Reasoning about Reasoning about Reasoning: From Logic to the Lab

KR 2023, September 2-8, 2023, Rhodos

Rineke Verbrugge
Department of Artificial Intelligence
Bernoulli Institute of Math, CS and AI

university of groningen

Trying to put yourself in their shoes

What is Theory of Mind (ToM)?

The ability to reason about *mental states* of others

This may concern their beliefs, thoughts, knowledge, intentions

People use it to explain, predict and manipulate behavior of others

People apply it recursively:
higher-order theory of mind

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *BBS, 1*,515-526
Dennett, D. *Intentional Systems*. Cambridge (MA), MIT Press, 1989

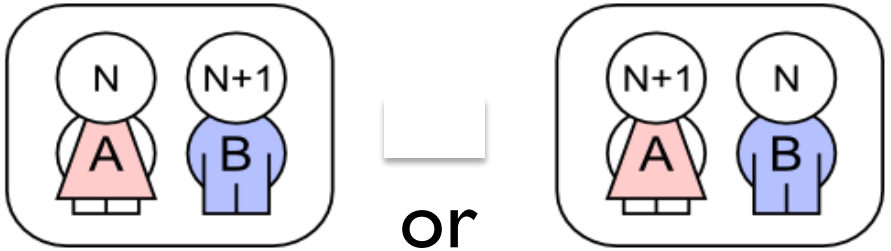
I. The orders of theory of mind

p: Ann published a novel under pseudonym
(zero-order)

- *first-order attribution:*
  *"Bob **knows** that* p"
  $K_B\ p$

- *second-order attribution:*
  *"Ann **does not know** that Bob **knows** that* p"
  $\neg\ K_A\ K_B\ p$

The riddle of the consecutive numbers



First question:
“If you know which number you have, please step forward”

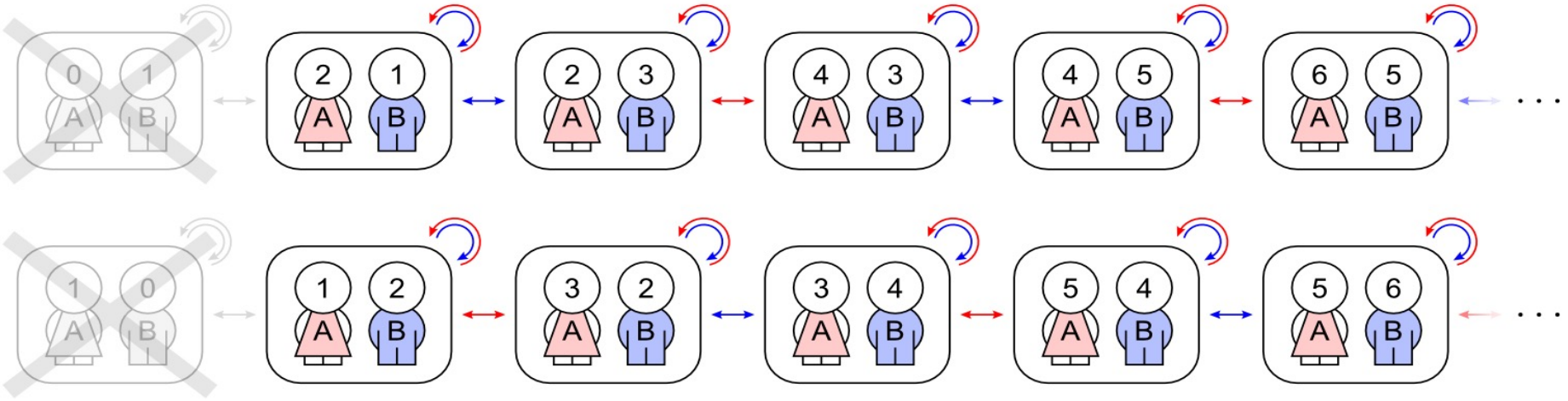
Second question:
“If you know which number you have, please step forward”

Third question:
“If you know which number you have, please step forward”

After the third question, Anja steps forward.

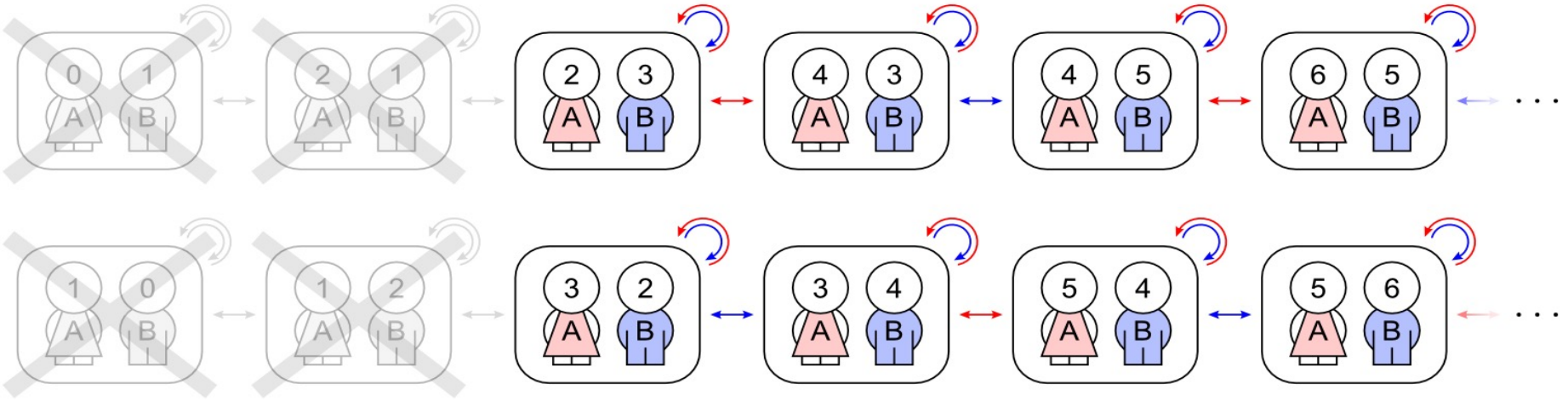
Starting situation

First question:
“If you know which number you have, please step forward”



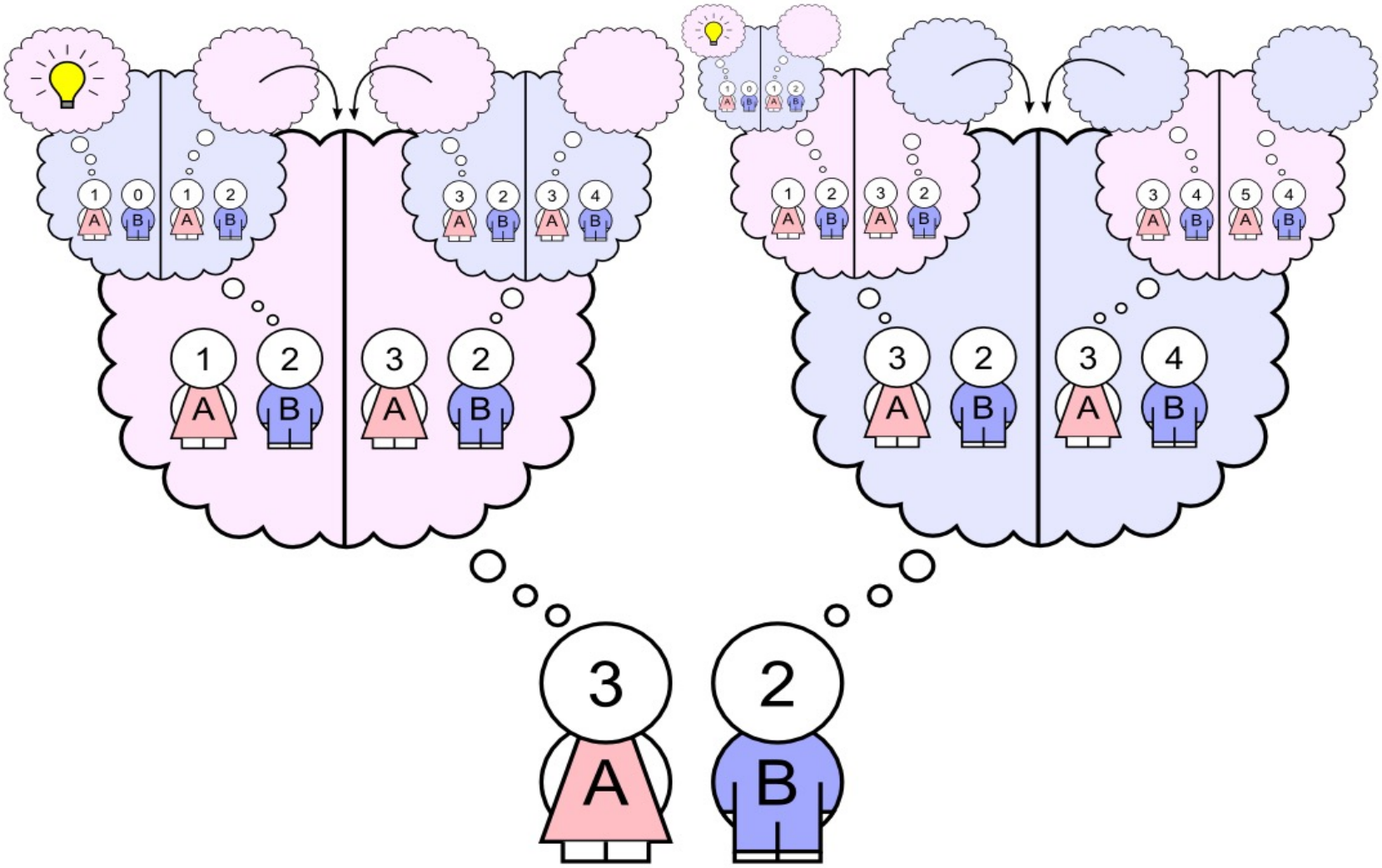
After the first question, nobody stepped forward.

Second question:
"If you know which number you have, please step forward"



After the second question, nobody stepped forward.

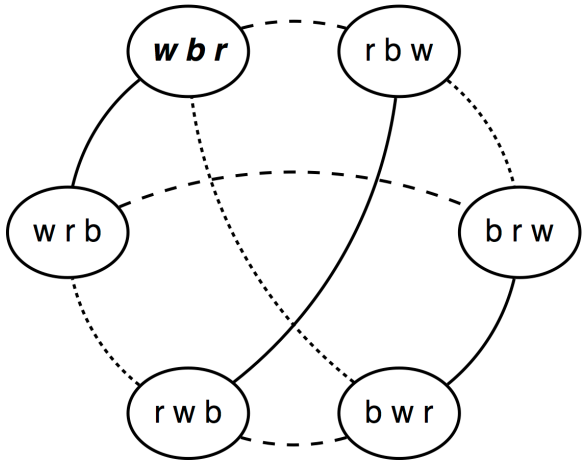
Third question: "If you know which number you have, please step forward".
After the third question, Anja steps forward and says: "I have 3".



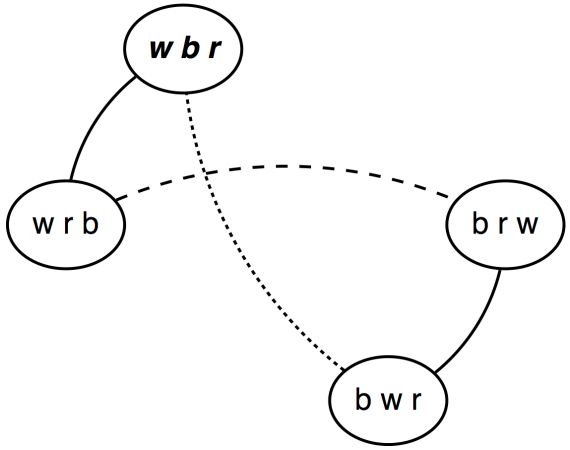
B. Kooi, H. van Ditmarsch en W. van der Hoek, *Dynamic Epistemic Logic*. Springer, Berlijn, 2007.

W. van der Hoek and R. Verbrugge, Epistemic logic: a survey. In: V. Mazalov and L. Petrosjan (eds.), *Game Theory and Applications, vol. 8*, Nova Science Publishers, New York, 2002, pp. 53-94.(4), 2008, 489-511

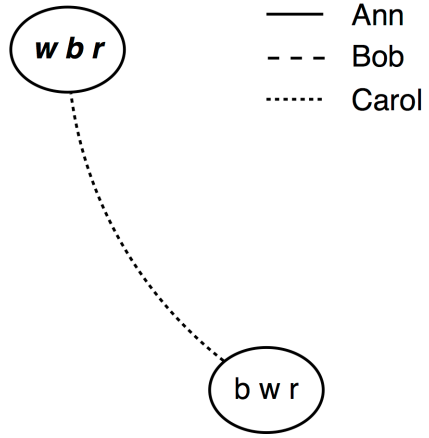
Models of dynamic-epistemic logic



Model 1



Model 2



Model 3

Truth definition for knowledge operator:

$$(M, w) \models K_A\varphi \text{ iff for all } v \text{ with } (w, v) \in R_A, (M, v) \models \varphi$$

In world *wbr* of Model I: Ann has white, w_{Ann}; Bob has blue, b_{Bob}; Carol has red, r_{Carol}

So in *wbr*, $\neg K_{Ann} \neg r_{Bob}$. Now public announcement: $\neg K_{Ann} \neg r_{Bob}$

Model 2 results. Next, suppose there is public announcement $\neg r_{Bob}$.

Model 3 results:

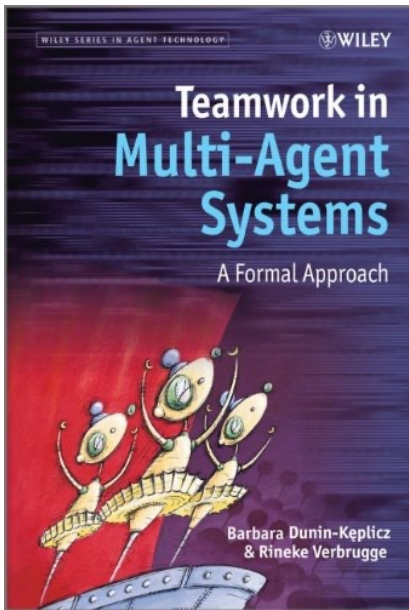
$$(M_3, wbr) \models K_{Ann}(w_{Ann} \wedge b_{Bob} \wedge r_{Carol}) \wedge \neg K_{Carol}(w_{Ann} \wedge b_{Bob} \wedge r_{Carol})$$

But do people really reason according to epistemic logics?

Computational complexity for multi-agent logics is high: the satisfiability problem is PSPACE-hard ($S5_n$, $KD45_n$)

If common knowledge is added, complexity of the satisfiability problem jumps to EXPTIME-hard.

Tractable cognition thesis:
People can only solve tractable problems (in P? fixed-parameter tractable?)



