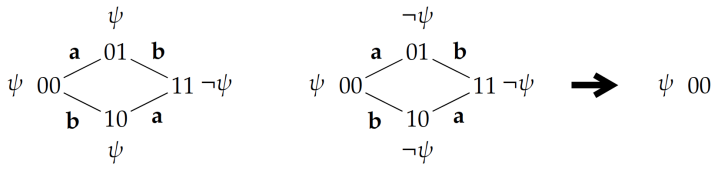


Figure 5: On the left, the generated epistemic structure $\mathcal{M}_{\mathcal{S}_2}(q)$. On the right, the simple generated epistemic structure $\mathcal{M}_{\mathcal{S}_3}(q)$, with update after the announcement of $\bigwedge_{a\in A} \hat{K}_a\psi$

The case of $\mathcal{S}_2$, from Example 5, is easy to evaluate. The announcement of $\bigwedge_{a\in A} \hat{K}_a\psi$ leaves the generated epistemic structure $\mathcal{M}_{\mathcal{S}_2}(q)$ unchanged (see Figure 5). Notice, however, that $K_{\mathbf{a}}\psi$ is true of $\langle 0, 0\rangle$, so $\mathcal{S}_2, q \vDash ((A))\,\psi$, as desired.

Now consider $\mathcal{S}3$ as discussed in Example 6, with q again denoting the state at which the agents choose an action. The generated epistemic structure $\mathcal{M}_{\mathcal{S}_3}(q)$ is represented in Figure 5, along with the updated structure after the announcement of $\bigwedge_{a\in A} \hat{K}_a\psi$. Since at every possible action other than $\langle 0, 0\rangle$ either $K_{\mathbf{a}}\neg\psi$ or $K_{\mathbf{b}}\neg\psi$, the structure is reduced to one possible action after this announcement. Since it is then true of $\langle 0, 0\rangle$ that $K_{\mathbf{a}}\psi$, we have a possible action $\sigma \in D(q)$ such that $\mathcal{M}_{\mathcal{S}_3}(q), \sigma \vDash_{PAL} \langle \bigwedge_{a\in A} \hat{K}_a\psi\rangle \bigvee_{a\in A} K_a\psi$. Thus, it is true that $\mathcal{S}_3, q \vDash ((A))\,\psi$, as desired (not to mention it is true that $\mathcal{S}_3, q \vDash \langle\langle A\rangle\rangle\psi$).

We leave it to the reader to verify that the semantics of SCL-I matches our intuitive assessment of the cases discussed in examples 7 and 8.

4 Some technical results

4.1 Expressivity of SCL-I

In this section, we are interested in studying the relative expressivity of SCL in comparison to CL. Throughout, we assume that a fixed set of agents $\mathcal{A}$ and set of atomic propositions Π is given. We begin by making precise the intuitively obvious overlap between the SCL-I and CL. A more pressing question is the following: the motivation for SCL is to present a new kind of stategic ability operator, and so an associated sense of strategic ability that is not captured by CL. We must be establish that this goal has in fact been accomplished. That is, the following question needs to be answered: is the logic SCL-I able to express, with the $((A))$ operator, a property that cannot be expressed in CL? In other words, is it true that there is no formula (complicated or oth-

erwise) in CL that is equivalent to a formula of the form $((A))\psi$ in SCL-I? To answer this question, we introduce some useful machinery (Ågotnes et al. 2007).

Definition 4.1 (Equivalence of Formulae)**.** Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two logical languages that are interpreted on the same set of models $\mathbb{M}$ (in accordance with a given satisfaction relation for each, respectively $\vDash_1$ and $\vDash_2$). Consider formulae $\varphi_1 \in \mathcal{L}_1$ and $\varphi_2 \in \mathcal{L}_2$. Then we say that φ_1 and φ_2 are *equivalent* just in case they are true in the same states (that is, for any $\mathcal{M} \in \mathbb{M}$ and $q \in \mathcal{M}$, we have that $\mathcal{M}, q \vDash_1 \varphi_1$ iff $\mathcal{M}, q \vDash_2 \varphi_2$). We denote this by $\varphi_1 \equiv \varphi_2$.

Definition 4.2 (Expressive Power)**.** Let two logical languages $\mathcal{L}_1$ and $\mathcal{L}_2$ that are interpreted in the same class of models be given.

- $\mathcal{L}_2$ is *at least as expressive as* $\mathcal{L}_1$ if and only if for every formula $\varphi_1 \in \mathcal{L}_1$ there is a formula $\varphi_2 \in \mathcal{L}_2$ such that $\varphi_1 \equiv \varphi_2$. We denote this by $\mathcal{L}_1 \preceq \mathcal{L}_2$.
- $\mathcal{L}_2$ is *more expressive than* $\mathcal{L}_1$ if and only if $\mathcal{L}_1 \preceq \mathcal{L}_2$ but $\mathcal{L}_2 \npreceq \mathcal{L}_1$. We denote this by $\mathcal{L}_1 \prec \mathcal{L}_2$.

Proposition 1. $\mathcal{L}_{CL} \preceq \mathcal{L}_{SCL\text{–}I}$.

Proof. It follows by a straightforward induction on the complexity of formulae in $\mathcal{L}_{CL}$ that for any $\psi \in \mathcal{L}_{CL}$ that $\psi \equiv tr(\psi)$, where $tr : \mathcal{L}_{CL} \longrightarrow \mathcal{L}_{SCL\text{–}I}$ is the obvious translation function. □

In service of our next result, we now present a notion of bisimulation for concurrent game structures, introduced in (Ågotnes et al. 2007).

Definition 4.3 (Bisimulation For CGSs)**.** Let CGSs $\mathcal{S}_1 = \langle k, Q_1, d_1, \delta_1, \Pi_1, \pi_1\rangle$ and $\mathcal{S}_2 = \langle k, Q_2, d_2, \delta_2, \Pi_2, \pi_2\rangle$ be given, with $\mathcal{A} = \{1, 2, \ldots, k\}$.

1. Let a set of agents $A \subseteq \mathcal{A}$ be given. A relation $\beta \subseteq Q_1 \times Q_2$ is a *(global) A-bisimulation* between $\mathcal{S}_1$ and $\mathcal{S}_2$, denoted $\mathcal{S}_1 \rightleftarrows^A_\beta \mathcal{S}_2$, iff for any $q_1 \in Q_1$ and $q_2 \in Q_2$, $q_1 \beta q_2$ implies that

   Local Harmony $\pi_1(q_1) = \pi_2(q_2)$.

   Forth For any A-strategy F^A_1 on $\mathcal{S}_1$, there exists an A-strategy F^A_2 on $\mathcal{S}_2$ such that for every computation $\lambda_2 \in out(\mathcal{S}_2, q_2, F^A_2)$ there exists a computation $\lambda_1 \in out(\mathcal{S}_1, q_1, F^A_1)$ such that $\lambda_1[1] \beta \lambda_2[1]$.

   Back Likewise, for 1 and 2 swapped.

2. If $\mathcal{S}_1 \rightleftarrows^A_\beta \mathcal{S}_2$ and $q_1 \beta q_2$, then we also say that β is a *local A-bisimulation* between $(\mathcal{S}_1, q_1)$ and $(\mathcal{S}_2, q_2)$, denoted $(\mathcal{S}_1, q_1) \rightleftarrows^A_\beta (\mathcal{S}_2, q_2)$.

3. If β is a A-bisimulation between $\mathcal{S}_1$ and $\mathcal{S}_2$ for every $A \subseteq \mathcal{A}$, we call it a *(full global) bisimulation* between $\mathcal{S}_1$ and $\mathcal{S}_2$, denoted $\mathcal{S}_1 \rightleftarrows_\beta \mathcal{S}_2$. Likewise, we define a *full local bisimulation* between $(\mathcal{S}_1, q_1)$ and $(\mathcal{S}_2, q_2)$, denoted $(\mathcal{S}_1, q_1) \rightleftarrows_\beta (\mathcal{S}_2, q_2)$.

