Predictive Theory of Mind Models Based on Public Announcement Logic

Top, Jakob Dirk; Jonker, Catholijn; Verbrugge, Rineke; de Weerd, Harmen

DOI
10.1007/978-3-031-51777-8_6

Publication date
2024

Document Version
Final published version

Published in
Dynamic Logic. New Trends and Applications - 5th International Workshop, DaLi 2023, Revised Selected Papers

Citation (APA)
Top, J. D., Jonker, C., Verbrugge, R., & de Weerd, H. (2024). Predictive Theory of Mind Models Based on Public Announcement Logic. In N. Gierasimczuk, & F. R. Velázquez-Quesada (Eds.), *Dynamic Logic. New Trends and Applications - 5th International Workshop, DaLi 2023, Revised Selected Papers* (pp. 85-103). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 14401 LNCS). Springer. https://doi.org/10.1007/978-3-031-51777-8_6

Predictive Theory of Mind Models Based on Public Announcement Logic

Jakob Dirk Top[1(✉)], Catholijn Jonker[2,3], Rineke Verbrugge[1], and Harmen de Weerd[1]

[1] University of Groningen, Groningen, The Netherlands
{j.d.top,l.c.verbrugge,harmen.de.weerd}@rug.nl
[2] Delft University of Technology, Delft, The Netherlands
c.m.jonker@tudelft.nl
[3] Leiden University, Leiden, The Netherlands

Abstract. Epistemic logic can be used to reason about statements such as 'I know that you know that I know that φ'. In this logic, and its extensions, it is commonly assumed that agents can reason about epistemic statements of arbitrary nesting depth. In contrast, empirical findings on Theory of Mind, the ability to (recursively) reason about mental states of others, show that human recursive reasoning capability has an upper bound.

In the present paper we work towards resolving this disparity by proposing some elements of a logic of *bounded* Theory of Mind, built on Public Announcement Logic. Using this logic, and a statistical method called Random-Effects Bayesian Model Selection, we estimate the distribution of Theory of Mind levels in the participant population of a previous behavioral experiment. Despite not modeling stochastic behavior, we find that approximately three-quarters of participants' decisions can be described using Theory of Mind. In contrast to previous empirical research, our models estimate the majority of participants to be second-order Theory of Mind users.

Keywords: Theory of Mind · Public Announcement Logic · Epistemic Logic · Behavioral Modeling · Random-Effects Bayesian Model Selection · Cognitive Science

1 Introduction

Theory of Mind (ToM) is the ability to attribute and reason about mental states of others, such as knowledge, beliefs, and intentions [10,30]. ToM can be used recursively. For example, if Amy knows that Ben knows that Amy knows that there will be a surprise party, Amy is using second-order ToM (ToM-2), by reasoning about the way Ben is using his theory of mind to reason about her own knowledge; and we are making a third-order attribution to Amy here. ToM is commonly used to navigate social situations, and can improve the outcomes of

N. Gierasimczuk and F. R. Velázquez-Quesada (Eds.): DaLí 2023, LNCS 14401, pp. 85–103, 2024.
https://doi.org/10.1007/978-3-031-51777-8_6

competitive [16,32], cooperative [13,28], and mixed-motive settings [39]. While human ToM capabilities develop over early childhood [41], and can be trained [1,38,40], it is generally found that there is a limit to human recursive ToM use, which often does not exceed level 2 [7,9,12,27], and sometimes fails entirely [23].

Epistemic logic, a variant of modal logic, is used to formalize the kind of recursive knowledge needed for ToM statements of the form 'I know that you know...' [19]. However, epistemic logics and their extensions classically assume logical omniscience, contrary to the commonly found limits on ToM. It has been suggested that these models should incorporate recursive reasoning limits [17,39], and there have been previous attempts to model similar aspects of bounded rationality [8,11,24,31]. The first formal attempt to incorporate ToM-like limitations in epistemic logic appears to be [22], which describes an approach close to our purposes: They define the *epistemic depth* of a formula based on the nesting of its modal operators. However, their approach does not cover Public Announcement Logic (PAL, introduced in Sect. 2.2), which we require for our purposes, and is a general approach that does not define how it can be used to encode the specific attributes of ToM.

While formal methods often do not take into account the ToM limits found in behavioral research, the latter does not regularly employ the tasks and models commonly used in epistemic logic, such as epistemic puzzles. Epistemic puzzles, like the Wise Men puzzle [25], Muddy Children puzzle [14], and the one described in Sect. 2.1, are puzzles where a set of agents, in a partially observable world, have to deduce unobservable facts using the epistemic statements of other agents. In the literature, reproducible experiments using these puzzles, especially ones yielding reusable data, appear sparse (see e.g. [8,18,20]).

The present paper attempts to bridge the gap between logic and (boundedly rational) cognition. We build on the work of Cedegao and colleagues [8] by adding ToM limitations to PAL, which we use to predict the answers of different ToM levels in the game of Aces and Eights (explained in Sect. 2.1). We validate our novel method on the data of Cedegao et al. [8] by using Random-Effects Bayesian Model Selection (RFX-BMS) [33], which we use to estimate the frequencies of different ToM levels among the participants of [8].

In recent work, parallel to ours, Arthaud and Rinard [2] create several logics of public announcements which place a limit on the number of nested knowledge operators an agent can understand. Before we continue, we note some key differences with our work. In [2], any nested knowledge operator increases a formula's depth, whereas we assume that only switching between knowledge operators for *different* agents requires higher ToM [39]. There should be a quantitative difference between recursively reasoning about your own knowledge, and that of others. In [2], a formula $K_a\varphi$ is false if the depth of agent a is lower than that of φ. Our ToM-0 agents act as if there are no relations for other agents. If an agent has no outgoing relations, it vacuously knows everything, so ToM-0 agents know that all other agents know everything. This could be similar to young children without ToM, who may think that their parents are all-knowing [5]. Lastly, we move beyond purely formal methods by fitting our models on human data.

In Sect. 2, we explain the tasks, data, and methods we use for predictive modeling. In Sect. 3, we present the results of our novel predictive modeling

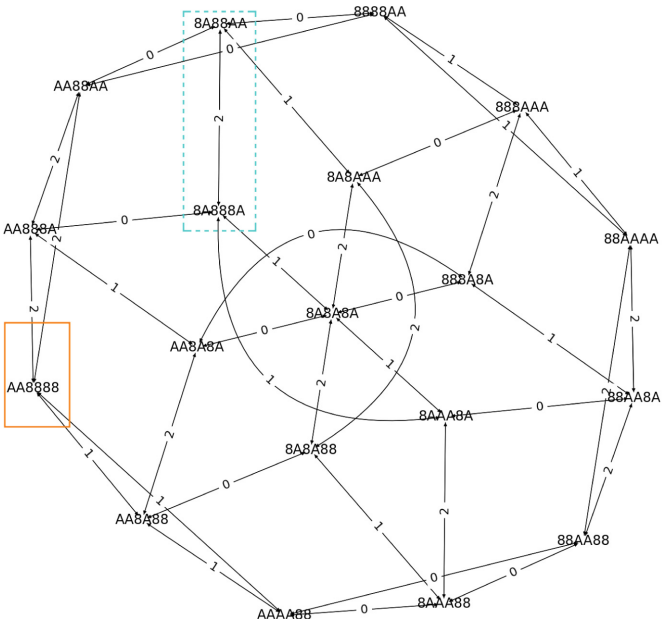

Fig. 1. Model before announcements. Reflexive edges omitted for clarity.