-M. Dziubiński, R. Verbrugge and B. Dunin-Kęplicz, Complexity issues in multi-agent logics. *Fundamenta Informaticae*, **75** (1-4), 2007, pp. 239-262.
-I. van de Pol, I. van Rooij, & J. Szymanik, (2018). Parameterized complexity of theory of mind reasoning in dynamic epistemic logic. *JoLLI*, 1-40.

Theory of mind is essential for hybrid intelligence

Aim: In Hybrid intelligence, humans and AI work together to jointly solve problems that neither could solve alone. AI does not replace but augment human intellect.

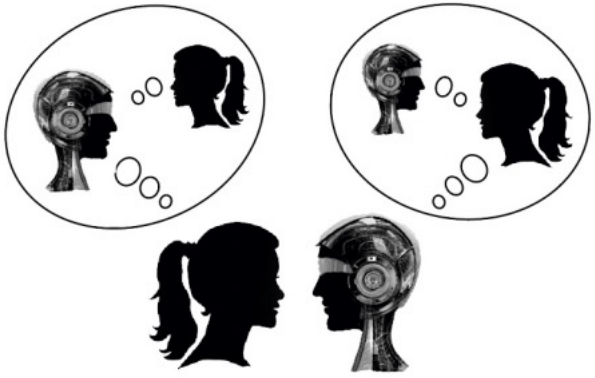


Fig. 3.1 Social Robots. Credit: Menah Wellen

AI needs to correctly model human theory of mind

Z. Akata, etc., 2020. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, *53*(8), 18-28.

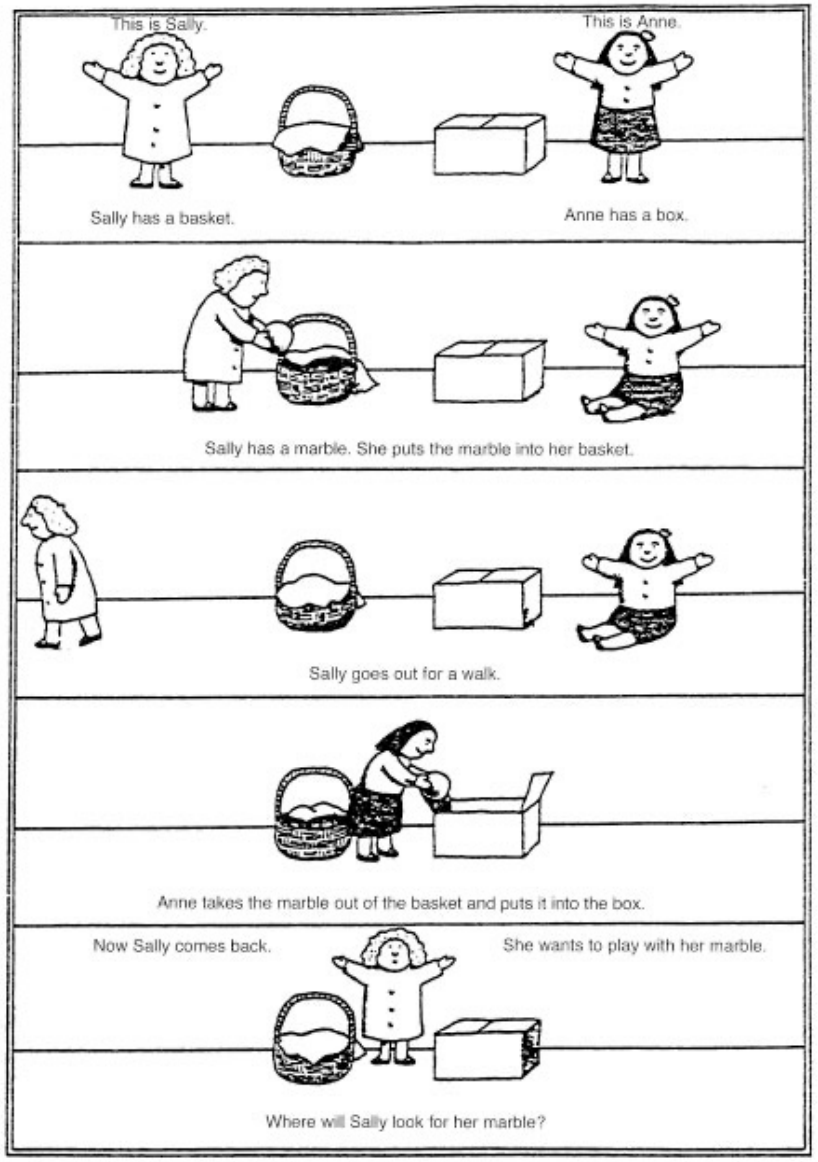
Overview rest of the talk

- **I. The first- & second-order false belief task**
  - Computational cognitive models of 5 year old children who are on the brink of developing second-order ToM
- **II. The Colored Trails game**
  - Using a software agent to entice students' theory of mind
- **III. The Marble Drop game**
  - What makes applying second-order ToM in a turn-taking game so difficult?
- **IV. Back to logic**
  - A logic for bounded reasoners
- **V. Does chatGPT 'have' theory of mind?**

I. Toddlers have trouble thinking about other people's beliefs

"If I cannot see them, then they cannot see me"

Theory of mind in children
A first-order false-belief task



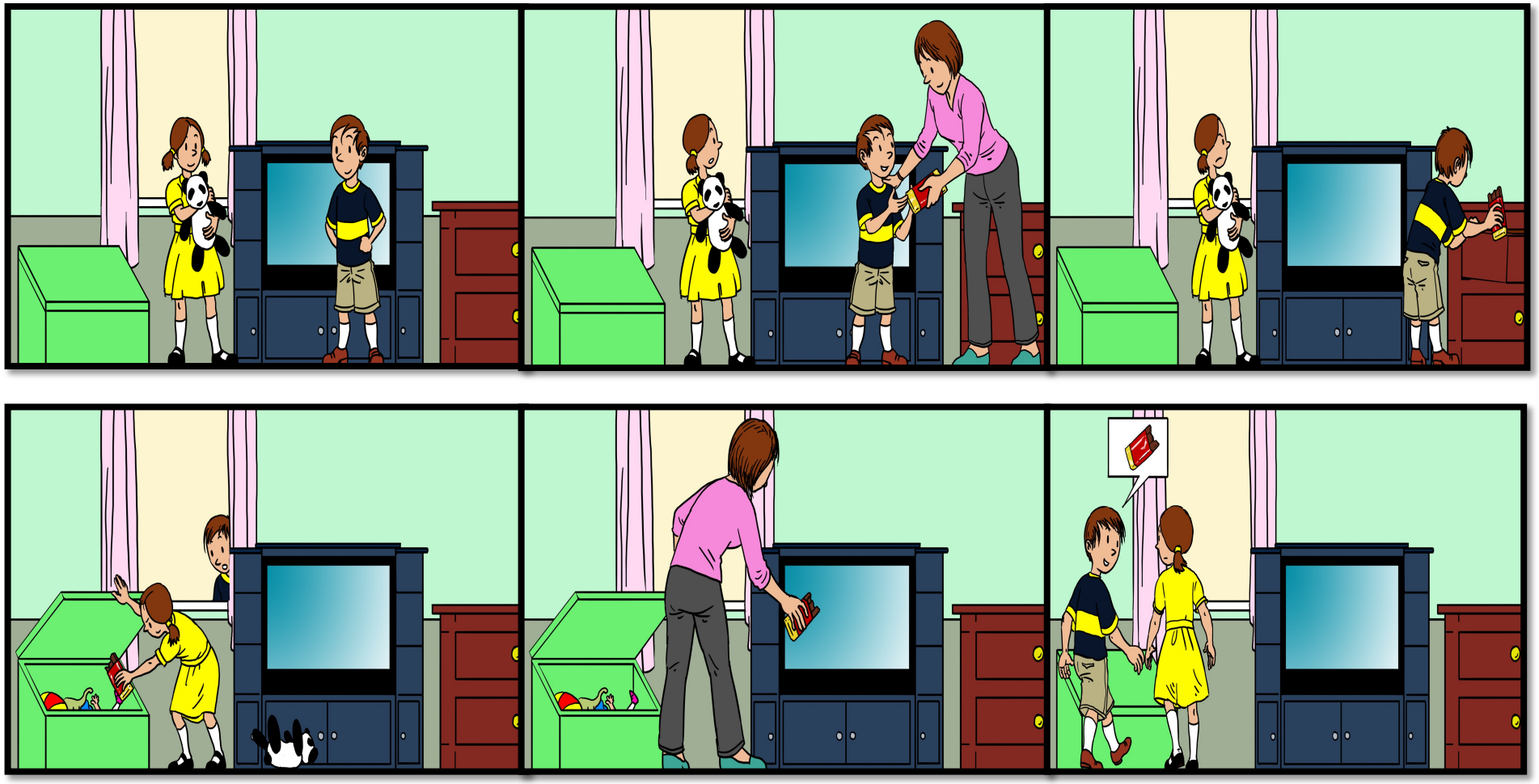
3-year old children: "Sally will look in the box."
(according to their own belief)

4-year old children: "Sally will look in the basket."
(according to Sally's false belief)

H. Wimmer en J. Perner, Beliefs about beliefs.
Cognition **13** (1), 1983, 103-12

Illustration Alex Scheffler (with permission)

Second-order false belief task ('Three locations')

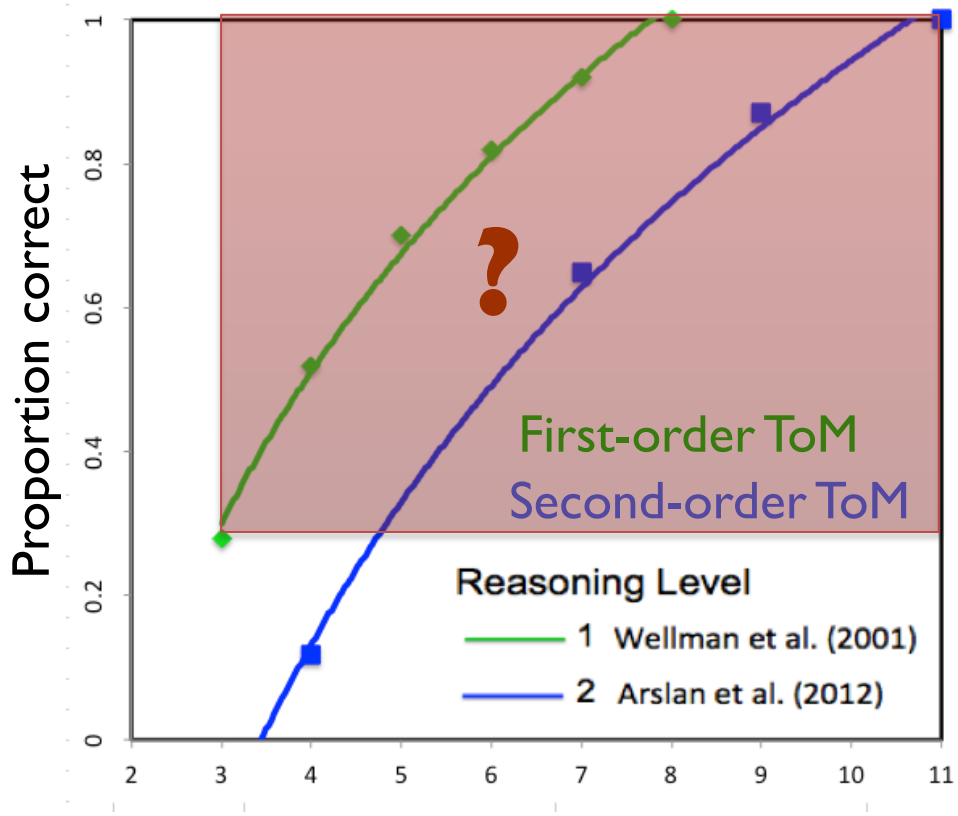


Reality control question: Where is the chocolate now? Zero-order (TV stand)

1st-order belief: Where will Murat look for the chocolate? First-order (Toy box)

2nd-order false belief: Where does Ayla think that Murat will look for the chocolate? Second-order (Drawer) Why does she think that?

Development of false belief reasoning



Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false-belief. *Child Development*, 72 (3), 655-684

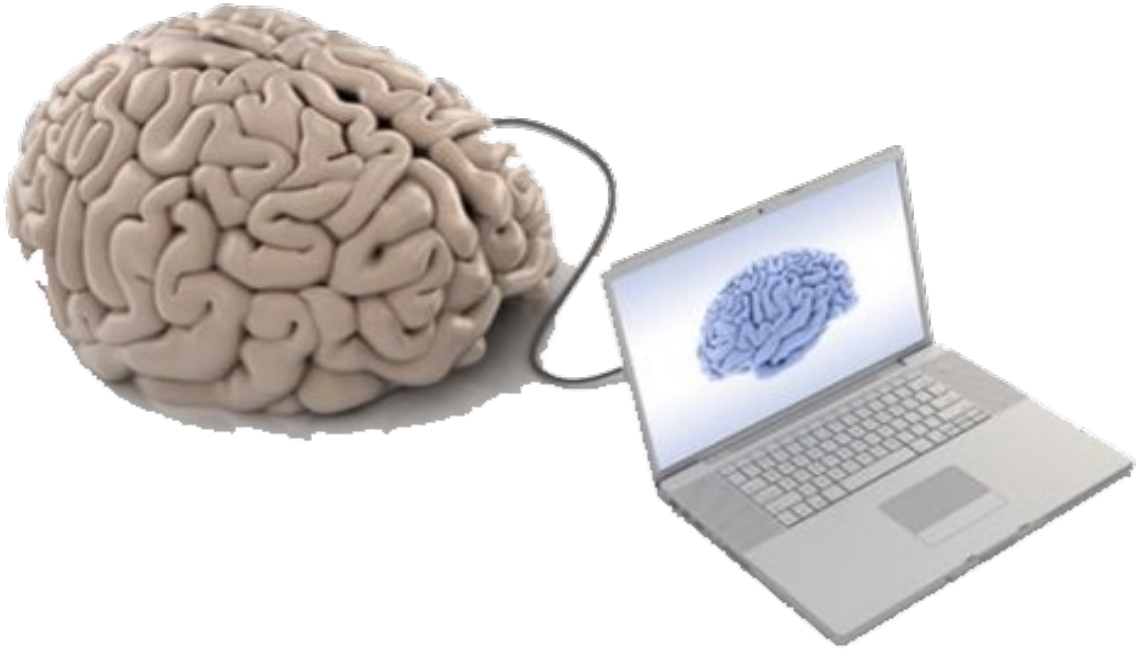
Arslan, B., Hohenberger, A., & Verbrugge, R. (2017). Syntactic recursion facilitates and working memory predicts recursive theory of mind. *PLoS ONE*, 12(1) (2017) e0169510

Research question about children's development of second-order theory of mind

How do children go through the reasoning transitions from their own point of view (zero-order) to taking into consideration another agent's beliefs (first-order), and later to taking into consideration another agent's beliefs about again another agent's beliefs (second-order)?