Definition 4.4. For a fixed $A \subseteq \mathcal{A}$, we denote by $\mathcal{L}_{CL}(\Pi[A])$ the fragment of $\mathcal{L}_{CL}(\Pi, \mathcal{A})$ consisting of only those formulae generated by

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle\langle A \rangle\rangle \mathcal{X}\varphi$$

Theorem 1. *If $\mathcal{S}_1 \rightleftarrows^A_\beta \mathcal{S}_2$ and $q_1 \beta q_2$, then, for every formula $\varphi \in \mathcal{L}_{CL}(\Pi[A])$, we have that $\mathcal{S}_1, q_1 \vDash_{CL} \varphi$ iff $\mathcal{S}_2, q_2 \vDash_{CL} \varphi$.*

Proof. See the appendix of (Ågotnes et al. 2007). □

Corollary 1. *If $\mathcal{S}_1 \rightleftarrows_\beta \mathcal{S}_2$ and $q_1 \beta q_2$, then, for every formula $\varphi \in \mathcal{L}_{CL}(\Pi, \mathcal{A})$, we have that $\mathcal{S}_1, q_1 \vDash_{CL} \varphi$ iff $\mathcal{S}_2, q_2 \vDash_{CL} \varphi$.*

With an appropriate notion of bisimulation in hand, a standard strategy for showing that a language has expressive power beyond that of CL presents itself: if we can find two models such that 1) the models are bisimilar to each other (and so unable to be distinguished by CL) and 2) that there is some formula from the language being compared to CL that holds on the one model but not the other, then we may conclude that this language can express properties that CL is unable to express. We shall follow precisely this strategy to show that SCL-I has expressive power beyond that of CL.

Theorem 2. *Suppose that $|\mathcal{A}| \geq 2$. Then $\mathcal{L}_{CL} \prec \mathcal{L}_{SCL-I}$.*

Proof. To begin, we assume that $k = 2$, i.e., that there are only two agents in the system. Now, consider the CGS $\mathcal{S}_6$, as depicted at the top in Figure 6 and the CGS $\mathcal{S}_3$, as discussed in motivating Example 6, and depicted (again) at the bottom of Figure 7. Both are CGSs in which $k = 2$. In $\mathcal{S}_6$, each agent has three moves from which to choose at the start state - namely, 0, 1 and 2. For the sake of readability, we label the states in the figure with the joint action that leads to that state.

Now, first, we know from the discussion of Example 6 that $\mathcal{S}_3, q \vDash ((A))\psi$, where q refers to the start state in $\mathcal{S}_3$ (for convenience, we refer to the start state in both structures as q). However, it is straightforward to check that $\mathcal{S}_6, q \nvDash ((A))\psi$: the generated epistemic model for $\mathcal{S}_6$ at q is updated to eliminate all joint actions in which some

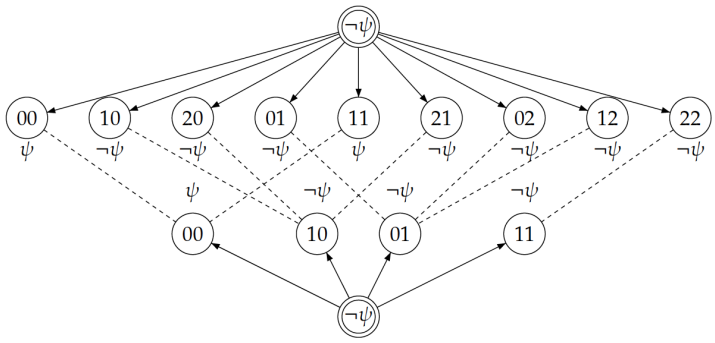


Figure 6: A depiction of the structure $\mathcal{S}_6$ (above), the structure $\mathcal{S}_3$ (below), and the bisimulation between them (the broken lines between states)

agent chooses action 2, but this just leaves us with the epistemic structure generated by $\mathcal{S}_1$ in motivating Example 4.

Now, we claim that a full global bisimulation exists between $\mathcal{S}_3$ and $\mathcal{S}_6$. We depict this bisimulation with the broken lines between states in the respective structures in Figure 6 (the start states in $\mathcal{S}_3$ and $\mathcal{S}_6$ are also related by this relation, but we omit depicting this in the figure for the sake of readability). It is straightforward, if a bit laborious, to check that the relation so depicted does in fact constitute a full global bisimulation between the two structures, and we leave this to the reader. From this we may conclude that there is no CL formula that can distinguish $\mathcal{S}_3$ from $\mathcal{S}_6$. Since the SCL-formula $((A))\psi$ can distinguish the two structures, we conclude that there is no equivalent formula in CL for the formula $((A))\psi$ in SCL-I.

Finally, we drop the assumption that $k = 2$ and consider a fixed, but arbitrary, number of agents k, where $k \geq 2$. In this case, the structures $\mathcal{S}_3$ and $\mathcal{S}_6$ can be recreated with the additional agents by simply assigning only one action at the start states of the structures to each agent i, where $i > 2$. The above reasoning can then be reproduced to achieve the general result. □

4.2 Validities and invalidities

In this section, we consider some signficant validities (and invalidities) of the logic SCL-I. We remark in passing that it is of interest to compare these systems to the (sound and complete) axiomatic systems for CL in (Pauly 2001) and ATL in (van Drimmelen and Goranko 2006).

Definition 4.5 (Significant Validities for SCL-I)**.** The set of validities for SCL-I includes all those based on the following schemata, where $A, B, \{\mathbf{a}\} \subseteq \mathcal{A}$:

- $\langle\langle\rangle\rangle$**-axioms:**

  $\bot_{\langle\langle\rangle\rangle} : \neg\langle\langle A\rangle\rangle\bot$ $\qquad$ $\top_{\langle\langle\rangle\rangle} : \langle\langle A\rangle\rangle\top$

  $\mathcal{A} : \neg\langle\langle\emptyset\rangle\rangle\neg\psi \rightarrow \langle\langle\mathcal{A}\rangle\rangle\psi$

  $\mathbf{S}_{\langle\langle\rangle\rangle} : \langle\langle A\rangle\rangle\psi_1 \wedge \langle\langle B\rangle\rangle\psi_2 \rightarrow \langle\langle A \cup B\rangle\rangle(\psi_1 \wedge \psi_2)$, where $A \cap B = \emptyset$

- $(())$**-axioms:**

  $\emptyset : \neg\,((\emptyset))\,\top$

  $((A))$-coalition-monotonicity: $((A))\,\psi \rightarrow ((A \cup B))\,\psi$

- ***Interaction Axioms:***

  Int1 : $((A))\,\psi \rightarrow \langle\langle A\rangle\rangle\psi$

  Int2 : $\langle\langle\mathbf{a}\rangle\rangle\psi \rightarrow ((\mathbf{a}))\psi$

  Int3 : $\Big(\bigwedge_{a\in A}\neg\langle\langle a\rangle\rangle\psi\Big) \wedge \Big(\bigwedge_{a\in A}\neg\langle\langle a\rangle\rangle\neg\psi\Big) \rightarrow \neg\,((A))\,\psi$

  Int4 : $((B \cup C))\psi \wedge \Big(\bigwedge_{b\in B}\neg\langle\langle b\rangle\rangle\neg\psi\Big) \wedge \Big(\bigwedge_{c\in C}\langle\langle c\rangle\rangle\neg\psi\Big)$
  $\rightarrow \bigvee_{b\in B}((C \cup \{b\}))\psi \vee ((C))\psi$

- ***Rules of Inference:***

  Modus Ponens: from ψ_1 and $\psi_1 \rightarrow \psi_2$, infer ψ_2

  $\langle\langle A\rangle\rangle$-monotonicity: from $\psi_1 \rightarrow \psi_2$, infer $\langle\langle A\rangle\rangle\psi_1 \rightarrow \langle\langle A\rangle\rangle\psi_2$

  $((A))$-equivalence: from $\psi_1 \leftrightarrow \psi_2$ infer $((A))\,\psi_1 \leftrightarrow ((A))\,\psi_2$

It is worth trying to capture informally what the interaction axioms suggest: the **Int1**-axiom says that if a coalition can enforce something without being able to communicate, then they can enforce it if they are able to communicate; the **Int2**-axiom says that when considering the abilities of individual agents, communication powers are irrelevant; the **Int3**-axiom says that, when considering a coalition of agents that can't communicate, if none of the agents can act individually to bring about the goal of the coalition and none of the agents can act individually to avoid actions that will definitely jeopardize the goal of the coalition, then the coalition is helpless to achieve its goal;

the **Int4**-axiom says that if a coalition are unable to communicate yet can achieve some goal, then there is some subcoalition of the coalition that are able to achieve that goal, where this subcoalition consists of the agents in the coalition who are able to individually perform (and therefore avoid) actions that will jeopardize the goal of the coalition and *at most* one other member of the coalition. These last two axioms go some way to capturing the spirit of the dynamic semantics of the $((A))$ operator in the language.