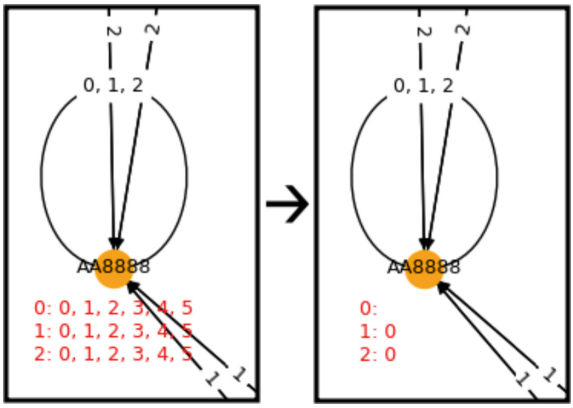

Fig. 2. State *AA*8888 in the ToM model before and after player 0 announces 'I do not know my cards'. This is a close-up of the orange rectangle in Fig. 1, with added ToM levels. Refer to Sect. 2.4 for an in-depth explanation.

method, and compare it to the results we obtain when applying Random-Effects Bayesian Model Selection to the models of [8]. Lastly, in Sect. 4, we discuss our findings and identify possible shortcomings and directions for future work.

2 Methods

Here, Sects. 2.1 through 2.3 describe existing work, leading into our novel work as described in Sects. 2.4 and 2.5.

2.1 The Game of Aces and Eights

Aces and Eights [14] is a three-player epistemic game where each player receives two cards out of a deck of four Aces and four Eights. Each player can only see the four cards held by the other two players. No player can see her own cards or the two remaining cards. Players take turns, in a fixed order, announcing whether or not they know the *ranks* of the cards they are holding — a card's suit does not matter. These announcements provide information that may allow players to work out which cards they have. Players are collectively informed of all these rules, allowing common knowledge of the game rules to arise.[1]

Let us introduce the notation employed throughout this paper. We use 'player 0', 'player 1', and 'player 2' (or, in short, '0', '1' and '2') for the player that makes the first, second, and third announcement each round, respectively. Suppose 0 has two aces (AA), 1 has two eights (88), and 2 has two eights (88). We denote the state of this game as $AA8888$, where the first two symbols are 0's cards, the second two symbols are 1's cards, and the third two symbols are 2's cards. In this state, 0 knows her cards. She sees that all available Eights are held by the other two players, so she must have two Aces. After 0 announces 'I know my cards', 1 and 2 can also know their cards, because they can attribute this reasoning to 0. For holding one Ace and one Eight (or one Eight and one Ace, as order does not matter), we write '8A'.

Cedegao et al. [8] discuss an experiment where each of 306 participants played ten games of Aces and Eights with two computer players that are perfect logical reasoners. Participants were recruited and played online, on the Prolific platform. The order and selection of games varied across participants, but each participant played one game requiring epistemic level 0 (EL-0, see Sect. 2.3) to solve, three games requiring EL-1, and two games each requiring EL-2, EL-3, and EL-4 (retrieved from their code). Participants switched between playing as player 0, 1, and 2 across games. Participants knew the rules and knew that the computer agents gave perfect answers. A game ended if the participant answered 'I know my cards', if the participant answered incorrectly (including answering 'I don't know' when they could have known), or if playing more rounds would not provide more information. Participants responding with 'I know my cards' also had to state the cards they thought they had. Participants were paid \$5 with a \$0.50 bonus for each game correctly solved. Participants were excluded if they failed more than 20% of attention checks, spent more than 87 min, gave impossible responses according to the rules, or had data recording errors. Following [8], this paper only uses the data for the remaining 211 participants.

[1] For solving the game of Aces and Eights, all players also need to be truthful, perfect logical reasoners, and there needs to be common knowledge of this.

2.2 Public Announcement Logic

Public Announcement Logic (PAL) [3,4,29] is an extension of epistemic logic that models how the knowledge of agents changes after public announcements are made. Here, the knowledge of all agents in some epistemic situation is encoded in a *Kripke model* (thus, assuming logical omniscience). A Kripke model can be represented using a directed graph. The graph for Aces and Eights is found in Fig. 1. Each node, or state, is a possible situation, such as the distribution of cards in Aces and Eights. Each edge is labelled with player(s), and indicates uncertainty for those players: A player i edge from state s_1 to s_2 means 'if s_1 is the *true* state (the state corresponding to the actual distribution of cards), then player i considers it possible that s_2 is the true state' (here, we may have $s_1 = s_2$). For example, if 2 sees that 0 has $8A$ and 1 has 88, then 2 considers it possible that she has either $8A$ or AA, so there is a symmetric player 2 edge between $8A888A$ and $8A88AA$, as well as reflexive edges at both states. This situation can be found in Fig. 1, where it is indicated with a cyan, dashed, rectangle (reflexive edges omitted). If, in state s, all outgoing player i edges connect to worlds where i has the same cards, then i knows her cards. An example of this is player 0 in $AA8888$, found in the solid orange rectangle in Fig. 1.

2.3 Bounded Models

Cedegao et al. [8] model an epistemic level l as follows: Take as an agent's *initial states* those states that the agent considers possible based on the game rules and the cards held by the other two players. For example, if agent 1 sees that 0 holds AA and 2 holds $8A$, then agent 1's initial states are $AA888A$ and $AA8A8A$. Modifying Definition 2.32 of [6], the height of a state is defined by induction: the height of *all* initial states is 0, and the states of height $n+1$ are the immediate successors (states that can be reached in one step along any outgoing edge) of states of height n that have not yet been assigned a height smaller than $n+1$. States with height l are marked peripheral states, and their outgoing edges are removed. States with a height exceeding l are removed entirely. When an announcement is made, a bounded model is updated by removing those *non-peripheral* states (and connecting edges) where the announced formula is false. Answers are based on the remaining initial states. Since our models differ from those in [8], we use 'ToM order' when talking about our models, and 'epistemic level' (EL) when talking about the models of [8].

Since all states *other than* the peripheral states have the same relations as the full model, which is an $S5_{(3)}$-model, Cedegao's models allow for paths with an infinite number of switches between *different* agents (e.g., $\ldots(s_{n-1}, s_n) \in R(0), (s_n, s_{n+1}) \in R(1), (s_{n+1}, s_{n+2}) \in R(0), \ldots$). We argue that paths with infinitely many perspective switches are contrary to human recursive ToM limits. Furthermore, an agent with epistemic level 4, playing Aces and Eights, uses the same graph as a logically omniscient agent. In contrast to Cedegao and colleagues [8], we instead attempt to limit the number of recursive reasoning steps an agent can use, as outlined in the next section.

2.4 Theory of Mind Models

This section introduces our novel methods for modeling ToM, in a logic we call TOMPAL. We work in the language $\mathcal{L}_{K[]}(A, P)$, taken directly from [37]:

Definition 1. The language of public announcement logic is inductively defined

$$\mathcal{L}_{K[]}(A, P) \ni \varphi ::= \quad p \quad | \quad \neg\varphi \quad | \quad (\varphi \wedge \varphi) \quad | \quad K_i\varphi \quad | \quad [\varphi]\varphi$$

with $i \in A$, a set of agents, and $p \in P$, a finite set of propositional atoms.