Burcu Arslan

A computational cognitive model of a second-order false belief task

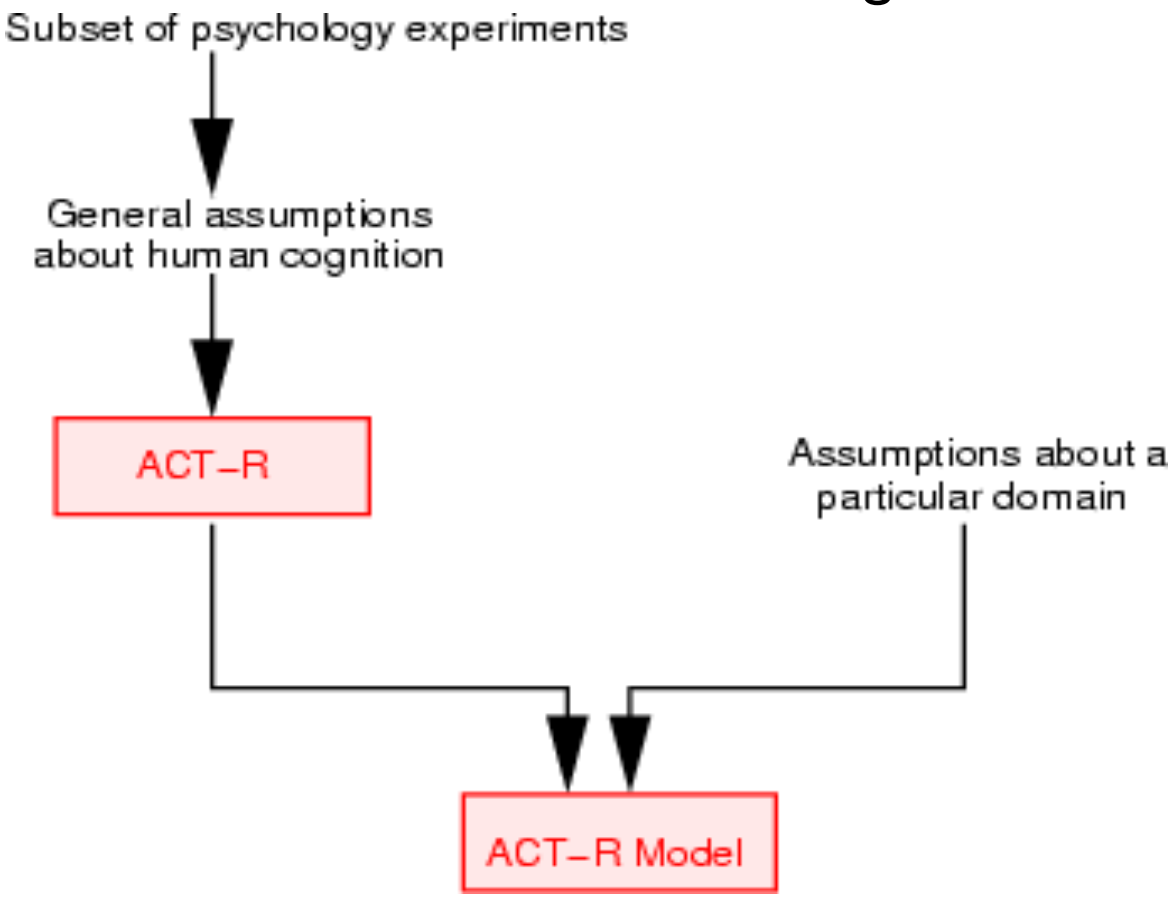


Using a computational cognitive model allows to make specific predictions about children's accuracy, reaction times, points of attention, active brain regions.
These predictions can be tested empirically.

Arslan, B., Taatgen, N. A., & Verbrugge, R. (2013). Modeling developmental transitions in reasoning about false beliefs of others. In R. West & T. Stewart (eds.), *Proceedings CogSci*, 77-82.

ACT-R

ACT-R is a **cognitive architecture**: a theory about how human cognition works.

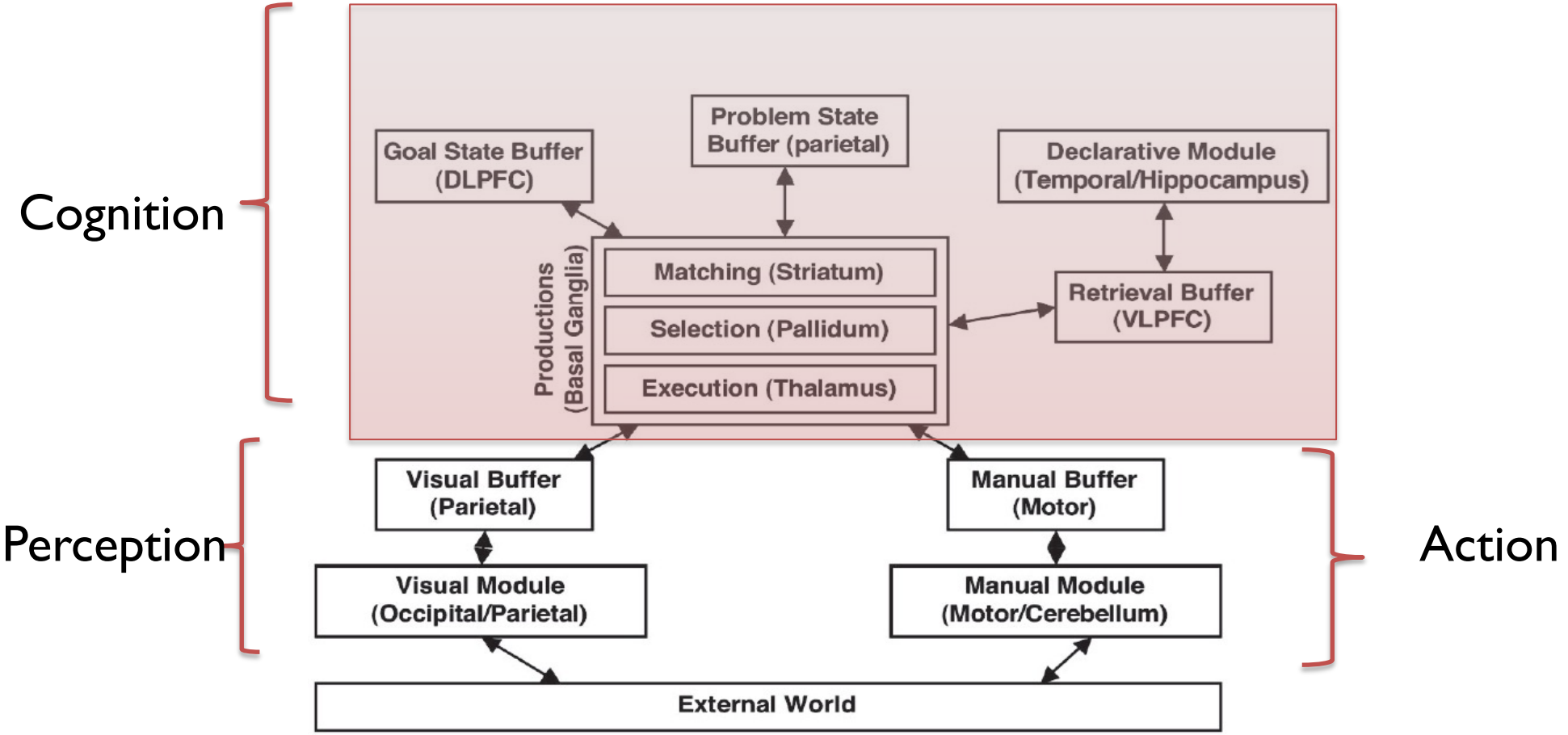


- ACT-R has been used for numerous different tasks:
  - ✓ memory for text
  - ✓ multi-tasking,
  - ✓ high school algebra,
  - ✓ air traffic control,
  - ✓ children's learning of irregular verbs
  - ✓ children's pronoun interpretation (him/himself)

Anderson, J.R & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum.

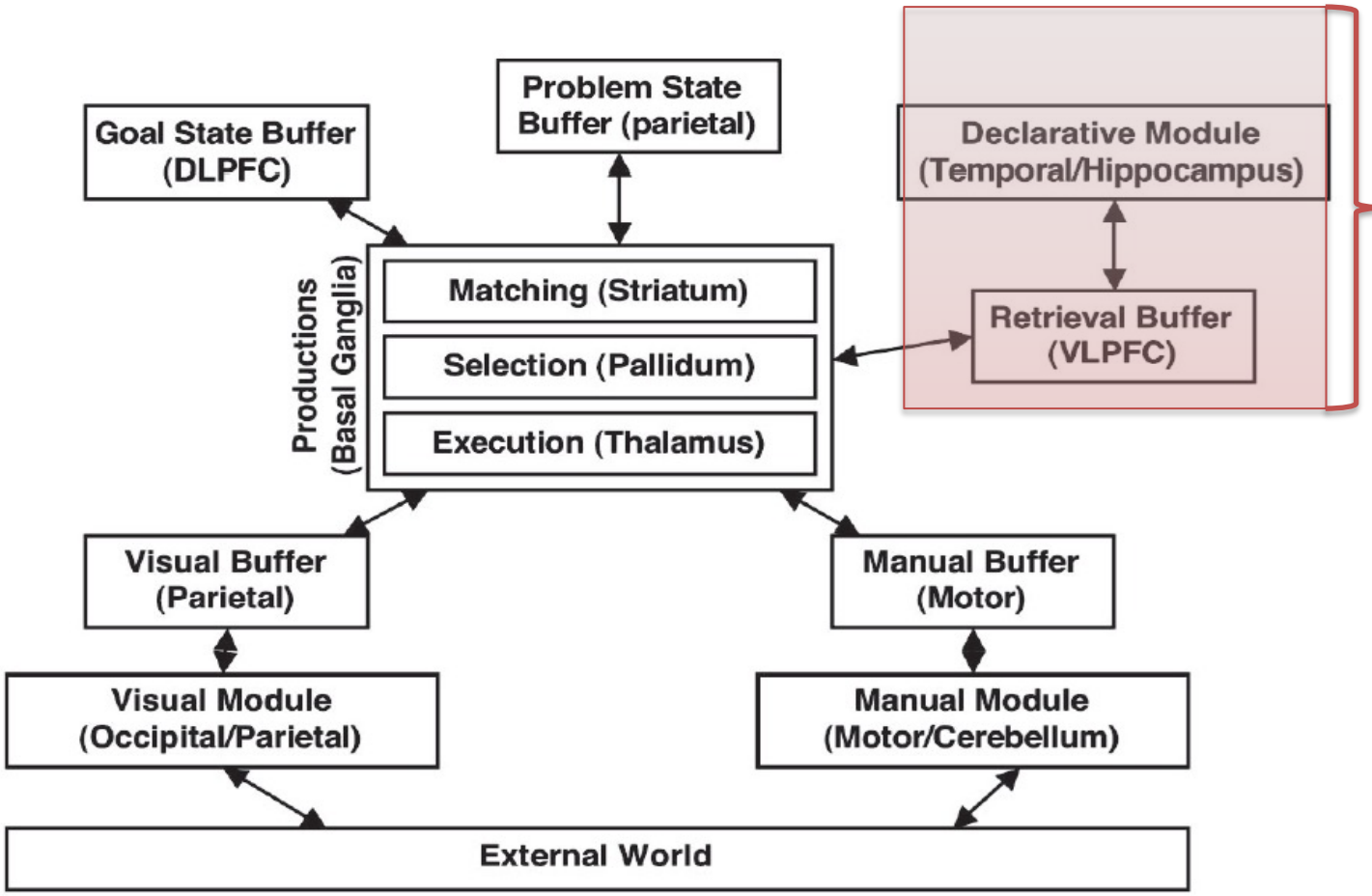
Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

How does ACT-R work?



Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

How does ACT-R work?



Factual knowledge represented in chunks

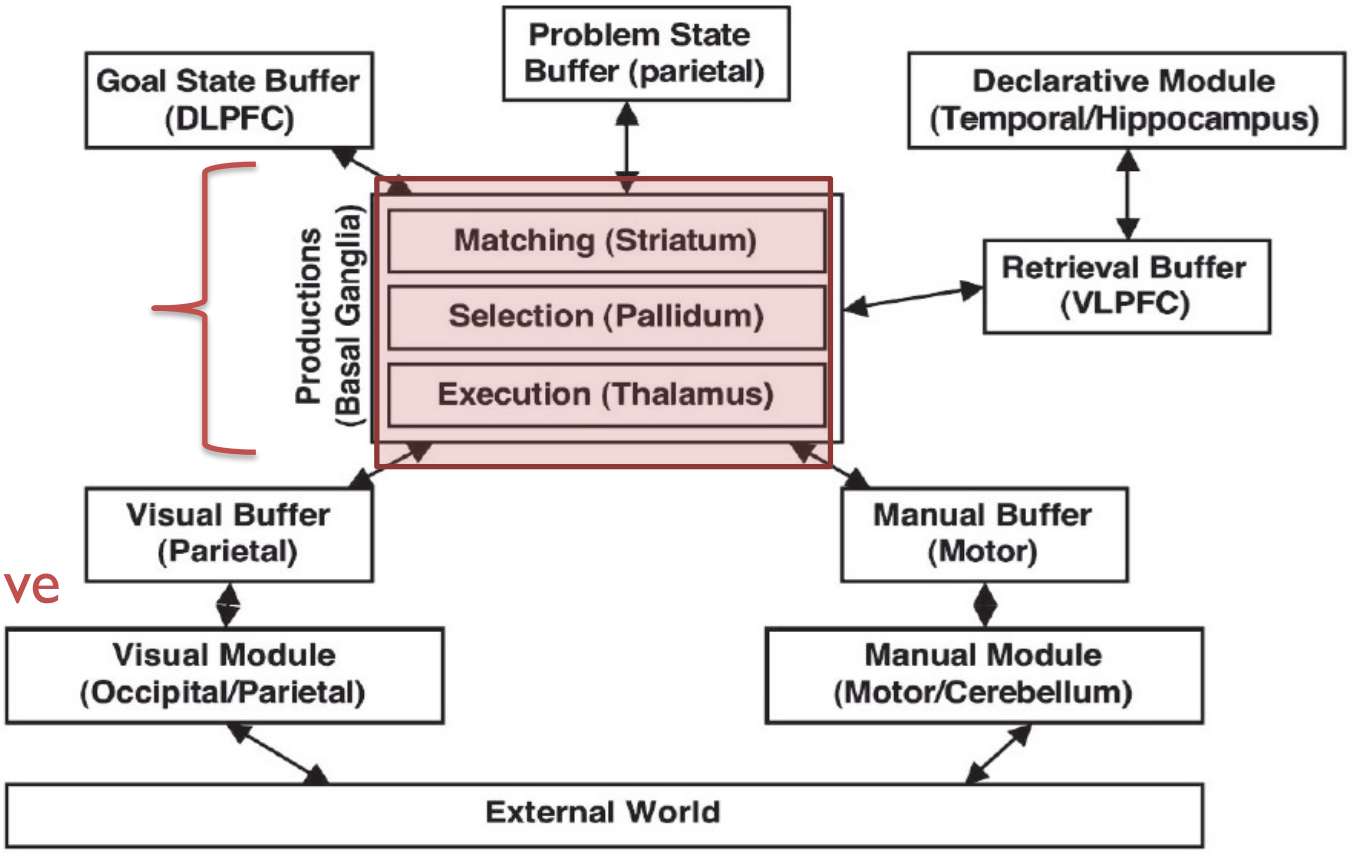
e.g. 2+3 = 5

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

How does ACT-R work?