Proposition 2 (Soundness)**.** *Each above axiom is valid, and each inference rule preserves validity.*

Proof. Throughout the proof, let $\mathcal{S}$ refer to an arbitrary CGS, q be an arbitrary state in $\mathcal{S}$ and $\mathcal{M}$ refer to the generated epistemic structure $\mathcal{M}_{\mathcal{S}}(q)$.

The proof for each axiom is straightforward if fussy. We prove three of the results and leave the rest as an exercise for the reader.

$\emptyset$: for $\mathcal{S}, q \vDash ((\emptyset))\top$ to hold, there must exist a joint move $\sigma \in D(q)$ such that $\mathcal{M}, \sigma \vDash_{PAL} \langle \bigwedge_{a\in\emptyset} \hat{K}_a \top\rangle \bigvee_{a\in\emptyset} K_a \top$. Since $\bigwedge_{a\in\emptyset} \hat{K}_a \top \equiv \top$ and $\bigvee_{a\in\emptyset} K_a \top \equiv \bot$, it follows that there can be no such move.

Int3: suppose that

$$\mathcal{S}, q \vDash \bigwedge_{a\in A} \neg\langle\langle a\rangle\rangle\psi$$

and that

$$\mathcal{S}, q \vDash \bigwedge_{a\in A} \neg\langle\langle a\rangle\rangle\neg\psi.$$

Now, by definition, there exists $\sigma \in D(q)$ such that $\mathcal{M}, \sigma \vDash K_a\neg\psi$ just in case $\mathcal{S}, q \vDash \langle\langle a\rangle\rangle\neg\psi$. Hence, for all $\sigma \in D(q)$ and $a \in A$, we have that $\mathcal{M}, \sigma \nvDash K_a\neg\psi$. Thus, the update following the announcement of $\bigwedge_{a\in A} \hat{K}_a\psi$ leaves the model unchanged. This means that $\mathcal{S}, q \vDash ((A))\,\psi$ can hold only if there exists $\sigma \in D(q)$ such that $\mathcal{M}, \sigma \vDash K_a\psi$ for some $a \in A$. However, there exists $\sigma \in D(q)$ such that $\mathcal{M}, \sigma \vDash K_a\psi$ just in case $\mathcal{S}, q \vDash \langle\langle a\rangle\rangle\psi$, which is assumed to be false for all $a \in A$. So, for all $\sigma \in D(q)$ and $a \in A$, it follows that $\mathcal{M}, \sigma \nvDash K_a\psi$, and so $\mathcal{S}, q \nvDash ((A))\,\psi$.

Int4: suppose that $\mathcal{S}, q \vDash ((A))\,\psi$ and that $C = \{a \in A \mid \mathcal{S}, q \vDash \langle\langle a\rangle\rangle\neg\psi\}$. From the former supposition, we may conclude that there exists $\mathbf{a}$ and $\sigma_* \in D(q)$ such that $\mathcal{M}|_{(\bigwedge_{a\in A} \hat{K}_a\psi)}, \sigma_* \vDash K_{\mathbf{a}}\psi$. From the latter supposition, it may be concluded that

$$C = \{a \in A \mid \exists\sigma \in D(q)\ s.t.\ \mathcal{M}, \sigma \vDash K_a\neg\psi\}\,.$$

It follows easily that $\mathcal{M}|_{(\bigwedge_{a\in A} \hat{K}_a\psi)}$ is the same structure as $\mathcal{M}|_{(\bigwedge_{a\in C\cup\{\mathbf{a}\}} \hat{K}_a\psi)}$, and so that $\mathcal{M}|_{(\bigwedge_{a\in C\cup\{\mathbf{a}\}} \hat{K}_a\psi)}, \sigma_* \vDash K_{\mathbf{a}}\psi$. The result follows.

□

Proposition 3 (Invalidities)**.** *The following are* not *valid axioms or rules for SCL-I:*

(1) $((A))\,\psi_1 \wedge ((B))\,\psi_2 \rightarrow ((A \cup B))\,(\psi_1 \wedge \psi_2)$*, where* $A \cap B = \emptyset$

(2) $\neg((\emptyset))\,\neg\psi \rightarrow ((\mathcal{A}))\,\psi$

(3) from $\psi_1 \rightarrow \psi_2$*, conclude* $((A))\,\psi_1 \rightarrow ((A))\,\psi_2$

Proof. (1) Motivating Example 7 provides a counter-example: take $A = \{\mathbf{a}\}$ and $B = \{\mathbf{b}\}$.

(2) Motiviating Example 4 provides a counter-example: from the $\emptyset$-axiom, it follows that $\neg((\emptyset))\,\neg\psi$ is a validity and so true in $\mathcal{S}_1$ at q. However, it is false that $((A))\,\psi$ at q.

(3) Motivating Example 8 provides a counter-example: it is valid that $\psi_1 \rightarrow (\psi_1 \vee \psi_2)$, yet for the structure $\mathcal{S}_5$ of Example 8 it holds that $\mathcal{S}_5, q \vDash ((A))\,\psi_1 \wedge \neg((A))\,(\psi_1 \vee \psi_2)$.

□

5 Conclusion and further work

We have introduced and motivated a new variation on Coalition Logic, which we have called Strategic Coordination Logic (Mark I). The purpose of this logic is to deal more effectively with situations of almost perfect information, and the accompanying effects on the ability of coalitions of agents to coordinate. We have put forward a precise syntax and semantics for SCL-I, and provided technical results concerning expressivity and validity for the logic that illuminate differences between CL and SCL-I.

There is much scope for further research. Certain technical matters are not dealt with in this paper, such as a presentation of a sound and complete axiom system, or results on the complexity of model-checking for SCL-I (such results are in our possession, but cannot be presented here due to limitations of space). Another route is to explore further refinements of Strategic Coordination Logic. For instance, SCL-I can express joint ability under the assumption of reliable communication within coalitions, and under the assumption of a lack of reliable communication within coalitions. What of intermediate cases, however, such as where certain agents in the coalition can communicate, but not others?

Acknowledgements Thanks to Dmitry Shkatov, Valentin Goranko, Wes Holliday, Thomas Icard, and various anonymous referees for helpful comments on earlier versions of this paper.

References

R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Journal of the ACM*, (49):672–713, 2002.

J. van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):13–45, 2007.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, 2008.

G. van Drimmelen and V. Goranko. Complete axiomitization and decidability of alternating-time temporal logic. *Theoretical Computer Science*, (353):93–117, 2006.

R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

H. Ghaderi, Y. Lesperance, and H. Levesque. A logicial theory of coordination and joint ability. In *Proceedings of Twenty-Second Conference on Artifical Intelligence (AAAI07)*, pages 421–426, Vancouver, BC, 2007.

V. Goranko and W. Jamroga. Comparing semantics for logics of multi-agent systems. *Synthese*, 139 (2):241–280, 2004.

A. R. Mele. Agent's abilities. *Nous*, 37(3):447–470, 2003.

M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

M. Pauly. *Logic for Social Software*. PhD thesis, University of Amsterdam, 2001.

W. van der Hoek and W. Jamroga. Agents who know how to play. *Fundamenta Informaticae*, (62):1–35, 2004.

W. van der Hoek and M. Pauly. Modal logic for games and information. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*. Elsevier, 2007.

W. van der Hoek and M. Wooldridge. Cooperation, knowledge and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica*, (75):125–127, 2003.