The usual abbreviations are used for $\vee$, $\rightarrow$, and $\leftrightarrow$. For $\neg K_i \neg\varphi$ we use $M_i\varphi$.

We consider that repeated nestings of knowledge operators for the same agent do not require additional ToM levels to be understood (see [39]), and that reasoning about one's own knowledge does not require ToM at all. Instead, we assume only *switching* to the perspective of a different agent requires an additional level of ToM. For example, player 0 needs ToM-2 to reason about the sentence $K_0K_0K_1K_1K_1K_0p$.[2] When an agent switches perspectives, she attributes her own order, minus one, to the other agent. To keep track of this, we modify the definition of models in [36] by adding a map T, as follows:

Definition 2. A ToM model $M = (S, R, V, T)$ consists of a non-empty set of states S, an accessibility function $R : A \rightarrow \mathcal{P}(S \times S)$, a valuation $V : P \rightarrow \mathcal{P}(S)$, where $V(p)$ is the set of states where p is true, and a ToM map $T : S \rightarrow \mathcal{P}(A \times \mathbb{N})$ (with $0 \in \mathbb{N}$), which maps each state to a set of tuples (i, l) with $i \in A$ and $l \in \mathbb{N}$. For $s \in S$, $i \in A$, and $l \in \mathbb{Z}$, the pair $(M, (s, (i, l)))$ is a **perspective state**.

Intuitively, having $(i, l) \in T(s)$ means 'agent i, at ToM order l, has not yet eliminated state s due to new information'. Conversely, $(i, l) \notin T(s)$ means 'agent i, at ToM order l, either due to some previous announcement no longer considers state s to be possible, or did not consider it possible to begin with'.

Visually, to each state in the model found in Fig. 1 we add one row for each player, consisting of the player's name, followed by a colon, followed by

[2] Note that this differs from [11], where the *horizon* of a player i at (M, s) contains all states player i can 'reach' by taking one step along one of her own edges, followed by any number of steps along any agent's edges. Closer to our intentions, but more general, is the notion of *admissibility on* E [22,24].

that player's possible ToM levels, e.g., '0: 0, 1, 2, 3, 4, 5' at state s means $(0,0) \in T(s), (0,1) \in T(s), \ldots, (0,5) \in T(s)$. An example for state $AA8888$ can be found in the leftmost half of Fig. 2. Here, considering it possible that the actual distribution of cards is $AA8888$ is consistent with reasoning at ToM levels 0 through 5 for all players. In our software implementation of Aces and Eights, we ignore ToM levels beyond 5, because these yield identical answers to ToM-5.

A *perspective state* is an epistemic state viewed from the perspective of agent i at ToM order l; such states are used in our semantics. The semantics of TOMPAL are a modification of those in [37] and are as follows:

Definition 3. Assuming a ToM model $M = (S, R, V, T)$, $i \in A$, and $l \in \mathbb{Z}$:

$$
\begin{array}{ll}
M, (s,(i,l)) \models p & \Leftrightarrow s \in V(p) \\
M, (s,(i,l)) \models \neg\varphi & \Leftrightarrow M, (s,(i,l)) \not\models \varphi \\
M, (s,(i,l)) \models \varphi \wedge \psi & \Leftrightarrow M, (s,(i,l)) \models \varphi \text{ and } M, (s,(i,l)) \models \psi \\
\text{for } i = j : M, (s,(i,l)) \models K_j\varphi & \Leftrightarrow M, (t,(j,l)) \models \varphi \text{ for all } (t,(j,l)) \text{ with} \\
 & \quad (s,t) \in R(j) \text{ and } (j,l) \in T(t) \\
\text{for } i \neq j : M, (s,(i,l)) \models K_j\varphi & \Leftrightarrow M, (t,(j,l-1)) \models \varphi \text{ for all } (t,(j,l-1)) \text{ with} \\
 & \quad (s,t) \in R(j) \text{ and } (j,l-1) \in T(t) \\
M, (s,(i,l)) \models [\varphi]\psi & \Leftrightarrow M, (s,(i,l)) \models \varphi \text{ implies } M|\varphi, (s,(i,l)) \models \psi
\end{array}
$$

where the model restriction $M|\varphi = (S, R, V, T')$ is defined as $(i,l) \in T'(s)$ iff $(i,l) \in T(s)$ *and* $[M, (s,(i,l)) \models \varphi$ *or* $[l \leq 0$ and φ contains an operator K_j with $i \neq j]]$.

We make three deviations from the usual semantics for public announcement logic: first, formulas are interpreted at a perspective state $M, (s,(i,l))$. They are true or false from the perspective of a specific agent with a specific ToM order. Secondly, our knowledge operator has two clauses: when an agent reasons about her own knowledge, she does not switch perspectives. When an agent reasons about the knowledge of a different agent, she switches perspectives to the other agent, and attributes her own ToM order, minus one, to the other agent. In doing so, a ToM-0 agent attributes ToM-(-1) to other agents. Since by definition we have $(i,-1) \notin T(s)$ for all i and s, a ToM-0 agent reasons as if there are no outgoing relations for other agents. Lastly, we modify the model restriction such that tuples (i,l) are removed instead of states. A ToM-0 agent cannot switch perspectives, and therefore 'ignores' announcements that she cannot understand because they contain K-operators for other agents.[3]

Next, we show some theorems that capture the properties of TOMPAL. First, we want ToM-0 agents to ignore announcements they do not understand. From a ToM-0 agent's perspective, no tuples are removed due to such announcements:

[3] We use $l = 0$ as the only special case, but for situations other than Aces and Eights we need a more general solution, found in Appendix A. Furthermore, our semantics can be made equivalent to one with the usual knowledge operator if we 'unfold' our models such that we have $R : (A \times \mathbb{N}) \to \mathcal{P}(S \times S)$.

Theorem 1. *If φ contains a K_j operator, then for all $M, (s, (i, 0))$ with $i \neq j$:*

$$M, (s, (i, 0)) \models (\varphi \to \psi) \leftrightarrow [\varphi]\psi.$$

Proof. The key point is showing that $T' = T$ and hence $M|\varphi = M$. Details are left to the reader. □

Secondly, ToM-0 agents should act as if there are no outgoing relations for *other* agents, so we should have:

Theorem 2. *For all $M, (s, (i, 0))$ with $i \neq j$: $M, (s, (i, 0)) \models K_j\varphi$.*

Proof. The key point is that there are no $(t, (j, -1))$ with $(s, t) \in R(j)$ and $(j, -1) \in T(s)$, due to the definition of T. Details are left to the reader. □

Note that Theorem 2 implies that $M, (s, (i, 0)) \models K_j\varphi \wedge K_j\neg\varphi$ when $i \neq j$.