Procedural knowledge

e.g. how to drive



Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

Production rules:

symbolic representation of procedural knowledge

```
(P example1
    A
    ==>
    B
)
```

Utility: 3

```
(P example2
    A
    ==>
    D
)
```

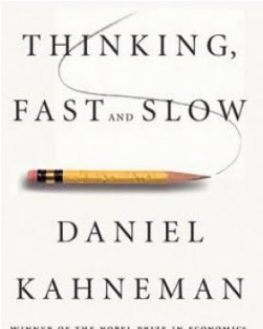
Utility: 2

```
(P example3
    B
    ==>
    C
)
```

Utility: 1

Production rules correspond with 'fast' decisions: system 1

Reasoning strategies in declarative memory correspond with 'slow' decisions: system 2



Two types of learning in ACT-R

1. Instance-based learning

Adding new 'reasoning strategy' chunks to declarative memory. In the beginning there is only reasoning 'from my own perspective' (zero-order).
Chunks will be retrieved more easily according to frequency and recency of use

2. Reinforcement learning

Reasoning strategies all exist from the start as production rules. The utility of a production rule increases when it plays a role in a successful answer, it decreases otherwise.

What the two models have in common

The relevant story facts +
some reasoning rules independent from the story:

i) The location of an object changes by an action towards that object.
ii) 'Seeing leads to knowing' (acquired by children around the age of 3).
iii) Inertia, e.g.: People search for an object at the location where they have last seen it, unless they are informed that there is a change in the location of the object.
iv) Other people reason 'like me'.

-Pratt & Bryant, 1990 Pratt, C., & Bryant, P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development*, *61* (4), 973-982.
-Stenning K. & van Lambalgen M. (2004). *Human Reasoning and Cognitive Science*. MIT Press,.

Instance-based learning model

Where does Ayla think that Murat will look for the chocolate?

s1 s2 s5

Retrieve a story fact that has an action

s5

If the retrieved story fact is not the latest fact, request the latest one.

Model's own perspective

(reasoning level 0)
(reasoning level 1)
(reasoning level 2)

Retrieve a strategy chunk

Drawer
Toybox
Tvstand

Give an answer accordingly.

WRONG!
CORRECT!
(reasoning level 1)
(reasoning level 2)

Based on the feedback add a higher-level strategy or strengthen the current one.

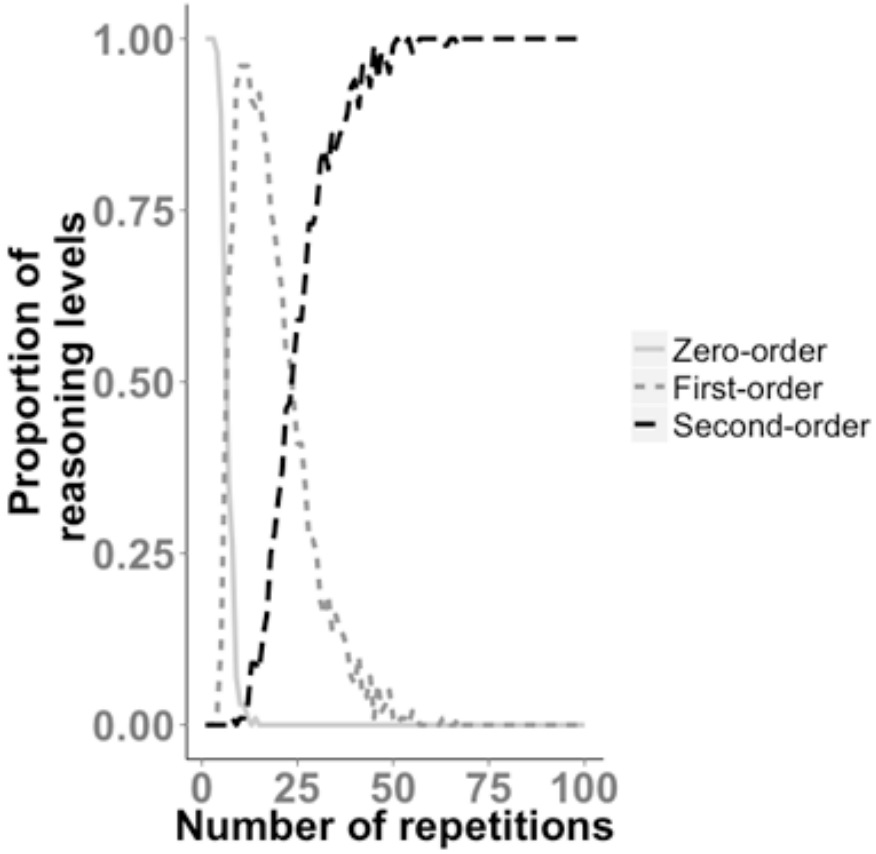
Declarative Memory

(s1 murat put chocolate drawer time 1)
(s2 ayla put chocolate toybox time 2)
(s3 murat saw ayla time 2)
(s4 ayla did not see murat time 2)
(s5 mother put chocolate tvstand time 3)

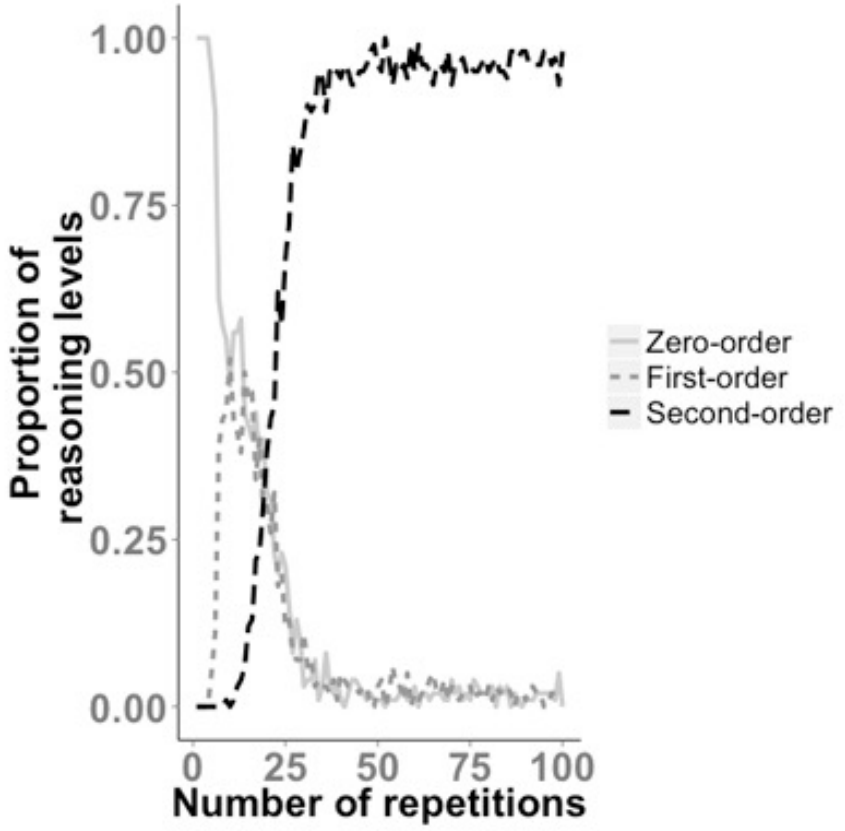
(reasoning level 0) (reasoning level 1) (reasoning level 2)

Results of the models

Instance-based learning model



Reinforcement learning model



For each learning model, we let 100 virtual children repeat the second-order false-belief task 100 times (20.000 runs in total)

Predictions of the models

Instance-based learning model

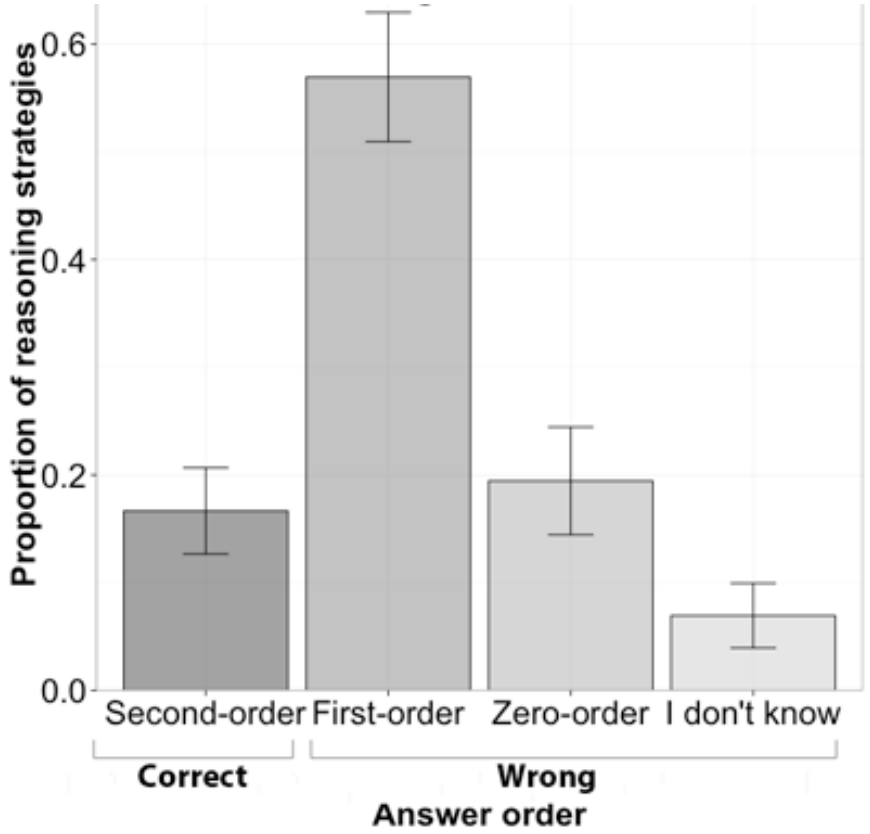
Children who have enough experience with first-order false belief reasoning but not with second-order reasoning do NOT give ZERO-ORDER answers BUT FIRST-ORDER answers to the second-order false belief question.

Reinforcement learning model

Children who do NOT have enough experience with second-order reasoning will EQUALLY give ZERO-ORDER and FIRST-ORDER answers to the second-order false belief question.

Experimental validation of the instance-based learnig model

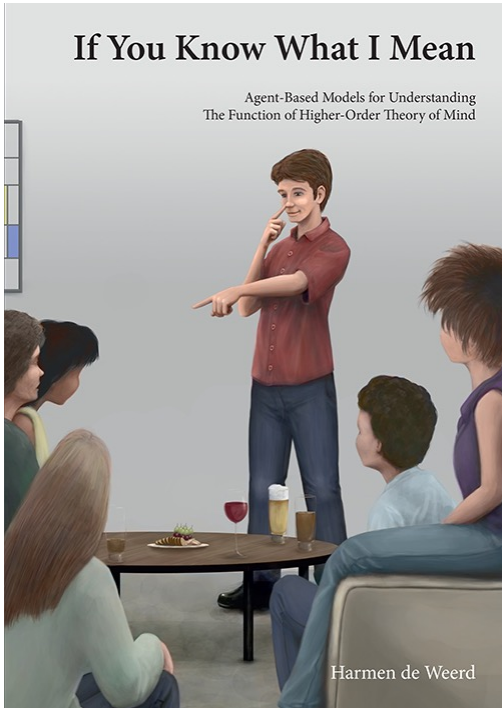
A sample of 79 Dutch 5-6 year-old children were recruited from a primary school in Groningen, the Netherlands.
(38 female, M_{age}=5.7 years, SE=0.04, range: 5.0 – 6.8 years).



Arslan, B., Taatgen, N. A., & Verbrugge, R. (2017). Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study. *Frontiers in Psychology*, 8

II. Students play a negotiation game against software agents

- Negotiations are situations with *mixed motives*, where participants have cooperative goals (to make a deal) & competitive goals (to get the most out of a trade)
- Is second-order theory of mind beneficial for agents in a negotiation game?
- Do students spontaneously use theory of mind in the negotiation game?



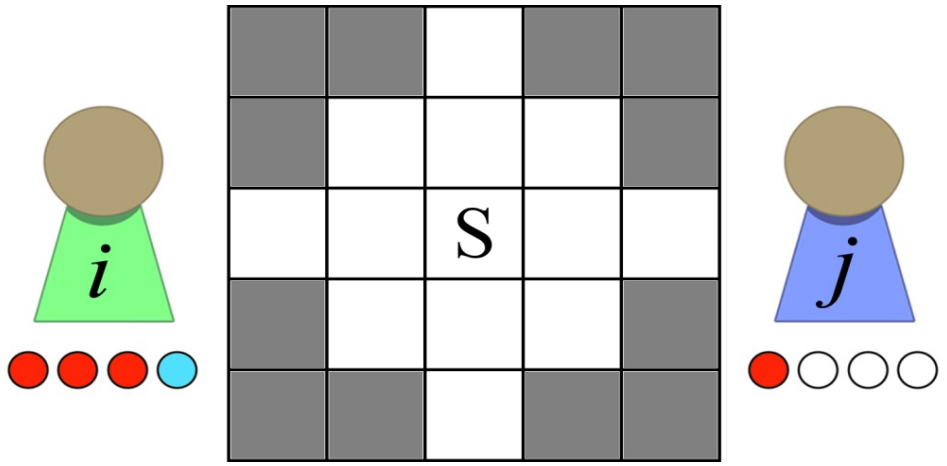
-de Weerd, H., Verbrugge, R., & Verheij, B. (2013). How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence*, *199*, 67-92.

Methodology for investigating ToM in a negotiation game

- Agent-based computational models
  - Simulate interacting agents
  - Introduce differences in the ability to use theory of mind
  - Compare performance among agents: Do higher orders of ToM allow agents to achieve better outcomes?
- Behavioral experiments
  - Let participants play against theory of mind agents
  - Use a higher-order ToM agent to determine to what extent human participants use ToM and whether they dynamically adapt their level to their opponent's use of ToM

-Verbrugge, R. (2009). Logic and social cognition: The facts matter, and so do computational models. *JPL* 38, 649-680.