T. Ågotnes, V. Goranko, and W. Jamroga. Alternating-time temporal logics with irrevocable strategies. In D. Samet, editor, *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge (TARK XI)*, pages 15–24. Presses Universitaires de Louvain, 2007.

Computation as Social Agency: What and How

Johan van Benthem

University of Amsterdam
Stanford University
johan.vanbenthem@uva.nl

1 Views of computing, from "what" to "how"

This Turing Year has been the occasion for lively debates about the nature of computing. Are we on the threshold of new styles of computation that transcend the limitations of the established paradigm? Let us briefly recall three classical themes of the golden age when Turing and his generation made computation a subject for mathematical inquiry, and hand in hand with that, for practical development. First of all, by analyzing the bare basics of mechanical computing, Turing defined a *Universal Machine* that can compute the result of any algorithm on any matching input, when both are presented in suitably encoded form. This notion then supported the subsequent development of Recursion Theory, bringing to light both the basic structures and powers of effective computation, but also its limitations as exemplified in the undecidability of the Halting Problem. On the basis of this and other, equivalent models proposed at the time, *Church Thesis* then claimed that all effectively computable functions over the natural numbers (a canonical domain that can mimic non-numerical computation by various encodings going back to Gödel and others), coincide with the 'recursive functions', that can be computed on Turing machines. As the power of this paradigm became clear, it was suggested in the famous *Turing Test* that computation might well emulate human cognition, to the extent that in conversation, humans would not be able to tell whether they are interacting with another human or a machine.

Now, 80 years later, computer science and information technology have transformed the world of both machines and humans in sometimes wholly unpredictable

ways. Given the experience obtained over this period, can we spring the bounds of the classical age, and compute a larger class of functions/problems after all? There are interesting and lively current debates in the US and Europe on this theme, with proposals ranging from using infinite machines to letting the physical universe do the computing in its own ways (Cooper and Sorbi 2008, Cooper 2011). I am not going to enter these debates here, except for one basic comment. It seems important to make a distinction here between two issues:

(a) What can we compute, and (b) how can we compute it?

Somewhat apodictically, my view is this. I see no evidence in current debates that we can compute more than before, forcing us to extend the calibration class of recursive functions. But then, this 'What' question is not of great interest. Of much greater interest is a 'How' question, not addressed by Church's Thesis, namely, what are natural styles of computing? Or if you insist on 'what' questions after all: do not ask what is *computable*, but what is *computing*, viewed as a kind of process producing characteristic forms of *behavior*.

Right from its start, the history of computer science has shown an outburst of ideas on these themes, and this paper will be about one of these: computation as social agency. My discussion will have a logical slant, reflecting my own work in this area (van Benthem 2008; 2011), and I am not claiming to represent public opinion in computer science.

2 Computer science as a hot spring of ideas

Before I start with my own theme, here is some very quick background that not all of my fellow logicians interested in the foundations of computing seem aware of.

Logic and fine-structuring views of computing Turing machines have opaque programs telling the machine in complete detail what to do in machine code, making heavy use of that old enemy of perspicuity called 'go to' instructions (Dijkstra 1968). Real computer science took off when higher programming languages were invented, that represent higher-level ideas on the sort of computation taking place. One can think of programs in such languages as 'algorithms' that describe the essence of some computational task at some more suitable abstraction level. Different programming languages have given a wealth of perspectives on this, often drawing on traditions in logic.

For instance, imperative programs like those of Algol or C^+ may be viewed as a 'dynamified' version of logical formulas in standard formalisms like predicate logic, telling the machine what sort of results to achieve given certain preconditions. Such

systems lend themselves well to model-theoretic semantics in the usual logical style (first-order, modal, or otherwise), witness the development of Hoare Calculus and Dynamic Logic. On the other hand, there are functional programming languages like LISP or Haskell, akin to systems of lambda calculus and type theory, closer to the proof-theoretic and category-theoretic traditions in logic. And of course, there are many other styles that do not fall simply into this dichotomy, including object-oriented programs, logic programs, and so on. The semantics for this large family programming languages have provided a wealth of matching process models that offer many deep answers to the issue of how we compute.

Distributed computation and process theory One major challenge around 1980 was a theoretical reflection on the practice of distributed computing emerging at the time. One major development here, moving up the abstraction level beyond programming languages, was the invention of Process Algebra by Milner, Bergstra, and others (cf. Bergstra and Smolka 2001), an abstract view of processes as graphs modulo some suitable notion of structural behavioral invariance, often some variant of bisimulation. While it is true to say that no consensus has emerged on one canonical model of distributed computation, comparable in breadth of allegiance to Turing machines, a deep process theory did emerge with many features that have no counterpart in the classical theory of sequential computing (van Emde Boas 1990). Abstract process theories are still emerging in the foundations of computation. A noticeable new development has been the birth of co-algebra, as a theory of computing on infinite streams, tied to fixed-point logics and category-theoretic methods (Venema 2006; 2007). My point here is very modest: thinking about the foundational hows of computation is a productive line of thought that shows no signs of abating yet[1]. For instance, in the last 15 years, a striking model for multi-agent distributed computing has been the introduction of game models (Abramsky 2008), and in another paradigm (Thomas et al. 2002), leading to new encounters between computer science, logic, and game theory (van Benthem 2013a).

3 Computation and social agency

It is often said that Turing took the human out of the term 'computer', extracting only the abstract procedures behind their pencil-and-paper activity. In that light, the Turing

[1] My subsidiary point is addressed more to some of my fellow logicians, who sometimes think that no deep insights worth our august attention have ever come out of actual computer science. My point to them is simply that there is life after Turing: deep thinkers on the foundations of computation such as McCarthy, Dijkstra, Hoare, Pnueli, Milner, Pratt or Abramsky have a lot to teach us.

Test then added insult to injury, since the computer thus defined might then even dispel the mystery of our other intelligent tasks just as well. And even without grand reductionist aims, it is undeniably true that computational models have proved of immense value in studying what might be considered typical human activities, such as conversation as a form of computed information flow driven by natural language functioning as a programming language (van Benthem 1996). I will mention more examples below.

But there is an opposite stream as well. Much of the history of computer science can also be seen as a case of the 'humans striking back'. Here is a mild instance of this phenomenon. Around 1980, Halpern and others started the TARK tradition of enlisting the delicate yet powerful understanding that we have of human agents with knowledge and social activity in support of modeling complex distributed protocols, how they function, what they do and do not achieve, and what might go wrong with them (Fagin et al. 1995). Now this may be a case of using metaphors, but the resulting revival of epistemic logic, broadly conceived, has had widespread repercussions in several disciplines. Likewise, and even much earlier, the development of Artificial Intelligence, though perceived by some as a reductionist replacement exercise, in fact gradually made computer scientists (and others) aware of the amazing subtlety of human behavior and skills in problem solving, knowledge acquisition, and social interaction. A wealth of logical systems arose out of this that also started influencing other areas far beyond computer science, including linguistics and philosophy. And finally, just consider the realities of computing today. What we see all around us are complex societies of humans and machines engaged in new interactive styles of behavior, nowadays even driven by conscious design of 'gaming', and it might even be thought that the real challenge today is understanding what makes this reality tick, rather than abstruse discussions about computing at infinity or in the Milky Way.

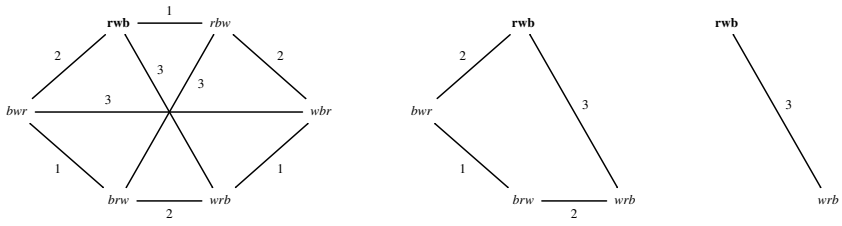
4 Conquering daily life: conversation as computation

Computational models are widely used in the study of human agency today, so much so, that the area of multi-agent systems combines features of computer science, epistemology, and cognitive science (Leyton-Brown and Shoham 2009). Of the vast literature, I mention only one current strand as an illustration: 'dynamic-epistemic logics' (van Ditmarsch et al. 2007, van Benthem 2011).