Lastly, there should be no paths which infinitely alternate between *different* agents, as ToM puts a limit on the number of times any agent can switch perspectives:

Theorem 3. *For all non-empty sequences $(M_{j_1}M_{j_2}, \ldots, M_{j_{n-1}}, M_{j_n})$ of M-operators such that $|\{k : j_k \neq j_{k+1}\}| > l \geq 0$, respectively for all $M, (s, (i, l))$ and for all $M, (s, (i, l + 1))$:*

$$\begin{array}{lll} \textit{Clause 1: } M, (s, (i, l)) & \models \neg M_{j_1}M_{j_2} \ldots M_{j_{n-1}}M_{j_n}\psi & \textit{for } i = j_1 \\ \textit{Clause 2: } M, (s, (i, l + 1)) & \models \neg M_{j_1}M_{j_2} \ldots M_{j_{n-1}}M_{j_n}\psi & \textit{for } i \neq j_1 \end{array}$$

Proof. First, we denote $M_{j_1}M_{j_2} \ldots M_{j_{n-1}}M_{j_n}$ as M^n. We rewrite $\neg M^n\psi$ as $K^n\neg\psi$, which, as we prove for all $\psi \in \mathcal{L}_{K[]}$, we rewrite to $K^n\psi$. We prove the theorem through mutual induction over l.

Base case, clause 2: our base case is that for all $M, (s, (i, 0))$ with $i \neq j_1$: $M, (s, (i, 0)) \models K_{j_1} \ldots K_{j_n}\psi$, which is shown in Theorem 2 by taking K_{j_1} as K_j and $K_{j_2} \ldots K_{j_n}\psi$ as φ.

Inductive step from clause 2 to clause 1: our **induction hypothesis** is that for some arbitrary $l \geq 0$, for all M, s, i with $i \neq j_1$: $M, (s, (i, l)) \models K^n\psi$. We have to show that, for some non-empty sequence $(K_i, \ldots, K_i)$, $M, (s, (i, l)) \models K_i \ldots K_iK^n\psi$. For s we write s_1, for $(K_i, \ldots, K_i)$ we write $(K_{i_1}, \ldots, K_{i_m})$. We omit all text after the first 'for all':

$$\begin{array}{lll} M, (s_1, (i, l)) & \models K_{i_1}K_{i_2} \ldots K_{i_m}K^n\varphi & \Leftrightarrow \\ M, (s_2, (i, l)) & \models K_{i_2} \ldots K_{i_m}K^n\varphi \text{ for all } (s_2, (i, l)) \text{ with} & \\ & \quad (s_1, s_2) \in R(i) \text{ and } (i, l) \in T(s_2) & \Leftrightarrow \\ \vdots & \vdots & \vdots \\ M, (s_m, (i, l)) & \models K_{i_m}K^n\varphi \text{ for all } \ldots & \Leftrightarrow \\ M, (s_{m+1}, (i, l)) & \models K^n\varphi \text{ for all } \ldots & \end{array}$$

The latter holds because of our induction hypothesis.

Inductive step from clause 1 to clause 2: our **induction hypothesis** is $M, (s, (i, l)) \models K^n\psi$ for some arbitrary $M, (s, (i, l))$ with $l \geq 0$, and $i = j_1$.

We have to show that for $i \neq k$, $M, (s, (k, l+1)) \models K^n\psi$. Through a series of equivalences, it can be shown that both are equivalent to $M, (t, (j_1, l)) \models K_{j_2} \dots K_n\psi$ for all $(t, (j_1, l))$ with $(s, t) \in R(j_1)$ and $(j_1, l) \in T(t)$.

By starting at our base case for clause 2 and *alternating* between both inductive steps, any instance of the theorem can be constructed. No base case for clause 1 is needed. For all $M, (s, (i, l))$ with $l < 0$, $K^n\psi$ holds vacuously, as by definition $(i, l) \notin T(s)$ and $(i, l-1) \notin T(s)$. □

Aces and Eights. For Aces and Eights, we use $A = \{0, 1, 2\}$ and $P = \{88_0, 8a_0, aa_0, 88_1, 8a_1, aa_1, 88_2, 8a_2, aa_2\}$, where 88_0 means 'agent 0 is holding two eights', $8a_1$ means 'agent 1 is holding an Ace and an Eight', et cetera. S and R are as depicted in Fig. 1. V is as would be expected. For example, $V(aa_0) \cap V(88_1) \cap V(88_2) = \{AA8888\}$. We have $(i, l) \in T(s)$ for all $s \in S$, $i \in A$, and $l \in \mathbb{N}$ (though we do not consider $l > 5$). Agent i announcing 'I know my cards' is a public announcement of $K_i88_i \vee K_i8a_i \vee K_iaa_i$, announcing 'I do not know my cards' is a public announcement of its negation.

Consider state $AA8888$ in the leftmost half of Fig. 2, with for $AA8888$ only $(AA8888, AA8888) \in R(0)$. As an example, we show what happens to this state when agent 0 announces that she does *not* know her cards ($AA8888$ may not be the *true* state). For brevity, we use the simpler announcement 'I do not know that I have two Aces'. We compute $T'(AA8888)$ for $M|\neg K_0aa_0$ (and hence $M|\neg K_0aa_0$ itself). We consider each type of tuples on a case by case basis:

For tuples of the type $(i, 0)$ with $i \neq 0$, the formula contains an operator K_j with $i \neq j$ and $l = 0$, so, by definition, these tuples are not removed.

For tuples of the type (i, l) with $i \neq 0$ and $l > 0$, we have that $l \neq 0$, so we have to check whether $M, (AA8888, (i, l)) \models \neg K_0aa_0$. If not, they are removed. We use a series of equivalences:

$$
\begin{array}{ll}
M, (AA8888, (i, l)) \models \neg K_0aa_0 & \Leftrightarrow \text{(definition of } \neg) \\
M, (AA8888, (i, l)) \not\models K_0aa_0 & \Leftrightarrow \text{(def. of } K) \\
M, (t, (0, l-1)) \not\models aa_0 \text{ for some } (t, (0, l-1)) \text{ with} & \\
\quad (AA8888, t) \in R(0) \text{ and } (0, l-1) \in T(AA8888) & \Leftrightarrow \text{(def. of } R(0)) \\
M, (AA8888, (0, l-1)) \not\models aa_0 \text{ for } (0, l-1) \in T(AA8888). &
\end{array}
$$

We have $(0, 0), (0, 1), \dots, (0, 5) \in T(AA8888)$ and $AA8888 \in V(aa_0)$, so $M, (AA8888, (i, l)) \models \neg K_0aa_0$ is false for any $i \neq 0$ and $l > 0$. Hence, all tuples of the type (i, l) with $i \neq 0$ and $l > 0$ are removed. For similar reasons, all tuples of the type $(0, l)$ for *all* l are also removed. The resulting $T'(AA8888)$ can be found in the rightmost half of Fig. 2.

Answers. With these TOMPAL models, we can model which answer any player i with ToM level l would give, given a distribution of cards (corresponding to state s) and a sequence of previous announcements, as follows: Using the methods previously described in this section, update the model with all previous announcements in order. Then, if exactly one of $M, (s, (i, l)) \models K_i88_i$,

$M, (s, (i, l)) \models K_i 8a_i$, and $M, (s, (i, l)) \models K_i aa_i$ holds, player i answers 'I know my cards', and states the cards she has. In any other case, player i answers that she does not know her cards. Note that this deviates from standard epistemic logic where, if there are no outgoing edges for an agent i, all statements of the type 'agent i knows φ' are true, whereas all statements 'agent i does not know φ' are false. Recall from Sect. 1 that we will use TOMPAL to predict the answers and usage of different ToM levels in [8]'s data of Aces and Eights. To be able to employ our statistical methods, we need our models to give single answers. Not only is 'I do not know' the most common answer in the data, but it is also an intuitively good response when you consider nothing to be possible.