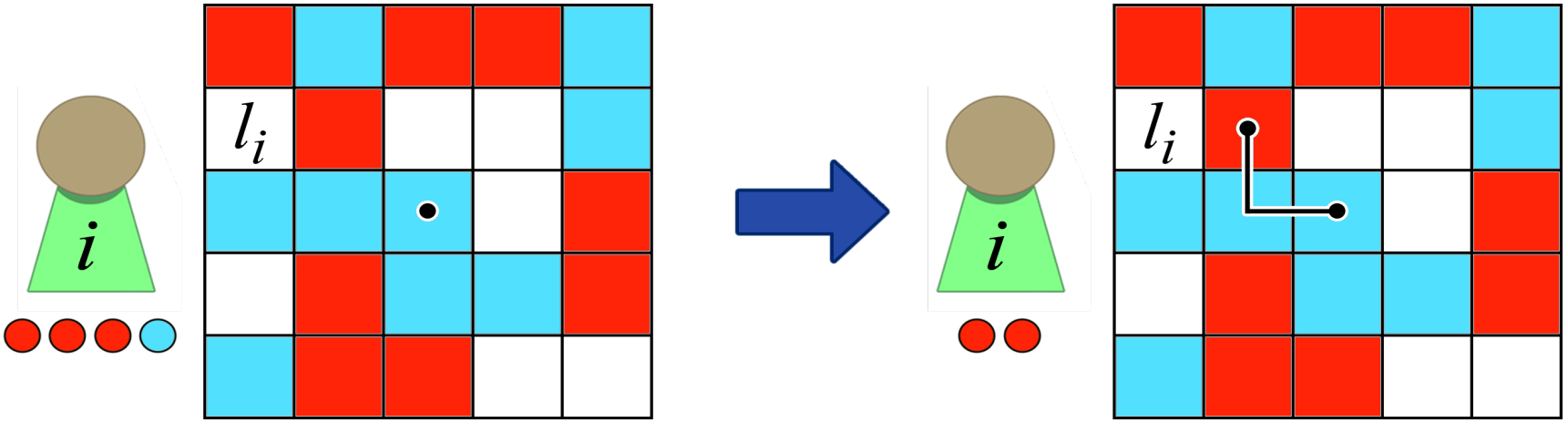
Colored trails: Negotiation game outline



- Each player has an initial location, goal location and set of chips
  - Each agent starts at the central square (marked S)
  - Goal locations are assigned randomly, > 2 steps away (gray squares)
  - Agents know their own goal location, but do not know the goal location of their trading partner (imperfect information)

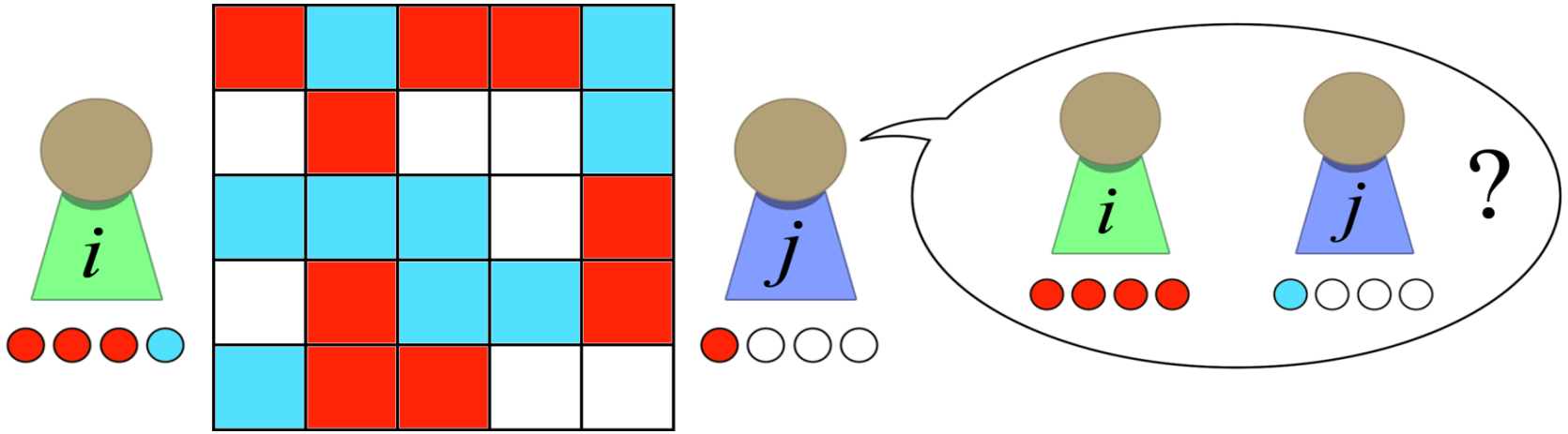
Grosz, B., Kraus, S., Talman, S., Stossel, B., Havlin, M.
The influence of social dependencies on decision-making: Initial investigations with a new game. *Proceedings AAMAS* 2004

Colored trails: Scores



- A player can move to an adjacent tile by handing in a chip of the same color as the destination tile
- Players are scored based on their final location:
  - Each step towards the goal is worth 100 points
  - Reaching the goal is worth an additional 500 points
  - Unused chips are worth 50 points each

Colored trails: Negotiation



- Players alternately offer a redistribution of chips:
  - Negotiation continues as long as agents make offers
  - Negotiation succeeds if an offer is accepted
  - Negotiation fails if a player withdraws from negotiation; then each player's set of chips remains as originally allocated to them

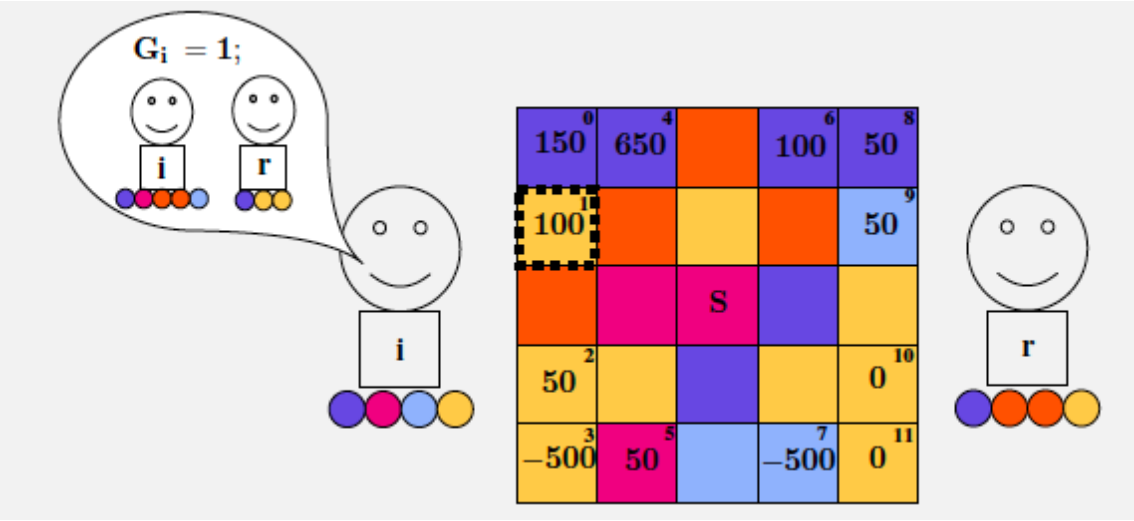
Results of second-order ToM agents in simulations

- ToM_2 agents outperform ToM_1 agents:
  - When a ToM_2 agent and a ToM_1 agent negotiate, the ToM_2 agent obtains at least as large a 'piece of the pie' as their trading partner
- Two ToM_2 agents work well together:
  - When two ToM_2 agents negotiate, they typically 'split the pie' into two equal pieces
  - Individual and collective incentives align, so behavior that yields a ToM_2 agent the highest gain also leads to highest collective performance

de Weerd, H., Verbrugge, R., & Verheij, B. Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *J. Autonomous Agents and Multi-Agent Systems*, 31(2): 250–287, 2017.

Recent results of 2nd-order ToM agents with the possibility to lie, in simulations

- In a recent variation on Colored Trails, some agents may tell their trading partner their goal location
- Some agents can lie



- It turns out that agents benefit more from a higher ToM level than from the ability to lie: honesty is the best policy

S. Brok, *The influence of lying in a negotiation setting: Colored Trails.* MSc thesis AI, U Groningen, 2023

Experiment on negotiations of students with ToM_0, ToM_1, and ToM_2 agents

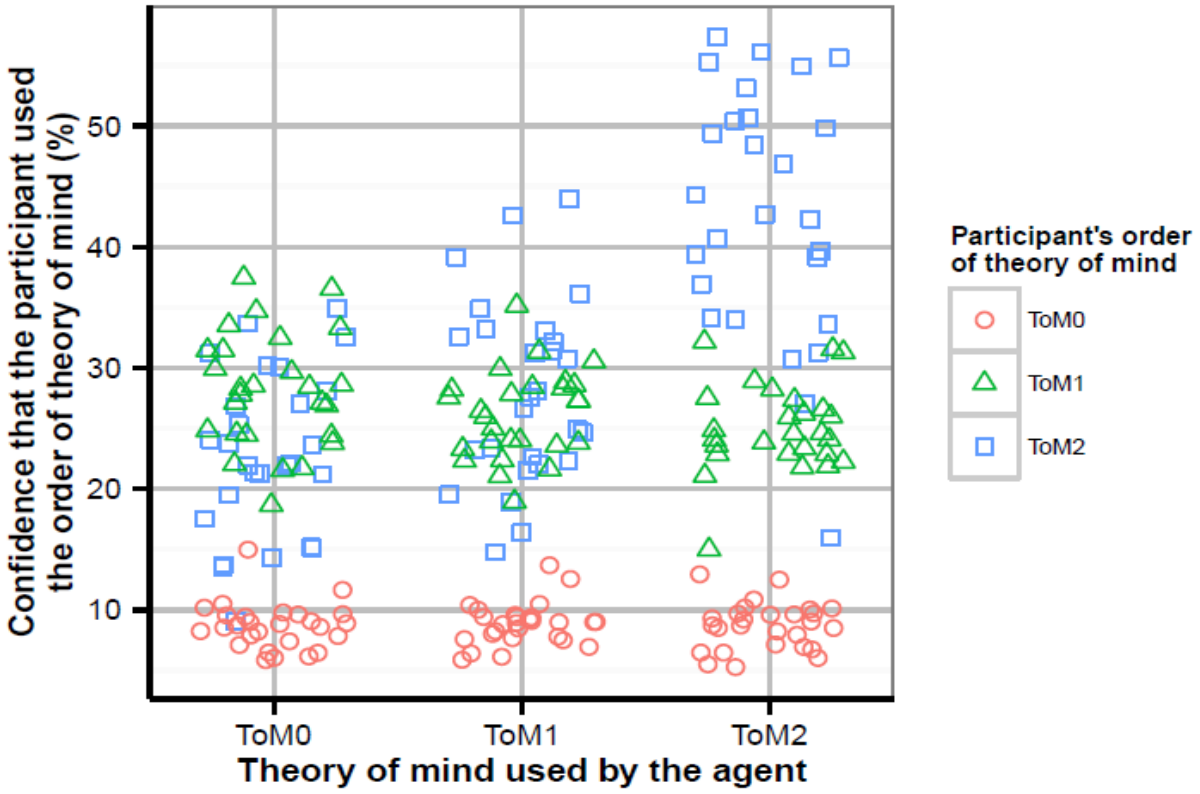
- Human participants play 24 Colored Trails games against computational agents
- Games are split up into 3 blocks of 8 games each
  - In each block, the theory of mind ability (ToM_0, ToM_1, ToM_2) of the computer player is different
  - Participants are not told that the computer agent changes
  - A negotiation game usually takes 4-6 rounds of offers and counteroffers
  - Participants have one minute to decide on each action (offer, accept, or withdraw)

Participant performance over all 24 games

Human subjects and agents usually come to win-win agreements; their scores do not differ significantly

Weerd, H. de, Broers, E., & Verbrugge, R. (2015) Savvy software agents can encourage the use of second-order theory of mind by negotiators. *Proceedings 37th Annual Meeting of the Cognitive Science Society*, pp. 542–547.

Orders of participants' theory of mind as estimated by a ToM_3 agent

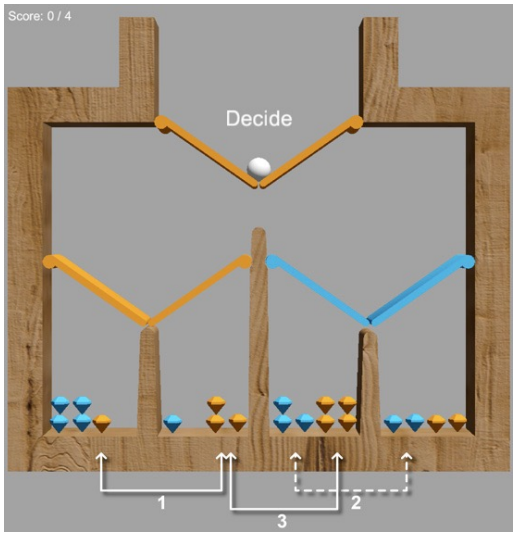


Participants are classified as a mix of ToM_1 and ToM_2.
When negotiating with ToM_2 agents, participants act more like ToM_2 agents

Student participants and ToM agents in the negotiation game

- Participants spontaneously use ToM_1, ToM_2 when they negotiate with agents
- A ToM_3 software agent can estimate, based on a number of different negotiation games, whether the participant plays ToM_0, ToM_1, or ToM_2
  - but it cannot discern other strategies.
- Participants adjust their ToM level to their trading partner