Conversation and information update Simple games are a good setting for studying communication. Three cards "red", "white", and "blue" are given to three children: 1 gets red, 2 white, and 3 blue. Each child sees his own card, not the others. Now 2 asks 1 "Do you have the blue card?", and the truthful answer comes: "No". Who

knows what now? Here is what seems the correct reasoning. If the question is genuine, player 1 will know the cards after it was asked. After the answer, player 2 knows, too, while 3 still does not. But there is also knowledge about others involved. At the end, all players know that 1 and 2, but not 3, have learnt the cards, and this is even 'common knowledge' between them.[2]

The Cards scenario involves a computational process of state change, whose basic actions are updates shrinking a current range. In the diagrams below, indexed lines indicate an uncertainty for the relevant agents. Informational events then shrink this range stepwise:



The first step is for the presupposition of the question, the second for the answer. In the final model to the right, both players 1 and 2 know the cards, but 3 does not, even though he can see that, in both of his remaining eventualities, 1, 2 have no uncertainties left. □

The geometry of the diagram encodes both knowledge about the facts and knowledge about others: such as 3's knowing that the others know the cards. The latter kind is crucial to social scenarios, holding behavior in place. Indeed, at the end of the scenario, everything described has become common knowledge in the group $\{1, 2, 3\}$.[3]

Dynamic logics of communication 'Dynamic epistemic' logics describing this informa-tion flow, and changes in what agents know from state to state, have been found on the analogy of program logics in computer science. First, what agents know about the facts, or each other, at any given state is described by a standard language of *epistemic logic*:

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_i\varphi$$

[2]This way of understanding the scenario presupposes that questions are sincere as seems reasonable with children. But our methods also cover the possibly insincere scenario.

[3]Cf. Fagin et al. 1995 for all these notions in games and computation.

while the corresponding epistemic models were tuples

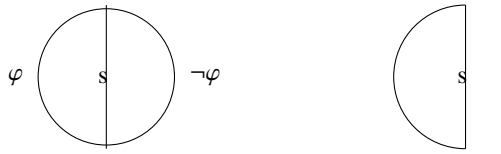
$$M = (W, \{\sim_i | \ i \in G\}, V)$$

with a set of relevant worlds W, accessibility relations $\sim_i$ and a propositional valuation V for atomic facts. Knowledge is then defined as having semantic information:[4]

$$M, s \models K_i\varphi \text{ iff for all worlds } t \sim_i s : M, t \models \varphi$$

Common knowledge $M, s \models C_G\varphi$ is defined as φ's being true for all t reachable from s by finite sequences of $\sim_i$ steps. If necessary, we distinguish an *actual world* in the model.

Update as model change The key idea is now that informational action is model change. The simplest case is a *public announcement* $!_\varphi$ of hard information: learning with total reliability that φ is the case eliminates all current worlds with P false:



From M To $M|\varphi$

We call this *hard information* for its irrevocable character: counter-examples are removed.

This dynamics typically involves truth value change for complex formulas. While an atom p stays true after update (the physical base facts do not change under communication), complex epistemic assertions may change their truth values: before the update $!p$, I did not know that p, afterwards I do. As with imperative programs, this may result in order dependence. A sequence $!\neg K_p; !p$ makes sense, but the permuted $!p; !\neg Kp$ is contradictory.

Public announcement logic The dynamic logic PAL arises by extending the epistemic language with a dynamic modality for public announcements, interpreted as follows:

$$M, s \models [!\varphi]\psi \text{ iff } \textit{ if } M, s \models \varphi, \textit{ then } M|\varphi, s \models \psi$$

[4]These epistemic models encode 'semantic information', a widespread notion in science, though other logical views of information exist (van Benthem and Martinez 2008).

The system of public announcement logic PAL can be axiomatized completely by combining a standard logic for the static epistemic base plus a recursion law for knowledge that holds after update, the basic 'recursion equation' of the system:

$$[!\varphi]K_i\psi \leftrightarrow \varphi \rightarrow K_i(\varphi \rightarrow [!\varphi]\psi)$$

Dynamics of other events Similar systems exist for updating agents' beliefs, defined in terms of truth in the *most plausible* epistemically accessible worlds. Here the variety of dynamic events increases. Beliefs can change under hard information, but also under *soft* information, where $\neg\varphi$-worlds are not eliminated, but made less plausible than φ-worlds. And similar methods again work for events modifying agents' *preferences* (Liu 2011).

Time and program structure Single events are just atomic actions that bunch together to form meaningful larger scenarios. Again, computational ideas are essential. Action and communication involve complex programs with operations of sequential composition: guarded choice IF THEN ELSE, and iteration WHILE DO. Even parallel composition || occurs when people act or speak simultaneously. Here is a well-known illustration:

The Muddy Children "After playing outside, two of a group of three children have mud on their foreheads." They can only see the others, and do not know their own status. Now the Father says: "At least one of you is dirty". He then asks: "Does anyone know if he is dirty?" The Children always answer truthfully. What will happen? As questions and answers repeat, nobody knows in the first round. But in the next round, each muddy child reasons thus: "If I were clean, the one dirty child I see would have seen only clean children, and so she would have known that she was dirty at once. But she did not. So I am dirty, too." This scenario falls within the above update setting, but we do not elaborate here. □

Clearly, there is a program here involving sequence, guarded choice and iteration:

> ! "At least one of you is dirty" ; WHILE not know your status DO (IF not know THEN "say don't know" ELSE "say know")

Temporal limit behavior Another interesting feature is the limit behavior in the puzzle, leading to a stable endpoint where updates have no further effect and agents' knowledge is in equilibrium. In particular, the children have common knowledge of

their status in the limit model $\#(M, \varphi)$ reached by the iterated updates $!\varphi$ of their ignorance assertion. So in the end, this statement 'refutes itself'. In other scenarios, like game solution procedures, the statement announced is 'self-fulfilling', becoming common knowledge in the limit.[5]

Limit features of computation over time can be studied in sophisticated fixed-point logics, but one simple case is just propositional dynamic logic PDL of basic imperative programs. The resulting setting has vast computational power (Miller and Moss 2005): the logic PAL with Kleene iteration of updates is Π^1_1-complete. Van Benthem (2008) has a positive interpretation of high complexity results for logics like these, namely that

conversation has universal computing power: any significant computational problem can be realized as one of conversation planning.

While this looks attractive as an observation about conversation as a paradigm for computation in general, there is a catch. The high complexity resides in the logic of reasoning about conversation, but as discussed in (van Benthem 2011), conversational algorithms themselves might have low complexity as far as computational procedures go.[6]

Our examples and glimpses of wider implications may have shown how computational notions and techniques arise all the way in a basic human activity like conversation. For many further examples of 'communication as computation', we refer to the cited literature.

5 Daily life strikes back: computation as conversation

Let us now reverse the perspective. Starting around 1980, Halpern and his colleagues in what is now sometimes called the TARK community have shown how human metaphors of knowledge and social interaction, if made precise in logical terms, can be a powerful tool for specifying and proving properties of complex protocols for multi-agent systems. The book (Fagin et al. 1995) is a landmark of the resulting program. But the borderline between a metaphor and the real thing may be thin. Increasingly, there seems to be a viable view that computing *is itself* a form of social behavior, mixing action and information much as humans do. Correspondingly, the same formal objects that act as programs for machines are also 'plans for humans. The two essential basic features of human agency then enter our understanding of computation: one is the *knowledge* of agents (perhaps also other attitudes, such as their beliefs and

[5] Limit features of belief revisions over time underlie formal learning theory (Gierasimczuk 2010).

[6] However, this complexity may go down on extended protocol models that constrain the admissible sequences of updates in each world.

preferences), and the other their social *interaction*. In what follows, we take up these two themes separately, showing how they enter our view of computing in natural ways. Our major tool for highlighting these phenomena will be transformations from standard algorithms to knowledge-based social procedures.