2.5 Random-Effects Bayesian Model Selection

Random-Effects Bayesian Model Selection (RFX-BMS) is a statistical method that estimates the frequencies of a set of strategies occurring in a population. Whereas fixed-effects Bayesian model selection methods assume there is a single strategy which best fits all participants, RFX-BMS assumes each subject was drawn from a fixed distribution of strategies, and estimates this distribution. Unlike Maximum Likelihood Estimation, RFX-BMS allows us to make more general claims about this distribution, and is robust to small differences between participants and strategies [9,33,38]. In our case, we estimate the frequencies of ToM levels in the participant population of Cedegao et al. [8]. RFX-BMS uses equation (14) of [33], which maximizes the log-likelihood of each participant using each ToM level by iteratively updating the strategy frequencies until convergence. This log-likelihood is $n(1-\varepsilon) \cdot \ln(1-\varepsilon) + n\varepsilon \cdot \ln(p \cdot \varepsilon)$, where a ToM level's error rate ε for a participant is its number of incoherent predictions for that participant, divided by n, the total number of decision points of the participant. A predicted answer is *coherent* if it is the same as the participant's answer, otherwise it is *incoherent*. A *decision point* is a turn in a game where a participant has to give an answer. The parameter p is a penalty coefficient, which is applied when a participant does *not* follow a certain ToM level, but *does* match its actions. We set it to 0.5. Predicted answers are generated as described at the end of Sect. 2.4. We deviate from [8], where models are fitted to full games instead of decision points. After all, participants can have multiple decision points in each game (one for each round).

In addition to ToM levels 0 through 5, we also fit a *random model*. We determine the best fitting random model by considering that each player guesses among the four options with a fixed but personal probability. The log-likelihood for the random model is

$$\sum_{a \in Ans} a \cdot \ln(\frac{a}{n})$$

where n is the total number of decision points, and $Ans = (k_\neg, k_{88}, k_{8A}, k_{AA})$ is a list of numbers, where we define $k_\neg$ as the number of times the participant answered 'I do not know my cards', k_{88} as the number of times the participant answered 'I know I have two Eights', et cetera.

Given these likelihoods, RFX-BMS estimates a vector α, containing one element for each ToM level and an additional element for the random model.[4]

3 Results

In Sect. 3.1, we explore the use of RFX-BMS by combining it with the epistemically bounded models of Cedegao and colleagues [8], as outlined in Sect. 2.3. In Sect. 3.2, we use the TOMPAL models introduced in Sect. 2.4 as models in RFX-BMS (as described in Sect. 2.5), which we use to predict the frequencies of each ToM level in the data of [8].

3.1 Predicted Epistemic Levels of Participants

Before employing our novel models, we validate the use of RFX-BMS by using it to estimate the relative frequencies of epistemic levels for subjects in [8] by using as model a non-stochastic version of SUWEB, the best-fitting model in [8], which employs the bounded models described in Sect. 2.3. SUWEB models have an *update probability*, the probability with which a state is removed after an announcement, and a *noise* parameter, the probability of the model guessing 'I know' when it does *not* know. We set these to 1 and 0, respectively. When SUWEB considers no states to be possible, it answers 'I know' or 'I don't know'

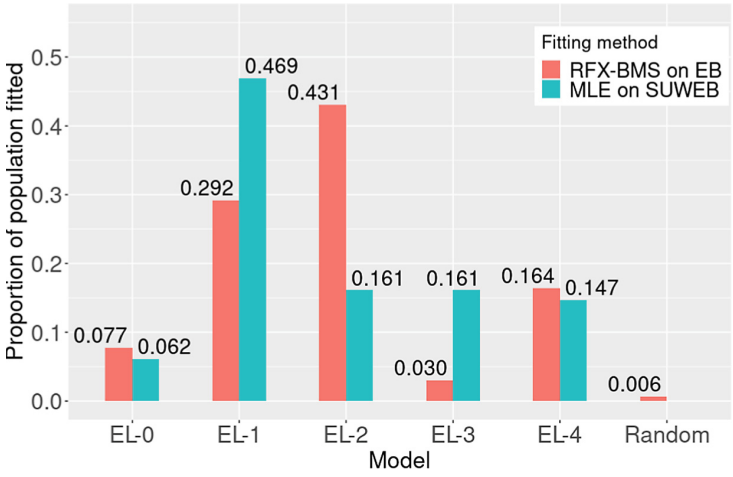

Fig. 3. In red, relative frequencies of each epistemic level and the random model as predicted by RFX-BMS, for [8]'s data, using bounded models. In blue, the original fit of [8]'s stochastic SUWEB models, which also are bounded models. (Color figure online)

[4] All code used for this article can be found at https://github.com/jdtoprug/EpistemicToMProject and doi: 10.5281/zenodo.8382660. Note that we implemented the model updates needed for Aces and Eights and related games, and not a general logical framework.

with equal probability. In these cases we have this non-stochastic SUWEB answer 'I don't know' instead. We combine this non-stochastic SUWEB with RFX-BMS as described in Sect. 2.5, in order to estimate the relative frequencies of each epistemic level, as well as the random model, across all 211 participants.

The predicted frequencies of epistemic levels in the population can be found in Fig. 3. Here, the blue bars are the original fit of [8], obtained by using Maximum Likelihood Estimation to estimate SUWEB's parameters and the epistemic level (EL) of each participant. The red bars are the predictions of RFX-BMS on non-stochastic SUWEB (EB), as explained in the previous paragraph. As a reminder, both red and blue bars use bounded models as explained in Sect. 2.3. For non-stochastic SUWEB, less than 1% of the population is classified as using the random model, which validates the epistemically bounded models presented in [8]. Over 40% of the population is classified as EL-2. This differs from the original SUWEB, which fits over 45% of participants to EL-1. We believe this is because many of the games that reportedly require levels 3 or 4 can be correctly solved by simply answering 'I don't know' in every round, which our non-stochastic EL-2 models consistently do, as opposed to the original SUWEB models, which sometimes answer 'I know' due to noise. Many participants that were fitted as EL-3 or EL-4 can be reclassified as EL-2 users who use this heuristic. For non-stochastic models, update probabilities are 1, which should make higher-level behavior less similar to lower-level behavior, as it causes models to say 'I don't know' less frequently. Zero noise may also decrease similarity between models, as noisy models are less likely to reach later rounds, where levels can be distinguished. These effects should be reflected in our findings.[5]

3.2 Predicted ToM Levels of Participants

In this section, we employ the same methods as described in Sect. 3.1, using our ToM models as described in Sect. 2.4, instead of [8]'s bounded models.

The predicted frequencies of ToM levels in the population can be found in Fig. 4. Less than 1% of the population is classified as using the random model, which shows that participant behavior is better described as ToM reasoning as described in Sect. 2.4 than it is described as guessing. Over 35% of the population is predicted to use ToM-2. A surprising result is the peak at ToM-5: it turns out that RFX-BMS estimates that 14% of the population fits ToM-5 better than any other ToM level. This is not dissimilar to [8], where 15% of participants is fitted to epistemic level 4 (the rightmost blue bar in Fig. 3). In our models, in order to solve *all* games, ToM-5 is needed, whereas in [8], non-stochastic EL-4 accomplishes the same.