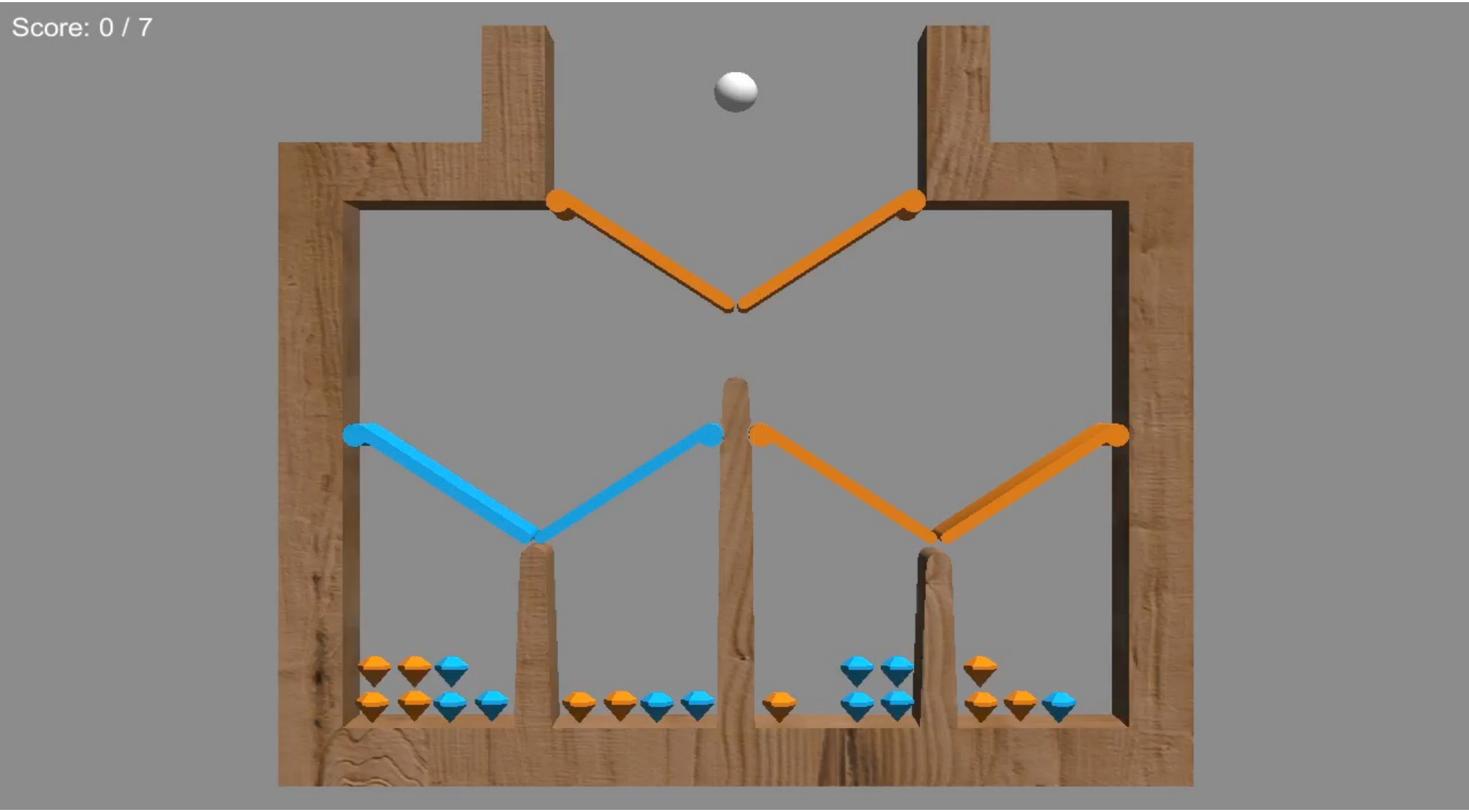
III. The Marble Drop game



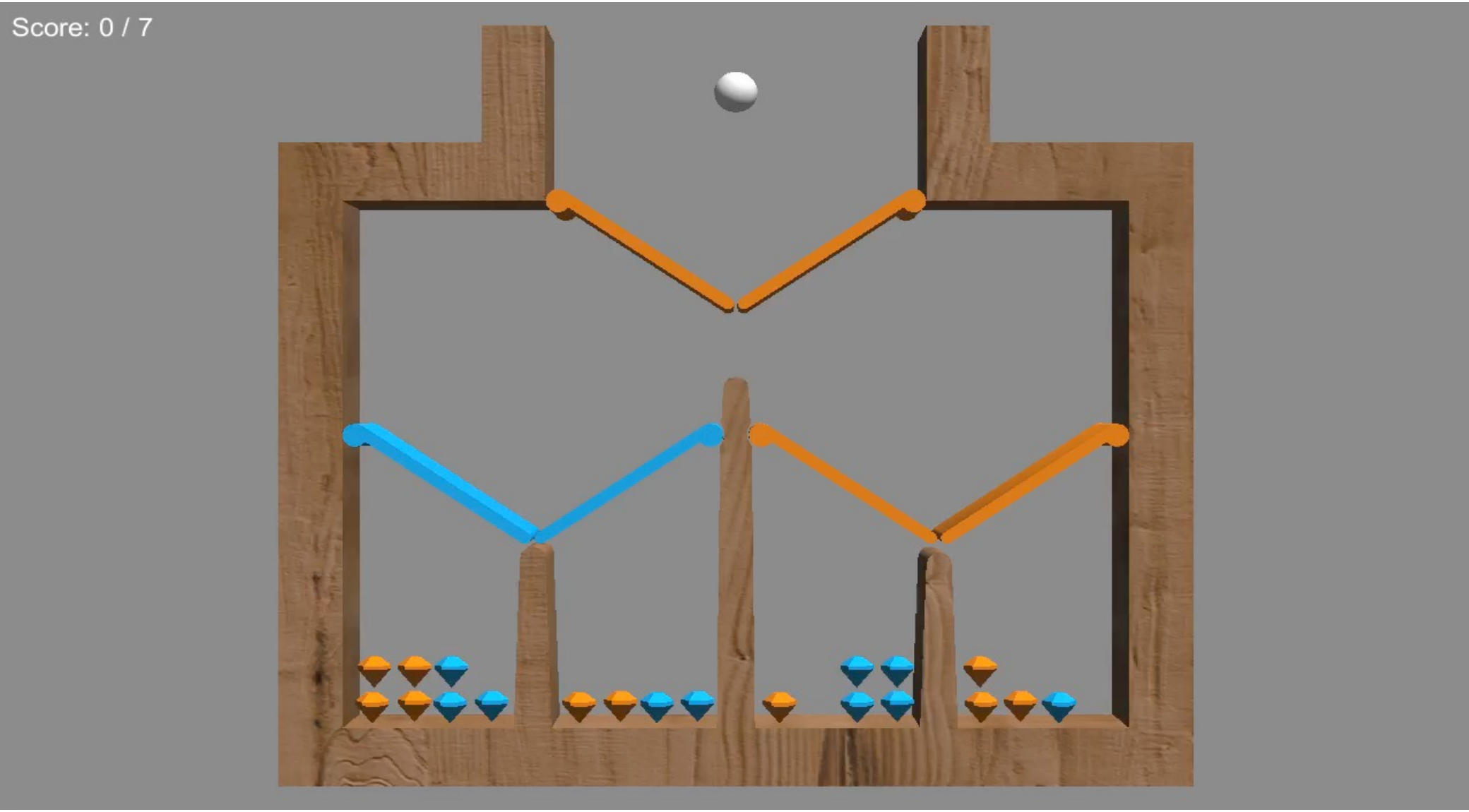
We designed the game Marble Drop:

- A turn-taking game between
  the participant (orange) and a computer player (blue)
- A white marble drops down. Players control the course of the marble by opening the left or right trapdoor of their color
- The participant wants the marble to drop into a bin in which there are as many orange diamonds as possible
- The computer wants the marble to drop into a bin in which there are as many blue diamonds as possible

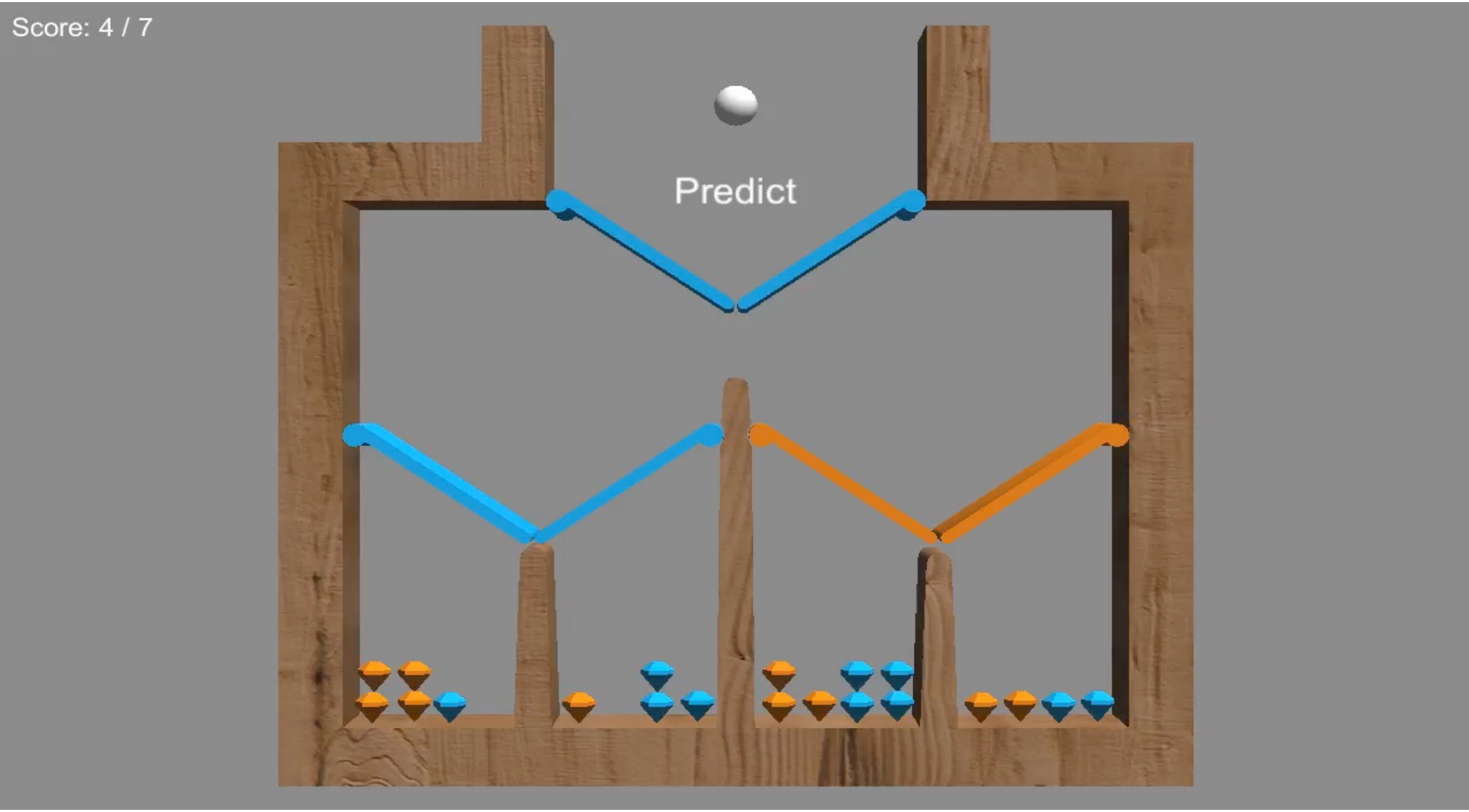
Our experiment: What do you, the orange player, decide?



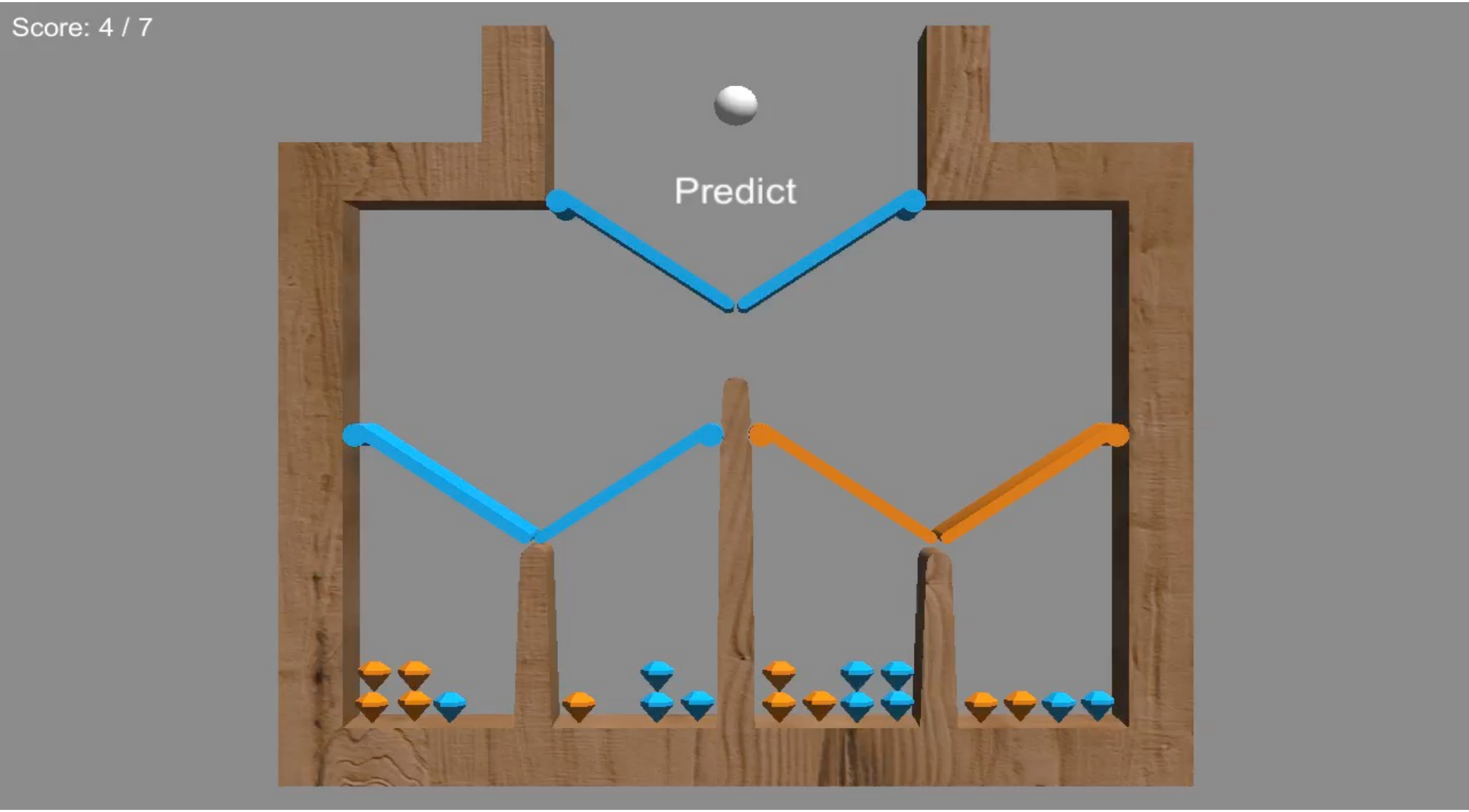
First-order games: “At blue doors, blue intends to go left”


Score: 0 / 7

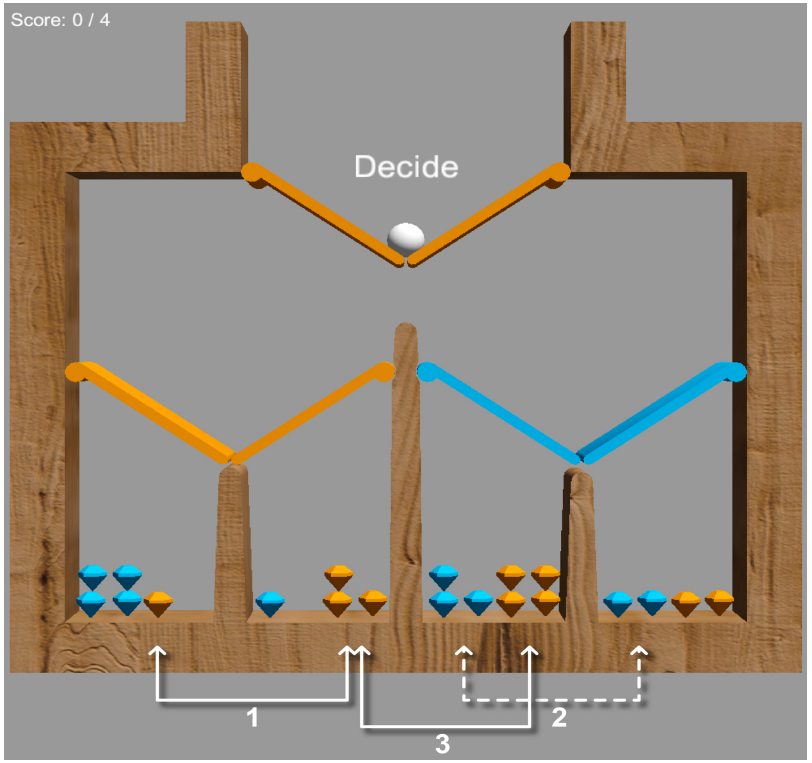
Predict what the blue opponent will decide