6 Epistemizing computational tasks

This section is about what we call the phenomenon of epistemization, the introduction of agents' knowledge at various parts in basic computational tasks.

Epistemizing algorithmic tasks Consider the key planning problem of Graph Reachability (*GR*). Given a graph G with points x, y, is there a chain of arrows from x to y? *GR* can be solved in *Ptime* in the size of G: a quadratic-time algorithm finds a path (Papadimitriou 1994). The same holds for reachability of a point in G satisfying a goal condition φ. The solution algorithm performs two related tasks: determining if a path exists at all, and giving a concrete way or plan for getting from x to y. We will now consider various natural ways of introducing knowledge and information in this setting.

Knowing you made it Suppose an agent is trying to reach a goal region defined by φ, with only limited observation of the terrain. The graph G is now a model $(G, R, \sim)$ with accessibility arrows, but also the earlier epistemic uncertainty links between nodes. It is natural to ask for a plan that will lead you to a point *that you know to be in the goal region* φ. Brafman and Shoham (1993) analyze a robot whose sensors do not tell her exactly where she is. They then add a *knowledge* test to the task, inspecting current nodes to see if we are definitely in the goal region: K_φ . Given the $P - time$ complexity of model checking for epistemic logic, the new search task remains *P-time*.[7]

Epistemizing social tasks Many algorithmic tasks themselves come from social scenarios, witness the area of computational social choice (Endriss and Lang 2006). Here, too, epistemization makes sense. Think of the basic computational task of *merging orderings*. In social choice theory, preferences of individual agents are to be merged into a preference order for the group as a whole. This way of phrasing started with Arrow's Theorem stating that no social choice procedure exists that satisfies some basic postulates of unanimity, monotonicity, context independence, and especially, absence

[7] A general model for epistemic robots relying on possibly limited or defective sensors is proposed in (Su et al. 2005). This approach has led to new 'evidence models' for human agency that are more fine-grained than the standard epistemic models of this paper.

of a 'dictator', an individual whose preference ranking always coincides with that of the group. These specifications are completely non-epistemic, which is somewhat surprising, since much of what we consider essential about democratic decision making has to do with privacy, and what agents may know or not know. But, there is even a mismatch between the usual base conditions and how they are interpreted intuitively in terms of agency. The existence of a dictator is problematic if we think of an individual who can abuse her powers: but for that, she should know that she is a dictator - and perhaps, others should also know (or not know) this. Thus, epistemic rethinking of the very scenario of social choice seems in order, and it is not even clear what a knowledge-based version of the basic theory would look like.

Other algorithmic tasks where similar points can be made occur in *game theory*. Indeed, the move from games of perfect information to games with imperfect information (Osborne and Rubinstein 1994) may be considered a case of epistemization in our sense.

Two aspects of epistemization Our examples show two different aspects of introducing knowledge. One is that the *specifications* of what an algorithmic task is to achieve may come to involve knowledge, like saying we must know we are at the goal, This does not necessarily mean that the algorithm itself has to be epistemic. Many social algorithms are purely physical, such as folding ballot slips, though they do have epistemic effects.

The second step, then, makes the *algorithms themselves* contain knowledge aspects. One obvious place where this happens is test conditions for conditional action. We find it obvious that a computer 'checks' in its registers whether, say, $x = 1$, before performing an IF THEN task: truth and knowledge are easily confused here. But for more complex conditions, such as 'the battery is functioning', we can only perform IF THEN instructions if we *know* which condition holds. And there may be yet more subtle aspects of knowledge involved. Turing (1937) says a machine should know which symbol it is reading, and even which state it is in: a rather human form of introspection.[8]

Epistemic programs Algorithms with conditions that we know to be true or false look like human plans. One format for epistemizing standard algorithms is the *knowledge programs* of Fagin et al. (1995), making actions dependent on conditions like

[8]Turing himself thinks that these epistemic properties are guaranteed by having only finite sets of symbols and states. This is not the notion of knowledge used in this paper, since it seems to refer more to perceptual discrimination (cf. Williamson 2000 on the latter notion in epistemology).

"the agent knows φ" that can always be decided given epistemic introspection.[9] The language of the programs now also explicitly contains our earlier epistemic operators. Knowledge programs make sense in epistemic planning (Bolander and Andersen 2011), and also as definitions for uniform strategies in imperfect information games (van Benthem 2013a).

A related way of epistemizing programs is offered by the earlier dynamic epistemic logics. Public announcements are closely related to 'test actions' that remove all epistemic uncertainty links between φ-worlds and $\neg\varphi$-worlds. The behavior of test actions, and that of many other ubiquitous informational actions, such as questions and answers, can be described in exactly the same logical style as before.

Further aspects of epistemization But once we entangle algorithms and knowledge, many further issues emerge, going beyond just opening a 'knowledge parameter' here and there. For a start, epistemic specifications or programs essentially refer to some agent performing the task, and then, the nature of those agents becomes a factor.

Different types of agent Epistemic algorithms may work for one type of agent but not for another. The literature on dynamic epistemic logic has mainly focused on agents with *Perfect Recall* who remember everything they knew at earlier stages of the process, and who also learn from observation only. But equally important are agents with bounded memory, such as finite automata. Various assumptions of this kind will be reflected in the epistemic logic of action. For instance, Perfect Recall holds for an agent iff the following commutation law for action and knowledge governs its behavior:

$$K[a]\varphi \rightarrow [a]K\varphi$$

This says that, if we know beforehand what an action is going to achieve, we will know its effect afterwards. This is crucial to consciously following a plan, though it can fail in other circumstances.[10] Other axioms that can be written in this language govern the behavior of finite automata. Clearly, such assumptions about agents influence what we can expect an epistemic algorithm to achieve - but I am not aware of any general theory.

Know-how, and knowing a program So far we followed the mainstream of epistemic logic in letting knowledge apply to propositions. But our setting suggests a

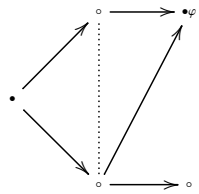
[9] Some of the surprising cognitive algorithms in (Gigerenzer and the ABC Research Group 1999) have this flavor.

[10] I may know that entering this enticing bar will lead to my swift moral downfall, but once I am inside, all such thoughts may have left my mind.

richer view. In addition to propositional *knowing-that*, there is know-how, embodied in algorithms, plans and procedures. Knowing how is the subject of much of our learning, perhaps even more than knowing that. And this know-how is related to an important notion in our natural language, that of knowing an *object*. In our setting, one obvious instance of this is what it means to 'know a program'. There seems to be no unique answer as to what this means, but here is a tie with propositional knowledge that seems relevant.

Suppose that we have an epistemic program or plan, knowing it seems to involve at least some clear grasp of its execution, not just being lucky. Should the agent know the plan to be successful: beforehand, and at all stages of its execution? There are two aspects to this mastery (see van Benthem 2013a) for further discussion and formal results). Suppose that the agent has an epistemic plan: does it follow that she knows its effects? It is easy to see that this is not always so, and hence we might use this as a stronger requirement on epistemized algorithms than we have imposed so far. But there is also another natural aspect to knowing a plan. Suppose that the agent knows now what the plan will achieve, will this knowledge persist over time as the plan is being followed?

Example For an illustration, recall the earlier problem of epistemized graph reachability. Let the agent at the root of the following graph trying to reach a φ-point:



The dotted line says that the agent cannot tell the two intermediate positions apart. A plan that reaches the goal is *Up; Across*. But after the first action, the agent no longer knows where she is, and whether moving *Across* or *Up* will reach the φ-point. □

Much more can be said about when intermediate knowledge of effects does hold, but we merely cite one result discussed in (van Benthem 2013a): agents with Perfect Recall have intermediate knowledge of effects for all knowledge programs in the earlier sense.