When comparing the RFX-BMS results for the epistemically bounded and ToM models, we see that the estimated frequency of ToM-2 users is lower than that of EL-2 users. We believe this is because there are some games where non-stochastic EL-2 correctly answers 'I do not know my cards' due to becoming

[5] We cannot test these predictions as we do not have access to the computational power required to fit the SUWEB model of [8] in a reasonable amount of time.

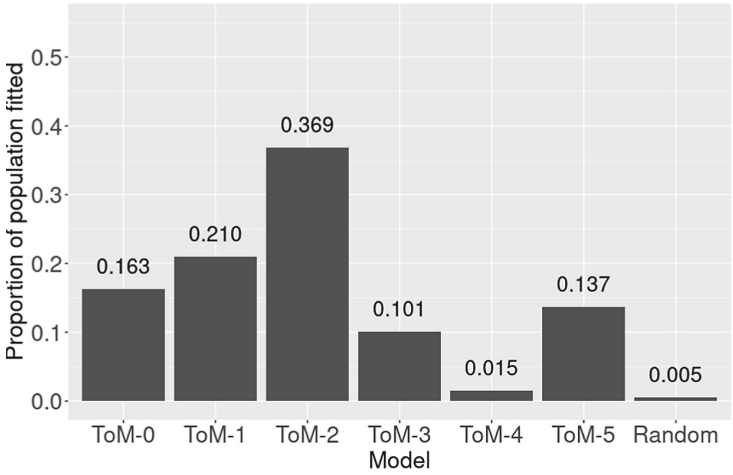

Fig. 4. Relative frequencies of each ToM level and the random model as predicted by RFX-BMS, for the data of [8], using ToM models.

'confused' and removing all non-peripheral nodes, whereas our ToM-2 models incorrectly answer 'I know my cards' due to mistakenly attributing ToM-1 to the other players (which are ToM-5). For one such example, see Appendix B.

To see how well, on average, our models' predictions correspond to participant behavior, the distribution of coherence across participants can be found in Fig. 5. A participant's coherence is the number of coherent predictions for that participant's best-fitting model, divided by that participant's total number of decision points. Coherence is at least .736 for over half of the participants, and only 15 participants have a coherence of 0.5 or lower. There are only six participants where the random model has the best coherence, which are indicated using an ×. Upon visual inspection of the data for the low-coherence outliers, it

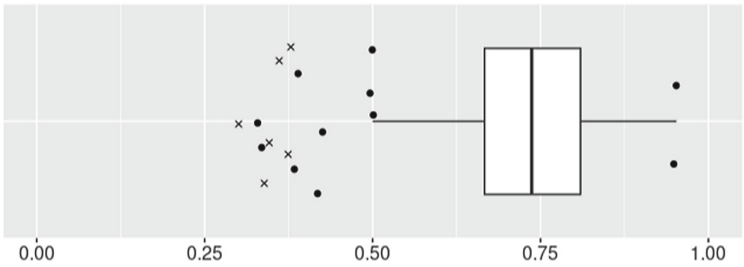

Fig. 5. Distribution of 1-ε for the best-fitting ToM levels for each of the 211 participants. Mean 0.723, median 0.737, IQR 0.143. Crosses indicate participants for whom the random model fits better than any of the ToM models. The vertical axis has no meaning and is used to separate data points for improved readability.

seems that these participants frequently answered 'I know my cards' when they could not, which our ToM models never do.

4 Discussion/Conclusion

Humans do not have the logical omniscience that modal logics based on Kripke models presuppose [21,39]. For one, human ToM is limited [23,27]. In this paper we propose a novel method of representing ToM limitations in Public Announcement Logic, building on the work of Cedegao et al. [8] (see also [22] and [11]). We use Random-Effects Bayesian Model Selection to predict the frequencies of ToM levels in the data of [8], and find some striking differences and similarities when comparing the estimates of ToM and epistemically bounded models.

We predict the majority of the participants of Cedegao and colleagues [8] to be using ToM-2, possibly bolstered by the heuristic of answering 'I don't know' in cases where a random answer would be given in the SUWEB model of [8]. For the latter, the majority of participants is fitted as Cedegao et al.'s epistemic level 1 (EL-1). We believe this difference is due to SUWEB's stochasticity, as well as EL-2 and higher overestimating human (recursive) reasoning capabilities. Our results are a refinement that show that participants are better described as ToM-2 than ToM-1, where the former lies between non-stochastic EL-1 and EL-2 in terms of game-solving capabilities. Our novel method also predicts a portion of participants to use ToM-5. However, since participants can solve many higher-level puzzles by always answering 'I don't know', it is difficult to distinguish higher-order reasoning from heuristics, so it is important to emphasize that the participants themselves may not necessarily be using fifth-order reasoning. We recommend employing games where to be correct, one must eventually answer 'I know' as diagnostic cases in future research.

A drawback of our approach is that we do not consider deviations from our ToM models' predictions, even though some participants exhibit clear guessing strategies where they answer 'I know my cards' when they cannot know. Also, our models do not consider the possibility that agents may attribute different levels of ToM reasoning to other players. For example, a ToM-2 model attributes ToM-1 to every other agent, and does not consider the possibility that one agent is using ToM-0, whereas another agent is using ToM-1. Furthermore, we assume that participants use a single ToM level throughout the experiment, but it could be possible that some participants switch ToM levels between games or even rounds. Lastly, recall that our models answer 'I do not know my cards' when there are no outgoing edges. When this answer is changed to a different answer, or any random distribution over the four answers, we find that mean coherence never drops under 0.72. However, we assume that all participants use the same strategy in such cases, whereas a richer model could try to find the best-fitting answering behavior for each player. In future work it may be possible to incorporate all these behaviors in our models, though even without covering these cases our models have a mean 0.723 coherence - a decent fit, and an indication that participant behavior can, at least partially, be described using our ToM models.

In Sect. 2.5, we calculate the log-likelihood of a ToM level fitting a participant by introducing a penalty for deviating from our models, the value of which strongly affects the relative fit of the random model compared to ToM models. Random-Effects Bayesian Model Selection must assign each participant to one of our defined models. Though we included ToM and random models, there may be other models that fit even better. For example, participants may be using a representation similar to the number triangles in [15], they may be generalizing such as the participants in [18], or they may be using other strategies. More research and data is needed to find all relevant behavioral features. Eye-tracking data could be used to distinguish between strategies, allowing for more accurate logically inspired models [26,34,35]. These models need not be based on formal logics: we also encourage cognitive scientists to model higher-order ToM in Aces and Eights. Nonetheless, we demonstrate that a large part of participant behavior can be attributed to ToM limitations as represented in our models.

Acknowledgements. This research was funded by the project 'Hybrid Intelligence: Augmenting Human Intellect', a 10-year Gravitation programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022. Lastly, we would like to thank our four anonymous reviewers and prof. dr. Hans van Ditmarsch for providing us with helpful comments, suggestions, and discussion.

Appendix A

This appendix describes how to extend our work beyond Aces and Eights.

In [22], concatenation of sequences is defined: $e \circ e' = (i_1, \ldots, i_m, j_1, \ldots, j_k)$ for $e = (i_1, \ldots, i_m)$, $e' = (j_1, \ldots, j_k)$. The empty sequence is ϵ, and $e \circ \epsilon = \epsilon \circ e = e$.