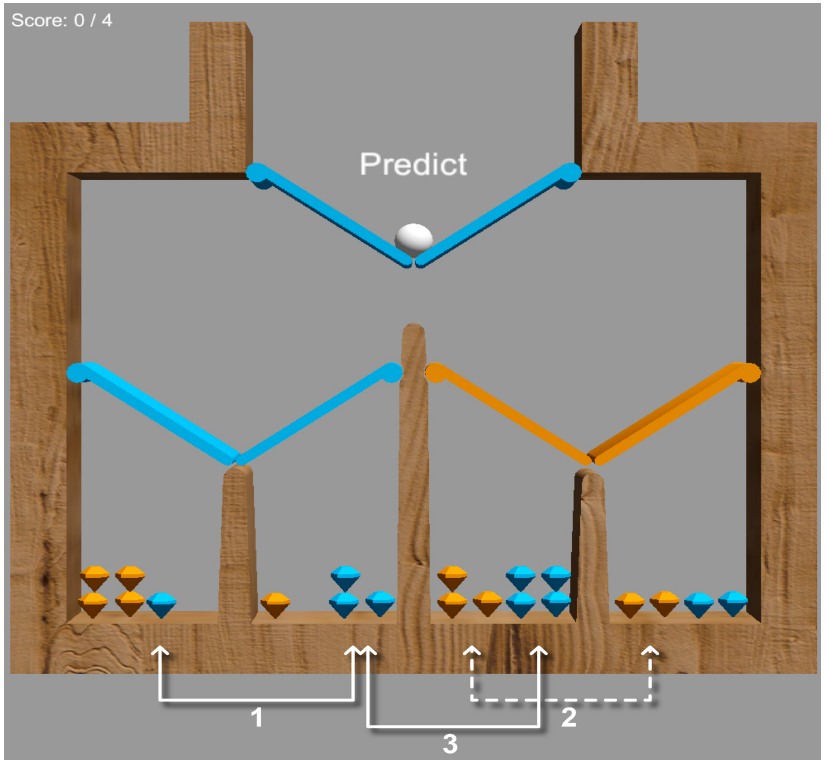


Second-order games: “Blue thinks that I intend to go left”


Score: 4 / 7
Predict

The two games are the same – except that orange and blue are exchanged





You have to make the same comparisons between numbers of marbles for both 'Decide' and 'Predict'.

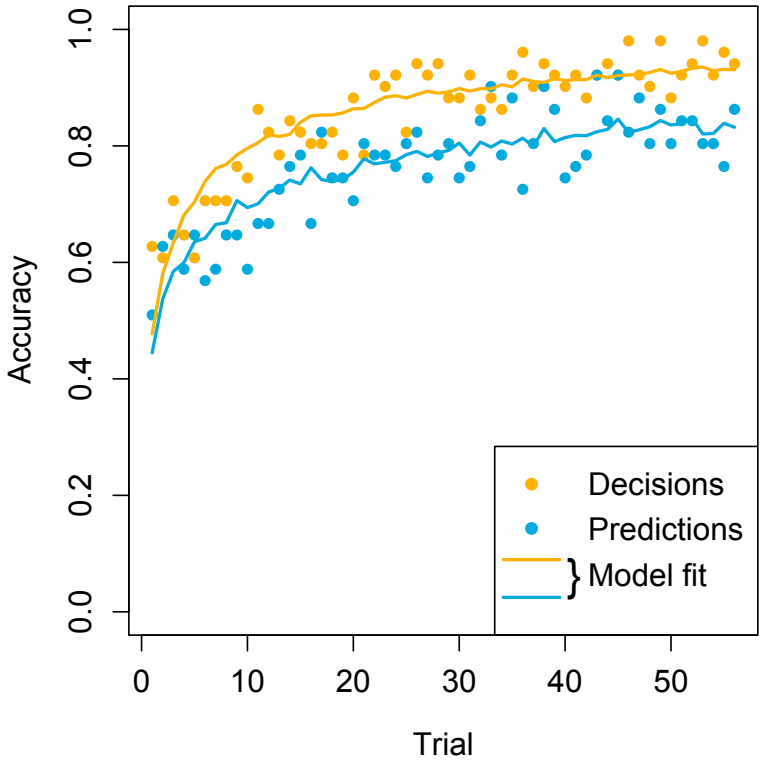
But for 'Decide', you apply first-order theory of mind:

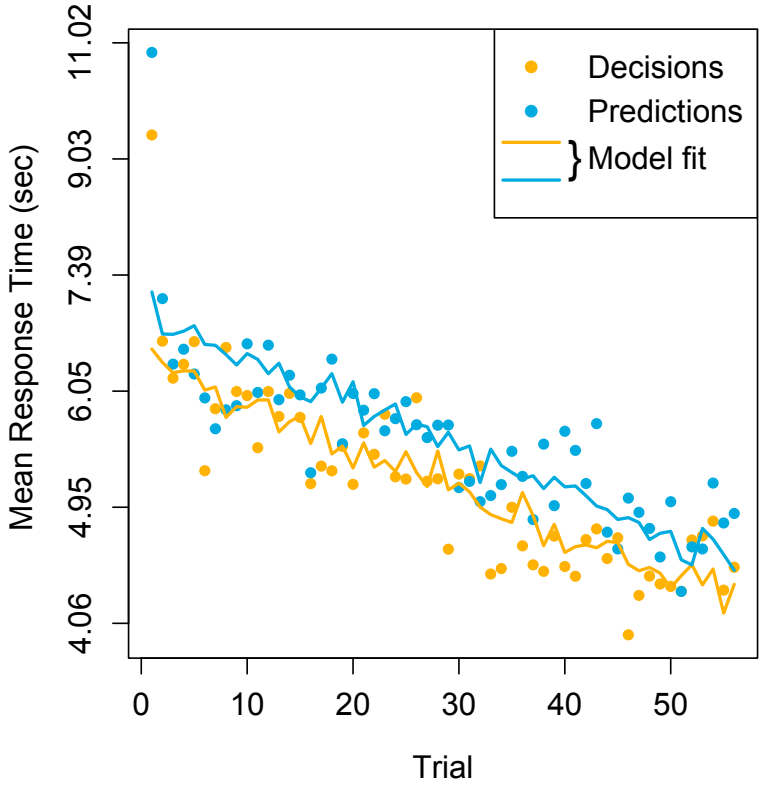
"What will blue *intend* to do if I go right?

For 'Predict', you need second-order ToM:

"What does blue *think* that I will *intend* to do if blue goes right?

Accuracy and reaction times





All participants play 56 first-order and 56 2nd-order games, in random order.
It remains difficult to 'put yourself in the other's shoes'.
Participants don't just think "Now I'm blue"/ "Now I'm orange!"

Verbrugge, Meijering, Wierda, van Rijn & Taatgen, It is hard to know your own mind when you stand in someone else's shoes: The costs of second-order theory of mind in turn-taking games. Under revision

IV Logic and theory of mind

"Very impressive, colleague…
But does it also work in theory?"

Logics have been used to explain behavior in false-belief tasks

- K. Stenning and M. van Lambalgen, *Human Reasoning and Cognitive Science*. MIT Press 2008
- H. van Ditmarsch & W. Labuschagne, My belief about your beliefs. *Synthese 2007*.
- T. Bolander, Seeing is believing: False-belief tasks in dynamic epistemic logic, *ECSI 2014*.
- T. Bräuner, P. Blackburn and I. Polyanskaya, Second-order false belief tasks: Analysis and formalization. *WOLLIC 2016*.
- I. van de Pol, I. van Rooij, & J. Szymanik, Parameterized complexity of ToM reasoning in dynamic epistemic logic. *JOLLI 2018*

Logic and theory of mind, cont.

"Very impressive, colleague…
But does it also work in theory?"

Logics have been used to explain behavior in epistemic riddles & games

- Z. Cedegao, H. Ham, W. Holliday: Does Amy know Ben knows you know your cards? *Proceedings CogSci*, 2021
- Anthia Solaki, Chapter 6, *Logical Models for Bounded Reasoners*, PhD thesis ILLC, 2021
- J.D. Top, C.M. Jonker, R. Verbrugge & H. de Weerd, Predictive theory of mind models based on public announcement logic, *Proceedings Workshop DaLí – Dynamic Logic*, 2023
- F. Arthaud & M. Rinard, Depth-bounded epistemic logic, *Proceedings TARK 2023*
- D. Longin & E. Lorini, Beliefs, time and space: A language for the Yokai board game. PRIMA 2020

TOMPAL: A bounded variant of dynamic epistemic logic

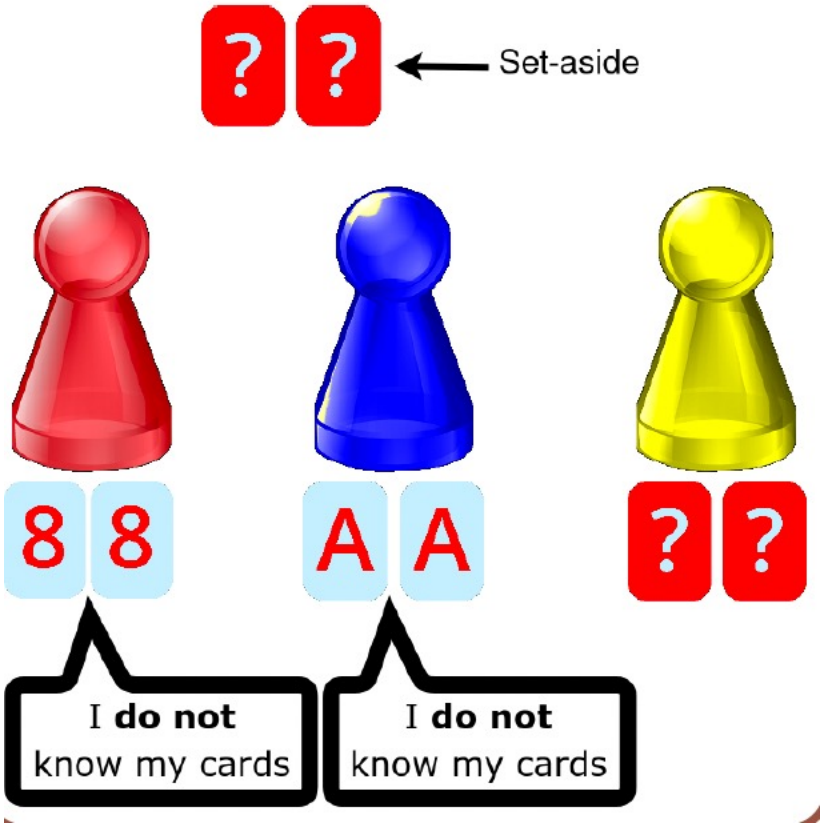
Goals:

- *Current:* Make logical models of bounded ToM reasoning in epistemic puzzles where all agents are truthful
- *Next*: Extend to situations in which some agents may lie
- These logics can be used to:
  - predict results of lab experiments
  - help design hybrid intelligent teams

J.D. Top, C.M. Jonker, R. Verbrugge & H. de Weerd, Predictive theory of mind models based on public announcement logic. In: Proceedings DaLí 2023

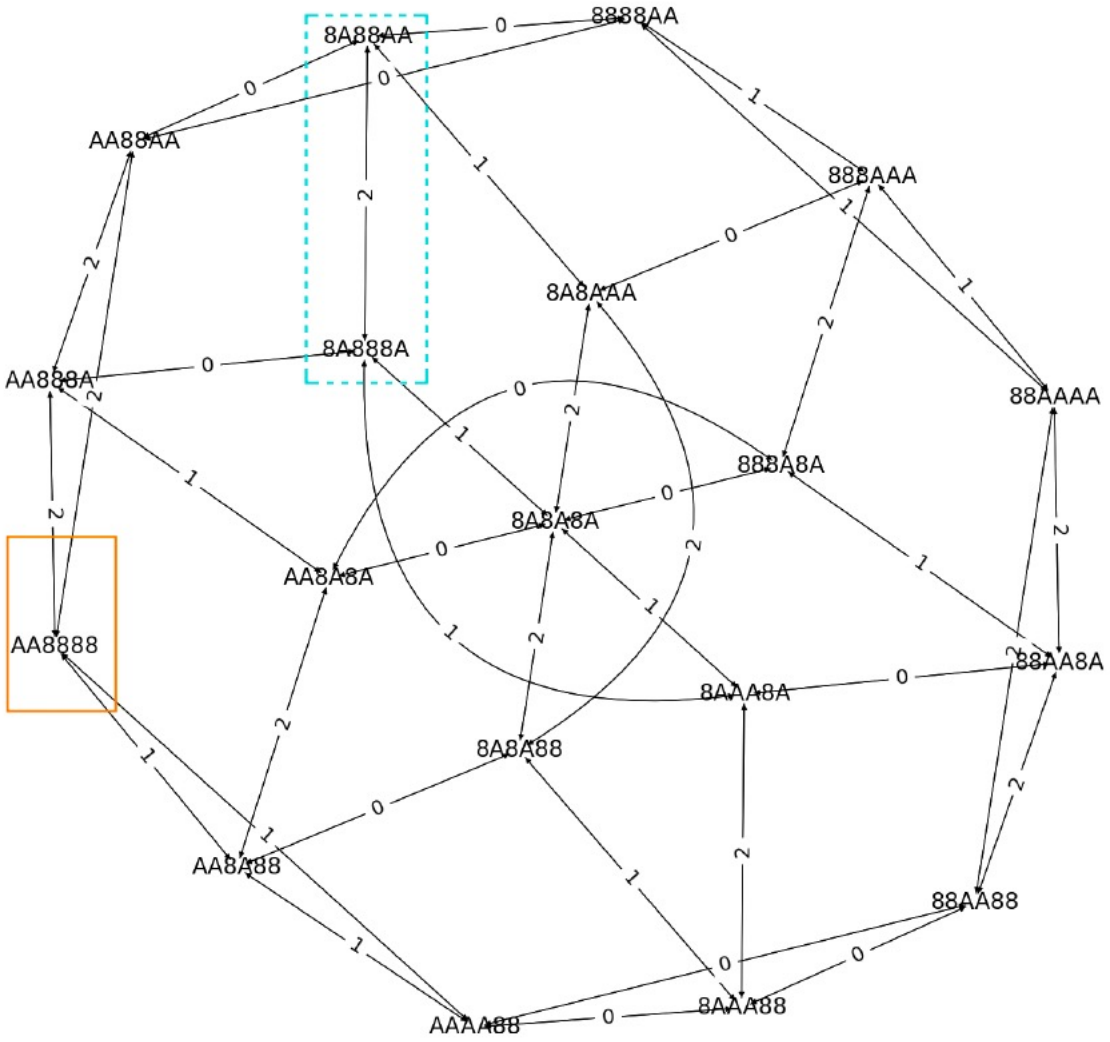
TOMPAL:
The game of Aces and Eights

- Three players
- Deck of 8 cards: 4 x A, 4 x 8
- Each player gets two cards
- You can only see the others' cards
- You are asked to announce, in turn, whether you know your cards
- Set-up is commonly known