From knowing to understanding In recent discussions, more stringent requirements have come up concerning knowing a program, sometimes under the heading of *understanding* what one is doing. In addition to propositional knowledge of effects of a plan,

or parts of it, another key feature is 'robustness': counterfactually knowing the effects of a plan under changed circumstances, or the ability to modify it as needed.[11] And there are yet other tests of understanding a subject, such as a 'talent for zoom': being able to describe a plan at different levels of detail, moving up or down between grain levels as needed.[12]

Epistemization in general We will not explore these issues further here, except to note that epistemizing algorithms seems to open up a rich and interesting area of investigation. Perhaps the first issue on the agenda here should be to *define epistemization* as a general transformation, or a family of these, on traditional algorithms and specifications, whose properties can then be studied as such. The next general issue would be what happens when we systematically epistemize major existing process theories of computation, such as process algebra or game semantics (Bergstra and Smolka 2001, Abramsky 2008).[13] There are bits and pieces in the literature, but I am not aware of general results in this spirit.[14]

Finally, it should be pointed out that epistemization is a more general phenomenon than just adding epistemic logic to the world of algorithms. Epistemic logic is one way of modeling knowledge, based, as we saw, on the notion of semantic information. However, various other views of information make sense for logic and computation, including more fine-grained syntactic accounts of information structure as code (van Benthem and Martinez 2008). The issues that we have raised in this section would still make sense then.

7 Interaction and games

The second essential feature of social agency that we mentioned earlier was multi-agent interaction. The typical paradigm for multi-agent action with many-mind knowledge are games, and what we will do know is look at a 'social transformation' of algorithmic tasks that might be called gamification (van Benthem 2008).

[11] Counterfactual robustness under a natural range of deviant circumstances is also well-known in the philosophical literature on definitions of knowledge (see Nozick 1981, Holliday 2012). In that literature, knowledge gets tied to policies for belief revision and it is an intriguing thought that really understanding a program or algorithm might also have to do with agents' beliefs about it.

[12] Similar issues arise in analyzing what it means for someone to understand a formal proof, and useful intuitions might be drawn from our experience with mathematical practice.

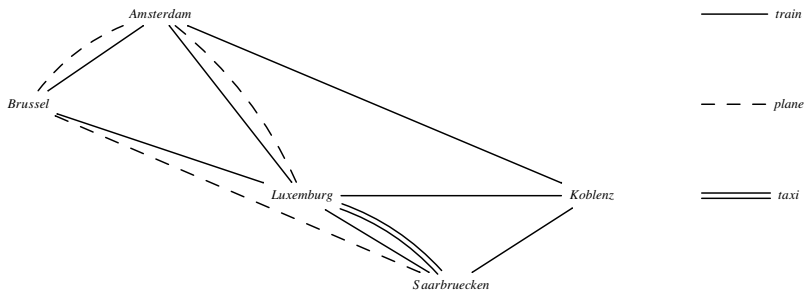
[13] Adding epistemic action to process algebra fits its emphasis on communication channels. Explicit epistemics also makes sense with game semantics of programming languages (Abramsky 2008).

[14] The earlier dynamic epistemic logics seem relevant to this enterprise, and so does the literature on computational complexity of epistemic action logics, cf. Halpern and Vardi 1989.

Multi agent scenarios and knowledge games Reaching a goal and knowing you are there naturally comes with social variants where, say, others should not know where you are. In the 'Moscow Puzzle' (van Ditmarsch 2002), two players must inform each other about the cards they have without letting a third party know the solution. More general knowledge games of this sort have been studied in (Ågotnes and van Ditmarsch 2011). One can think of these as extended semantic explorations of a given epistemic model, assigning different roles to different parties to model more interesting features of inquiry.

Reachability and sabotage Turning algorithms into games involves prying things apart with roles for different agents. Early examples are *logic games* in the style of Lorenzen, Ehrenfeucht, or Hintikka (cf. the survey in van Benthem 2013a), where traditional logical notions now involve a split between a player for truth (proof, analogy,...) versus a player for falsity (counter-model, difference, ...). The strategic game-theoretic powers of players in such games provide a more fine-structured analysis of many classical logical notions.

The sabotage game For a more purely algorithmic example, consider again the earlier Graph Reachability, now in a different scenario with two agents. The following picture gives a travel network between two European capitals of logic and computation:



It is easy to plan trips either way. But what if transportation breaks down, and a malevolent Demon can cancel connections, anywhere in the network? At every stage of our trip, let the Demon first take out one connection, while Traveler then follows a remaining link. This turns a one-agent planning problem into a two-player *sabotage game*. Simple game-theoretic reasoning shows that, from Saarbruecken, a German

Traveler still has a winning strategy, but in Amsterdam, Demon has the winning strategy against the Dutch Traveler.[15]

Sabotage, logic, and complexity The above suggests a transformation for any algorithmic task to a sabotage game with obstructing players. This raises general questions. First, there is logic (van Benthem 2005). One can design languages for these games and players' strategies in terms of "sabotage modalities" on models with accessibility relations R:

$$M, s \models \langle - \rangle \varphi \text{ iff there is a link } (s,t) \text{ in } R \text{ such that } M[R := R - \{(s,t)\}], s \models \varphi$$

In these unusual modal logics, models change in the process of evaluation, and indeed, one can show that sabotage modal logic, though axiomatizable, is undecidable: somehow the computational content of the logic has increased from standard modal logic. Next, there is computational complexity (Rohde 2005). For sabotaged Graph Reachability, the solution complexity of the game jumps from P-time for modal model checking to Pspace-completeness. This takes a polynomial amount of memory space, like Go or Chess.[16]

Still, the game need not always be more complex than the original algorithmic task.

Catch me if you can Now consider another game variant of *GR*. Obstruction could also mean that someone tries to stop me en route: "Starting from an initial position (G, x, y) with me at x and you at y, I move first, then you, and so on. I win if I reach my goal region in some finite number of moves without meeting you. You win in all other cases." This game, too, models realistic situations, such as avoiding some people at some receptions. The difference with the Sabotage game is that the graph remains fixed during the game. Sevenster (2006) proves that its computational complexity stays in P-time.

Adding knowledge and observation again But it also makes sense to combine all these games with our earlier epistemizations. For instance, sabotage as practiced in warfare involves limited observation and partial knowledge. If we turn algorithms into games of *imperfect information*, solution complexity may increase even further. Jones

[15] These games also have interesting interpretations in terms of learning, where a Teacher tries to trap a Student into a certain state of knowledge by blocking all escape routes (Gierasimczuk 2010).

[16] Ron van der Meyden, p.c., has pointed out that, while the sabotage game gamifies the original *reachability* task, there is still an additional issue of how the game solution procedure gamifies the original *algorithm* solving the task. Much remains to be understood at this second level.

(1978) gives a classic complexity jump in such a search task. Sevenster (2006) studies a broader array of epistemized gamified algorithms, linked with the 'IF logic' of Hintikka and Sandu (1997).

Gamification in general The general program behind these examples would be a theory of gamifying algorithmic tasks, and the study of their strategic properties as related to their earlier process properties. We mentioned knowledge games and sabotage games as specific instances - but as we have said, many further examples of successful gamification exist in logic and computer science.[17] A general understanding of this phenomenon might profit from current contacts between logic, computer science, and game theory.

What and how again Some fundamental issues that will play here are related to the central topics of van Benthem (2013a). One of these is the transition from logics of programs to *logics of strategies*. But also, an earlier issue that we raised at the beginning of this paper returns. A fundamental question in the logical foundations of game theory is *when two games are the same*. Answers to this question embody a view of a game as an interactive process, and hence, they embody a view of social computation. One persistent intuition here has been the possibility of simulating strategies in other games inside the current one, sometimes even in the brutal form of copying the same moves. But this computational idea is at the same time an intriguing intuition about the glue of social behavior.[18]

While these issues have pure versions, eventually, we want to look at epistemized ones. This brings us to the theory of *imperfect information games* (Osborne and Rubinstein 1994, Perea 2012). This meshes well with the dynamic epistemic logics that we have mentioned earlier, since they invite explicit analysis of the informational actions taking place during the game.[19] But there is also another dimension to this. Imperfect information games have bona fide solutions in terms of *Nash equilibria in mixed strategies*, letting players play moves with certain probabilities. Thus, perhaps surprisingly, epistemization and gamification may need foundations in terms of mixtures of logic and *probability theory*.[20]

[17] One should also mention the practical uses of gamification in the world of computer games, which seem to have developed very similar aims independently.