The epistemic depth $\delta(F)$ of a formula F is inductively defined as follows:

D0: $\delta(p) = \{\epsilon\}$ for any $p \in P$;
D1: $\delta(\neg F) = \delta(F)$;
D2: $\delta(F \to G) = \delta(F) \cup \delta(G)$;
D3: $\delta(\wedge\Phi) = \delta(\vee\Phi) = \cup_{F \in \Phi} \delta(F)$;
D4: $\delta(K_i(F)) = \{(i) \circ e : e \in \delta(F)\}$.
D5: $\delta([F]G) = \{f \circ e : e \in \delta(F), f \in \delta(G)\}$

We added D5, which is not present in [22]. Moving to novel work, we define the ToM structure $\mathcal{T}_{(p,l)}$, with $p \in A$ and $l \in \mathbb{N}$ inductively as follows:

Base Case: $e \in \mathcal{T}_{(p,l)}$ for every $e = (i_1, \ldots, i_m)$ where $0 \leq m \leq l$, and for every $i_j \in e$ we have that $i_j \in A$ and [if $0 < j < m$, then $i_j \neq i_{j+1}$]. If $m \leq 0$ then $e = \epsilon$.
Inductive Step 1: If $e \in \mathcal{T}_{(p,l)}$ and $l \geq 0$, then $(p) \circ e \in \mathcal{T}_{(p,l)}$
Inductive Step 2: If, for any e_1, i, e_2; $e_1 \circ ((i) \circ e_2) \in \mathcal{T}_{(p,l)}$, then $(e_1 \circ (i)) \circ ((i) \circ e_2) \in \mathcal{T}_{(p,l)}$

Our base case corresponds to our requirement that the number of 'perspective switches' is limited by an agent's ToM order. Inductive steps 1 and 2 correspond to not switching perspectives, not requiring additional ToM.

For zero or more repetitions of i we write i^*. As an example, consider $A = \{0, 1\}$. Then, $\mathcal{T}_{(0,2)} = \{\epsilon, (0^*), (1^*), (0^*, 1^*), (1^*, 0^*), (0^*, 1^*, 0^*)\}$.

We then modify our semantic definition of $[\varphi]\psi$ in Definition 3:

$$M, (s, (i, l)) \models [\varphi]\psi \quad \Leftrightarrow M, (s, (i, l)) \models \varphi \text{ implies } M|\varphi, (s, (i, l)) \models \psi$$

where we define the model restriction $M|\varphi = (S, R, V, T')$ with $(i, l) \in T'(s)$ iff $(i, l) \in T(s)$ *and* $[M, (s, (i, l)) \models \varphi$ *or* $[\delta(\varphi) \not\subseteq \mathcal{T}_{(i,l)}]]$.

Note that $\delta(\varphi) \not\subseteq \mathcal{T}_{(i,0)}$ is equivalent to "φ contains an operator K_j with $i \neq j$", as $\mathcal{T}_{(i,0)} = \{\epsilon, (i^*)\}$. With this substitution, our proofs for Theorems 1–3 hold, and our models can be used with any announcements.

Appendix B

There are two games where non-stochastic EL-2 answers correctly whereas our ToM-2 models answer incorrectly.[6] In both of these, the participant is player 0. The distribution of cards in these games is $AA8A88$ and $8A8AAA$. For the

Table 1. Tuples at each relevant state during a series of announcements.

AA8888	**AA8A88**	AAAA88	8AAA88	88AA88	8A8A88	next
0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	
1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	0: $k\neg$
2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	
0:	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	
1: 0	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1: $k\neg$
2: 0	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	
0:	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	0: 0,1,2,3,4,5	0: 0	0: 0,1,2,3,4,5	
1: 0	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1: 0,1,2,3,4,5	1:	1: 0,1,2,3,4,5	2: $k\neg$
2: 0	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	2: 0,1,2,3,4,5	2: 0	2: 0,1,2,3,4,5	
0:	0: 0,1,2,3,4,5	0: 0	0: 0,1,2,3,4,5	0: 0	0: 0,1,2,3,4,5	
1: 0	1: 0,1,2,3,4,5	1: 0	1: 0,1,2,3,4,5	1:	1: 0,1,2,3,4,5	0: $k\neg$
2: 0	2: 0,1,2,3,4,5	2:	2: 0,1,2,3,4,5	2: 0	2: 0,1,2,3,4,5	
0:	0: 0,1,2,3,4,5	0: 0	0: 0	0: 0	0: 0,1,2,3,4,5	
1: 0	1: 0,1,2,3,4,5	1: 0	1: 0,1	1:	1: 0,1,2,3,4,5	1: k
2: 0	2: 0,1,2,3,4,5	2:	2: 0,1	2: 0	2: 0,1,2,3,4,5	
0:	0: 0, 2,3,4,5	0: 0	0: 0	0: 0	0: 0, 3,4,5	
1:	1: 1,2,3,4,5	1:	1:	1:	1: 2,3,4,5	2: $k\neg$
2: 0	2: 0, 2,3,4,5	2:	2: 0	2: 0	2: 0, 3,4,5	
0:	0: 0, 2, 4,5	0: 0	0: 0	0: 0	0: 0, 3,4,5	
1:	1: 1,2, 4,5	1:	1:	1:	1: 2,3,4,5	
2: 0	2: 0, 3,4,5	2:	2: 0	2: 0	2: 0, 3,4,5	

[6] Because knowledge can be false, using 'knowledge' and K may not be entirely accurate. We use it because the model for Aces and Eights is $S5$, but for future work we recommend using 'beliefs' and B.

former, we show the removal of tuples after each announcement in Table 1, where each column is a relevant state, and each row corresponds to an announcement. Column ordering corresponds to the order of states in Fig. 1. The rightmost column shows the next announcement, where the index denotes the player, k is 'I know my cards', and $k\neg$ is 'I do not know my cards'. Tuples that will be removed after the next announcement are red. After six announcements, player 0 at ToM-2 will incorrectly answer 'I know my cards', whereas at ToM-5 she will answer 'I do not know my cards', which is the correct answer. When working through the example, it is recommended to use Fig. 1 as a companion.