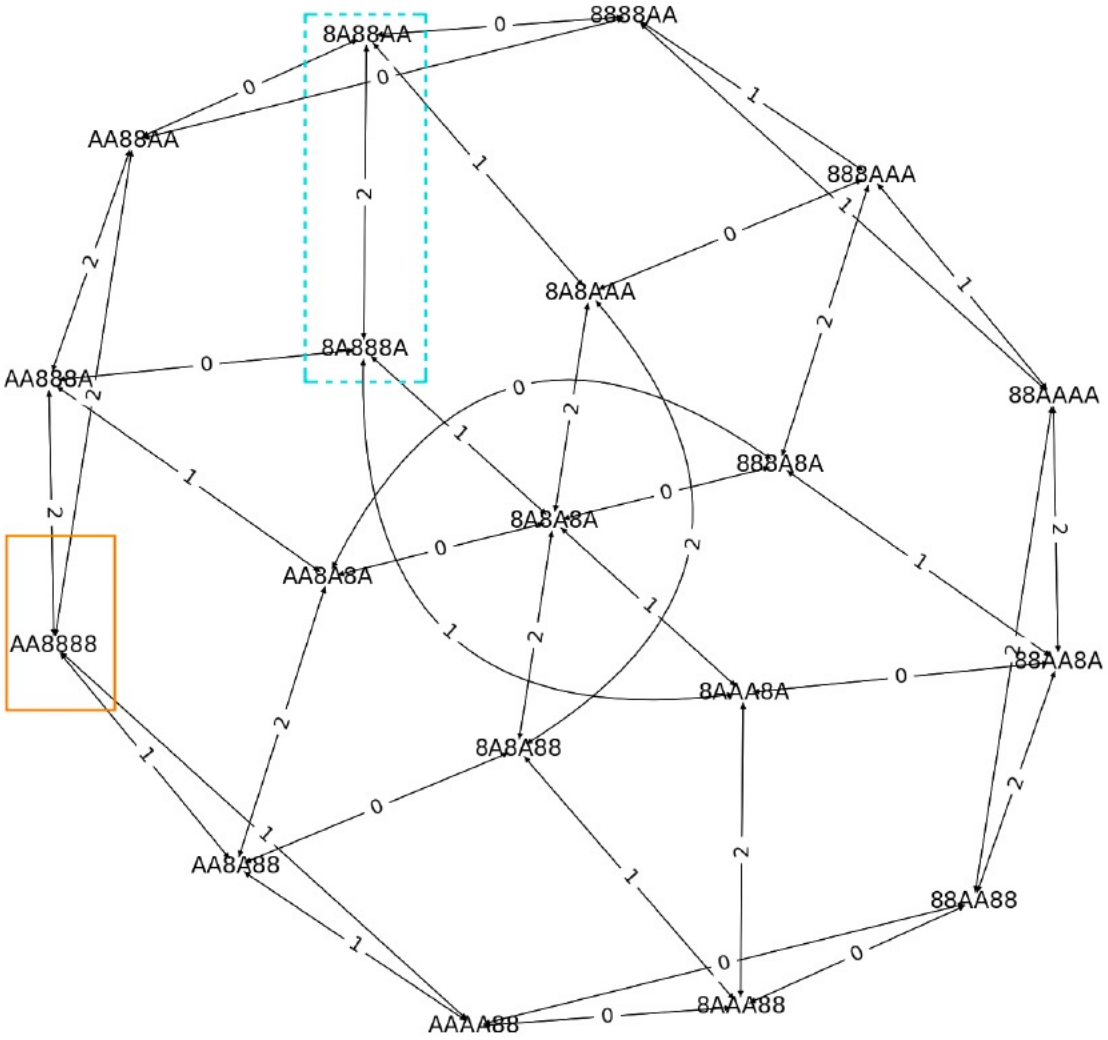
R. Fagin, J. Halpern, Y. Moses, & M. Vardi, Reasoning about Knowledge, MIT Press, 1995

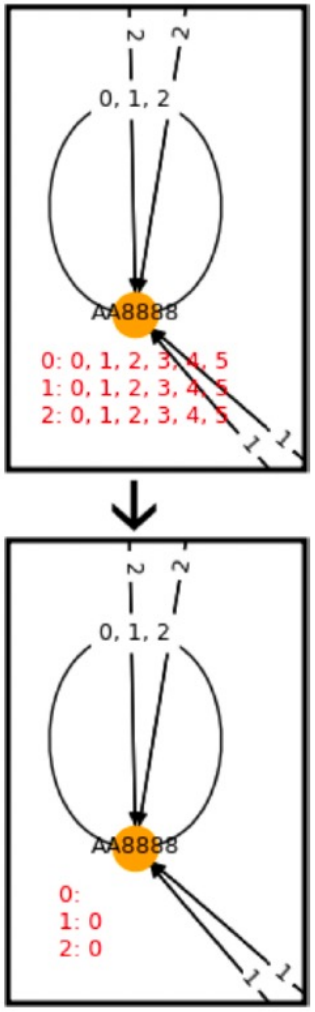
TOMPAL models for Aces and Eights



- In world AA8888, agent 0 knows that she has AA
- Agent 2 cannot distinguish worlds 8A88AA from world 8A888A

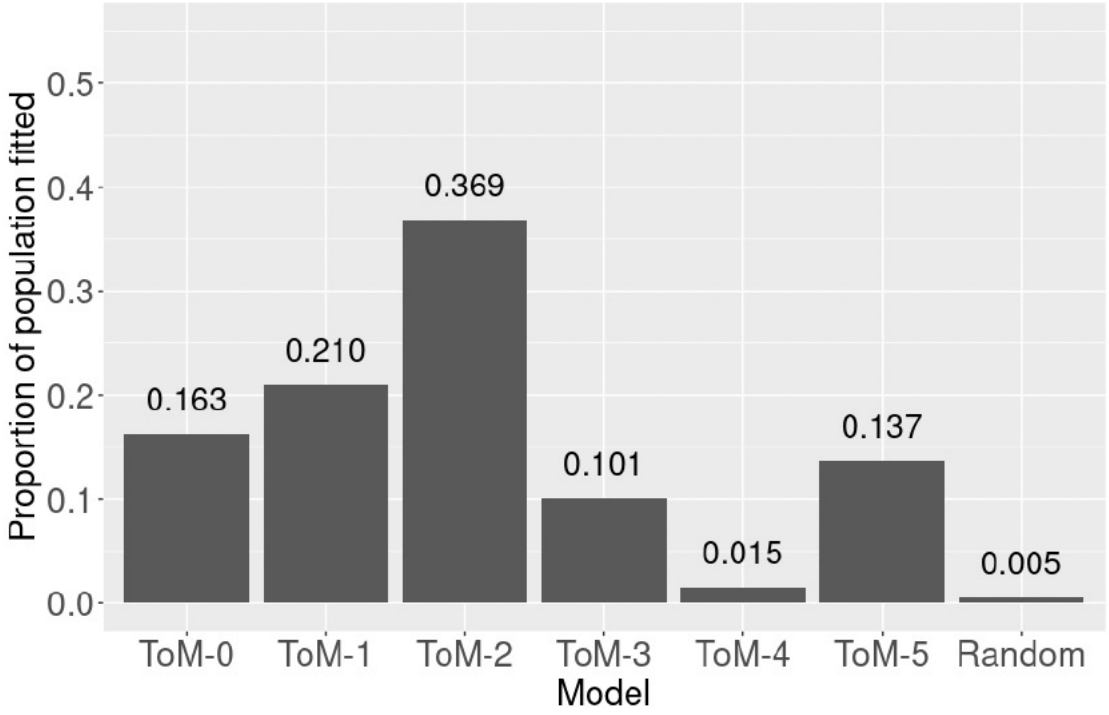
TOMPAL models for Aces and Eights





- Public announcement: "Player 0 does not know that she has AA"
  - For agents with ToM-1 and higher, world AA8888 is deleted
  - Agents with ToM-0 keep considering AA8888 possible

TOMPAL: Estimating participant ToM

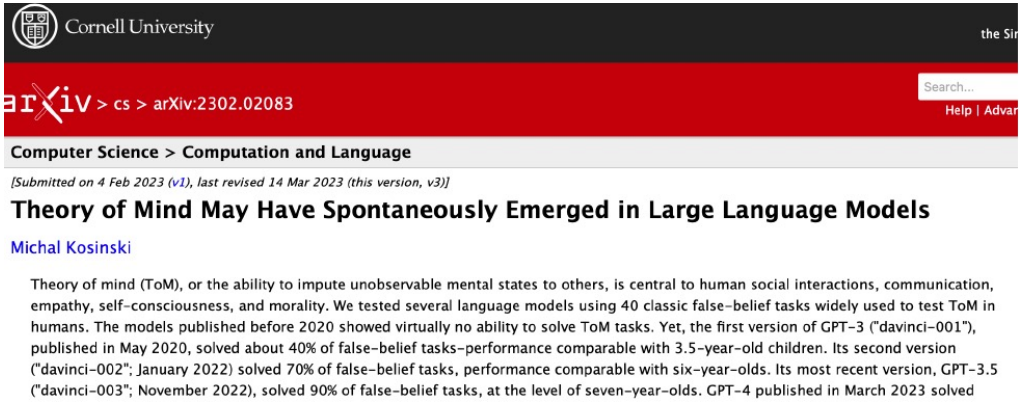


- Our estimated frequencies of strategies in Cedegao's data
- ToM-n: You can switch perspectives at most n times
- No limit on reasoning about your own knowledge
- Peak at ToM-2 is comparable to previous results

Z. Cedegao, H. Ham, W. Holliday: Does Amy know Ben knows you know your cards? *Proc. CogSci*, 2021

V. Do Large Language Models 'have' Theory of Mind?

Big claims by Kosinski 2023



Cornell University

the Sir

arXiv > cs > arXiv:2302.02083

Search...

Help | Advar

Computer Science > Computation and Language

[Submitted on 4 Feb 2023 (v1), last revised 14 Mar 2023 (this version, v3)]

Theory of Mind May Have Spontaneously Emerged in Large Language Models

Michal Kosinski

Theory of mind (ToM), or the ability to impute unobservable mental states to others, is central to human social interactions, communication, empathy, self-consciousness, and morality. We tested several language models using 40 classic false-belief tasks widely used to test ToM in humans. The models published before 2020 showed virtually no ability to solve ToM tasks. Yet, the first version of GPT-3 ("davinci-001"), published in May 2020, solved about 40% of false-belief tasks-performance comparable with 3.5-year-old children. Its second version ("davinci-002"; January 2022) solved 70% of false-belief tasks, performance comparable with six-year-olds. Its most recent version, GPT-3.5 ("davinci-003"; November 2022), solved 90% of false-belief tasks, at the level of seven-year-olds. GPT-4 published in March 2023 solved

More nuanced experiments

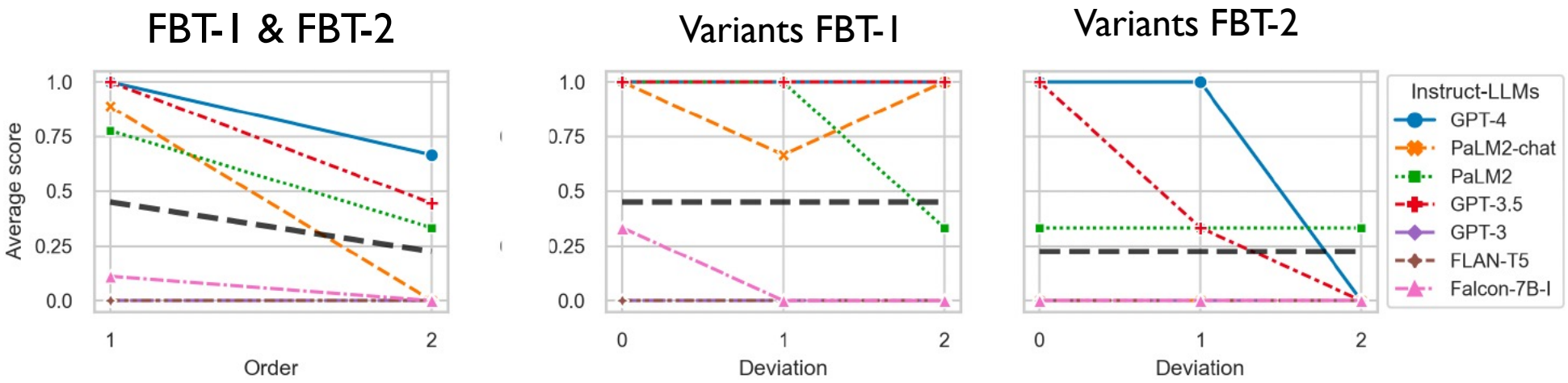
Bart van Dijk, Max van Duijn and Tom Kouwenhoven tested 'instruct-LLMs' such as chatGPT on a range of 1st-order and 2nd order false belief tasks and Happé's 'strange stories'

See e.g. Max van Duijn's presentation at EHBEA 2023, UCL and ongoing follow-up

Do Large Language Models have ToM?

Results of the Leiden group

1st- and 2nd order false belief tasks where given in:
0: original formulation; 1: close variant; 2: distant variant.
Performance of LLMs compared to 7-8 year olds (----):



Almost all LLMs struggle with FBT-2, except GPT-3.5 & -4.
For FBT-2 in a distant variant, even GPT-3.5 & -4 do worse than the 7-8 year old children.

Do Large Language Models have ToM?

Results of the Leiden group, continued

The researchers also show that GPT-3.5 & -4 do well on Happé's 'Strange Stories', including near and far variants; much better than 7-8 year old children.

They hypothesize that 'Strange Stories' require a willingness to be a cooperative communicator, which is rewarded in human interaction and in instruction of LLMs like GPT-3.5 & -4.

FBT-2 tasks are harder for LLMs because they rely on behaviorally-situated reasoning.

Conclusions: reasoning about reasoning about reasoning

- In hybrid human & AI multi-agent teams, the agents have different perspectives and need to think about others' mental states
- Computational cognitive models help to diagnose, understand, predict & train theory of mind
- 'Having theory of mind' is more than solving some known false belief tasks (chatGPT).

- Current research: Use ToM to help detect deception, including lies