[18] Maybe game-theoretic notions of equivalence also have to depend on the *types of agent* playing the games, with their ways of reasoning based on combining belief and preference.

[19] Van Benthem (2013a) develops this theme at length under the heading of *Theory of Play*.

[20] For quite different epistemic aspects of playing games, in terms of required knowledge of strategies, see van Benthem 2013b.

8 Foundations: the three pillars of computation once more

What does the more social perspective on computation sketched here tell us about the original grand questions about computation? We will go in reverse order.

As for the *Turing Test*, the issue of mimicking, or even replacing, humans by machines seems tedious and, despite some unholy attractions, ultimately uninteresting. Given how the world of computation has developed in reality, the real challenge today is understanding the diverse *mixed societies of computers and humans* that have sprung up all around us, and that have vastly increased the behavioral repertoire of humans (and machines).

More tenable today is the original *Church Thesis*. Given the close entanglement of social computation and our use of classical techniques of analysis in logic and complexity theory, we see no need to doubt its "What" answer: the recursive functions seem fine as the extensional view of what can be computed. But this may be the less interesting question eventually, if one's aim is to understand computation. As we said before, what we really want to understand is the "How" question of what constitutes computational behavior. And if we take the social perspective of information and interaction outlined here seriously, then some very fundamental questions are on the table: when are two social processes the same, and how do we factor in the essential role of the agents involved in them? What we really need is a convincing foundational theory of social behavior, and maybe the focus on computation of this paper will be a good way of making progress here.[21]

Finally, let us return to Turing's original contribution. The *Universal Machine* was, and remains, a crucial device for making our thinking about computation sharp, and allowing, for the first time in history, precise mathematical results on the power and limitations of what is computable. Can there be a similar universal format for the behavior produced by social computation in the sense of this paper? We may need a "new Turing" for this, but my guess is that the answer will come in the form of an abstract conceptual analysis of what is really means to be a *game* - beyond the details of current game theory.[22]

References

S. Abramsky. Tutorial on game semantics. LINT Workshop Amsterdam, Department of Computing, Oxford University, 2008.

[21] Admittedly, the lack of convergence to a unique view even in the more restricted area of concurrency may be a source of worry here. But see (Abramsky 2013) for some positive answers.

[22] This discussion paper is a version of an invited lecture at the Manchester ASL Logic Colloquium in the Turing Year 2012, the ESSLLI Summer School in Opole, and several later presentations.

J. van Benthem. *Exploring Logical Dynamics*. CSLI Publications, Stanford, 1996.

J. van Benthem. An essay on sabotage and obstruction. In D. Hutter and W. Stephan, editors, *Mechanizing Mathematical Reasoning*, volume 2605 of *Lecture Notes in Computer Science*, pages 268–276. Springer Verlag, 2005.

J. van Benthem. Computation as conversation. In S. Cooper, B. Lowe, and A. Sorbi, editors, *New Computational Paradigms, Changing Conceptions of What is Computable*, pages 35–58. Springer, New York, 2008.

J. van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, Cambridge UK, 2011.

J. van Benthem. *Logic in Games*. The MIT Press, Cambridge MA, 2013a.

J. van Benthem. Reasoning about strategies. In L. O. B. Coecke and P. Panangaden, editors, *Computation, Logic, Games, and Quantum Foundations*, volume 7860 of *Lecture Notes in Computer Science*, pages 336–347. Springer, 2013b.

J. van Benthem and M. Martinez. The stories of logic and information. In P. Adriaans and J. van Benthem, editors, *Handbook of the Philosophy of Information*, pages 217–280. Elsevier, 2008.

J. A. Bergstra and S. A. Smolka, editors. *Handbook of Process Algebra*. Elsevier Science, 2001.

T. Bolander and B. Andersen. Epistemic planning for single and multi-agent systems. *Journal of Applied and Non-Classical Logics*, 21(1):9–34, 2011.

R. Brafman and Y. Shoham. Towards knowledge-level analysis of motion planning. In *Proceedings AAAI 1993*, pages 670–675, 1993.

S. B. Cooper. Computability theory. In J. van Benthem and A. Gupta, editors, *Logic and Philosophy Today*, volume 1, pages 197–218. College Publications, London, 2011.

S. B. Cooper and A. Sorbi, editors. *New Computational Paradigms, Changing Conceptions of What is Computable*. Springer, 2008.

E. W. Dijkstra. Goto statement considered harmful. *Communications of the ACM*, 11 (3):147–148, 1968.

H. van Ditmarsch. The Russian cards problem: a case study in cryptography with public announcements. In *Proceedings of AWCL 2002 (Australasian Workshop on Computational Logic) Canberra*, pages 47–67, 2002.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, 2007.

W. Thomas, E. Gradel and T. Wilke, editors. *Automata, Logics, and Infinite Games*. Lecture Notes in Computer Science. Springer Verlag, 2002.

P. van Emde Boas. Machine models and simulations. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science, vol A, Algorithms and Complexity*, pages 3–66. Elsevier Science, 1990.

U. Endriss and J. Lang. *Proceedings of the 1st International Workshop on Computational Social Choice (COMSOC-2006), ILLC, University of Amsterdam*. 2006.

G. Gigerenzer and the ABC Research Group. *Simple Heuristics That Make Us Smart*. Oxford University Press, 1999.

N. Gierasimczuk. *Knowing One's Limits. Logical Analysis of Inductive Inference*. PhD Thesis, Institute for Logic, Language and Computation, University of Amsterdam, Dissertation ds-2010-11, 2010.

J. Halpern and M. Vardi. The complexity of reasoning about knowledge and time, I. Lower bounds. *Journal of Computer and System Sciences*, 38(1):195–237, 1989.

J. Hintikka and G. Sandu. *Game-Theoretical Semantics*, pages 361–410. Elsevier, Amsterdam, 1997.

W. Holliday. *Knowing What Follows: Epistemic Closure and Epistemic Logic*. PhD thesis, Department of Philosophy, Stanford University, Appeared in ILLC Dissertation Series DS-2012-09, 2012.

N. D. Jones. Blindfold games are harder than games with perfect information. *Bulletin of the EATCS*, 6:4–7, 1978.

K. Leyton-Brown and Y. Shoham. *Multiagent Systems: Algorithmic, Game Theoretic and Logical Foundations*. Cambidge University Press, Cambridge UK, 2009.

F. Liu. *Dynamic Logic of Preference Change*. Springer, 2011.

J. Miller and L. Moss. The undecidability of iterated modal relativization. *Studia Logica*, 97:373–407, 2005.

R. Nozick. *Philosophical Explanations*. Harvard University Press, Cambridge MA, 1981.

M. Osborne and A. Rubinstein. *A Course in Game Theory*. The MIT Press, 1994.

C. Papadimitriou. *Computational Complexity*. Addison-Wesley, Reading, 1994.

A. Perea. *Epistemic Game Theory: Reasoning and Choice*. Cambidge University Press, Cambridge UK, 2012.

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

P. Rohde. *On Games and Logics over Dynamically Changing Structures*. PhD thesis, Rheinisch-Westfalische Technische Hochschule, Aachen, 2005.

M. Sevenster. *Branches of Imperfect Information: Logic, Games, and Computation*. PhD thesis, Institute for Logic, Language and Computation, University of Amsterdam, ILLC Dissertation series DS-2006-06, 2006.

A. Turing. On computable numbers, with an application to the entscheidungsproblem. In *Proceedings of the London Mathematical Society, series 2*, volume 42, pages 230–265, 1937.

Y. Venema. *Algebras and Co-Algebras*, pages 331–426. Elsevier Science, Amsterdam, 2006.

Y. Venema. Lectures on the modal calculus. Institute for Logic, Language and Computation, University of Amsterdam, 2007.

T. Williamson. *Knowledge and Its Limits*. Oxford University Press, 2000.

T. Ågotnes and H. van Ditmarsch. What will they say? Public announcement games. *Synthese (KRA)*, 179(1):57–85, 2011.