References

1. Arslan, B., Verbrugge, R., Taatgen, N., Hollebrandse, B.: Accelerating the development of second-order false belief reasoning: a training study with different feedback methods. Child Dev. **91**(1), 249–270 (2020). https://doi.org/10.1111/cdev.13186
2. Arthaud, F., Rinard, M.: Depth-bounded epistemic logic. In: Verbrugge, L.C. (ed.) Proceedings of the 19th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 23), pp. 46–65 (2023). https://doi.org/10.4204/EPTCS.379.7
3. Baltag, A., Moss, L.S., Solecki, S.: The logic of public announcements, common knowledge, and private suspicions. In: Gilboa, I. (ed.) Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98), pp. 43–46 (1998)
4. Baltag, A., Moss, L.S., Solecki, S.: The logic of public announcements, common knowledge, and private suspicions. In: Arló-Costa, H., Hendricks, V.F., van Benthem, J. (eds.) Readings in Formal Epistemology. SGTP, vol. 1, pp. 773–812. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-20451-2_38
5. Barrett, J.L., Richert, R.A., Driesenga, A.: God's beliefs versus mother's: The development of nonhuman agent concepts. Child Dev. **72**(1), 50–65 (2001). https://doi.org/10.1111/1467-8624.00265
6. Blackburn, P., De Rijke, M., Venema, Y.: Modal logic (Cambridge tracts in theoretical computer science no. 53). Cambridge University Press (2001)
7. Camerer, C.F., Ho, T.H., Chong, J.K.: A cognitive hierarchy model of games. Q. J. Econ. **119**(3), 861–898 (2004). https://doi.org/10.1162/0033553041502225
8. Cedegao, Z., Ham, H., Holliday, W.H.: Does Amy know Ben knows you know your cards? A computational model of higher-order epistemic reasoning. In: Proceedings of the 43th Annual Meeting of the Cognitive Science Society, pp. 2588–2594 (2021)
9. De Weerd, H., Diepgrond, D., Verbrugge, R.: Estimating the use of higher-order theory of mind using computational agents. BE J. Theor. Econ. **18**(2) (2018). https://doi.org/10.1515/bejte-2016-0184
10. De Weerd, H., Verbrugge, L.C., Verheij, B.: How much does it help to know what she knows you know? An agent-based simulation study. Artif. Intell. **199–200**, 67–92 (2013). https://doi.org/10.1016/j.artint.2013.05.004
11. Dégremont, C., Kurzen, L., Szymanik, J.: Exploring the tractability border in epistemic tasks. Synthese **191**(3), 371–408 (2014). https://doi.org/10.1007/s11229-012-0215-7
12. Devaine, M., Hollard, G., Daunizeau, J.: The social Bayesian brain: Does mentalizing make a difference when we learn? PLoS Comput. Biol. **10**(12), e1003992 (2014). https://doi.org/10.1371/journal.pcbi.1003992

13. Etel, E., Slaughter, V.: Theory of mind and peer cooperation in two play contexts. J. Appl. Dev. Psychol. **60**, 87–95 (2019). https://doi.org/10.1016/j.appdev.2018.11.004
14. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.: Reasoning About Knowledge. MIT Press, Cambridge (1995)
15. Gierasimczuk, N., Szymanik, J.: A note on a generalization of the muddy children puzzle. In: Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge, pp. 257–264 (2011). https://doi.org/10.1145/2000378.2000409
16. Goodie, A.S., Doshi, P., Young, D.L.: Levels of theory-of-mind reasoning in competitive games. J. Behav. Decis. Mak. **25**(1), 95–108 (2012). https://doi.org/10.1002/bdm.717
17. Hall-Partee, B.: Semantics-mathematics or psychology? In: Bäuerle, R., Egli, U., Von Stechow, A. (eds.) Semantics from Different Points of View, SSLC, vol. 6, pp. 1–14. Springer, Berlin, Heidelberg (1979). https://doi.org/10.1007/978-3-642-67458-7_1
18. Hayashi, H.: Possibility of solving complex problems by recursive thinking. Jpn. J. Psychol. **73**(2), 179–185 (2002). https://doi.org/10.4992/jjpsy.73.179
19. Hintikka, J.: Knowledge and Belief: An Introduction to the Logic of the Two Notions. Cornell University Press, Ithaca, NY, USA (1962)
20. Jonker, C.M., Treur, J.: Modelling the dynamics of reasoning processes: Reasoning by assumption. Cogn. Syst. Res. **4**(2), 119–136 (2003). https://doi.org/10.1016/S1389-0417(02)00102-X
21. Kahneman, D., Slovic, P., Tversky, A.: Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge (1982)
22. Kaneko, M., Suzuki, N.Y.: Epistemic logic of shallow depths and game theoretical applications. In: Advances In Modal Logic, vol. 3, pp. 279–298. World Scientific (2002). https://doi.org/10.1142/9789812776471_0015
23. Keysar, B., Lin, S., Barr, D.J.: Limits on theory of mind use in adults. Cognition **89**(1), 25–41 (2003). https://doi.org/10.1016/S0010-0277(03)00064-7
24. Kline, J.J.: Evaluations of epistemic components for resolving the muddy children puzzle. Econ. Theor. **53**(1), 61–83 (2013). https://doi.org/10.1007/s00199-012-0735-x
25. McCarthy, J.: Formalization of two puzzles involving knowledge. Formalizing Common Sense: Papers by John McCarthy, pp. 158–166 (1990)
26. Meijering, B., van Rijn, H., Taatgen, N.A., Verbrugge, R.: What eye movements can tell about theory of mind in a strategic game. PLoS ONE **7**(9), 1–8 (2012). https://doi.org/10.1371/journal.pone.0045961
27. Nagel, R.: Unraveling in guessing games: An experimental study. Am. Econ. Rev. **85**(5), 1313–1326 (1995)
28. Paal, T., Bereczkei, T.: Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. Pers. Individ. Differ. **43**(3), 541–551 (2007). https://doi.org/10.1016/j.paid.2006.12.021
29. Plaza, J.: Logics of public announcements. In: Emrich, M., Pfeifer, M., Hadzikadic, M., Ras, Z. (eds.) Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems: Poster Session Program, pp. 201–216. Oak Ridge National Laboratory (1989)
30. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behav. Brain Sci. **1**(4), 515–526 (1978). https://doi.org/10.1017/S0140525X00076512
31. Solaki, A.: The effort of reasoning: modelling the inference steps of boundedly rational agents. J. Log. Lang. Inform. **31**(4), 529–553 (2022). https://doi.org/10.1007/s10849-022-09367-w

32. Stahl, D.O., II., Wilson, P.W.: Experimental evidence on players' models of other players. J. Econ. Behav. Organ. **25**(3), 309–327 (1994). https://doi.org/10.1016/0167-2681(94)90103-1
33. Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J.: Bayesian model selection for group studies. Neuroimage **46**(4), 1004–1017 (2009). https://doi.org/10.1016/j.neuroimage.2009.03.025
34. Top, J.D., Verbrugge, R., Ghosh, S.: An automated method for building cognitive models for turn-based games from a strategy logic. Games **9**(3), 44 (2018). https://doi.org/10.3390/g9030044
35. Top, J.D., Verbrugge, R., Ghosh, S.: Automatically translating logical strategy formulas into cognitive models. In: 16th International Conference on Cognitive Modelling, pp. 182–187 (2018)
36. Van Ditmarsch, H.: Dynamics of lying. Synthese **191**(5), 745–777 (2014)
37. Van Ditmarsch, H., van der Hoek, W., Kooi, B.: Dynamic Epistemic Logic, Synthese Library, vol. 337. Springer Science & Business Media, Dordrecht, Netherlands (2007). https://doi.org/10.1007/978-1-4020-5839-4
38. Veltman, K., de Weerd, H., Verbrugge, R.: Training the use of theory of mind using artificial agents. J. Multimodal User Interfaces **13**(1), 3–18 (2019). https://doi.org/10.1007/s12193-018-0287-x
39. Verbrugge, R.: Logic and social cognition: The facts matter, and so do computational models. J. Philos. Log. **38**(6), 649–680 (2009). https://doi.org/10.1007/s10992-009-9115-9
40. Verbrugge, R., Meijering, B., Wierda, S., Van Rijn, H., Taatgen, N.: Stepwise training supports strategic second-order theory of mind in turn-taking games. Judgm. Decis. Mak. **13**(1), 79–98 (2018). https://doi.org/10.1017/S1930297500008846
41. Wimmer, H., Perner, J.: Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition **13**(1), 103–128 (1983). https://doi.org/10.1016/0010-0277(83)90004-5