Modeling Theory of Mind in ACTransfer

Stefan M. Wierda[1], Burcu Arslan[1]

[1] Institute of Artificial Intelligence, University of Groningen, Groningen, The Netherlands
wierda.stefan@gmail.com, barslan.cogs@gmail.com

Keywords: Theory of Mind · ACT-R · ACTransfer · False-belief task · Strategic turn-based games · Cognitive Modelling

1 Introduction

Does playing strategic games improve your performance on other tasks in which you have to reason about someone else's beliefs and intentions, as in for example a false-belief task? Both tasks require theory of mind—the ability to reason about someone's desires, beliefs, and intentions. In reasoning about others' minds, we can distinguish different orders of reasoning. When we reason about the reality (e.g., this abstract), we talk about zeroth-order theory of mind. First-order theory of mind is when we reason about someone's mental states (e.g., I think you find this abstract interesting to read). Second-order is reasoning about someone who reasons about someone's mental states (e.g., I think that you think that I find this abstract interesting to read) and so forth. Several training studies have shown transfer between first-order theory of mind tasks [1,2]. Thus, training in one task improves performance on a different theory of mind task—a phenomenon called transfer. However, for higher-order theory of mind, the amount of transfer between tasks is still unclear. To explore the relation between two different kinds of tasks that require higher-order theory of mind, the overlap between two tasks that require second-order theory of mind is investigated.

The first task is a turn-based strategy-game called marble drop designed by Meijering et al. [3] that is played with two players—an orange player and a blue player. In marble drop, a white marble is dropped onto orange and blue trapdoors that are controlled by the players that have the corresponding color. Behind each trapdoor, there is either a second pair of trapdoors or a bin containing a certain number of orange and blue marbles. The goal is to let the white marble reach the bin that contains the most marbles of your own color, regardless of the number of marbles the other player gets. Often you cannot reach the goal by just controlling your own trapdoors. Thus, you have to reason about the other player's beliefs, intentions and desires to predict which trapdoor your opponent is going to open.

The second task is a false-belief task. In a false-belief task, a story is told while corresponding pictures are presented to the participant. In the story, an item is moved from one location to the other. This movement is not observed by all characters in the story. Thus, one or more characters have false beliefs about the location of the item. After the presentation of the story, participants are asked to reason about the beliefs of one of the characters in the story. In the story modeled in this study, a mother gives a

piece of chocolate to her son Murat. Her daughter Ayla, who witnesses the gift, gets angry because she does not get one. Now, Murat puts the chocolate into the drawer and leaves the scene. Ayla then takes the chocolate out of the drawer and hides it in the toy box—however, Murat is secretly looking through the window and sees Ayla hiding the chocolate, without Ayla noticing him. Now, Ayla also leaves the scene and the mother enters. At the end of the story, the mother finds the chocolate in the toy box and then puts it near the TV. Participants are then asked where Ayla thinks that Murat is going to look for the chocolate. To answer this question, second-order theory of mind is required because the participant has to reason about Ayla's belief about Murat's belief. In this case, Murat thinks that the chocolate is in the toy box, because he did not see the mother move the chocolate. Ayla however, unaware of the fact that Murat observed her, thinks that Murat will look in the drawer—the location where he originally put it. The zeroth-order question in this story is "Where is the chocolate?" The correct answer to this question is the TV—the location where the mother last put the chocolate.

In 1901, Woodsworth and Thorndike [4] argued that transfer occurs when two tasks share identical elements of knowledge. However, they did not specify the identical elements. Decades later, Singley and Anderson [5] specified the identical elements as cognitive procedures and showed that transfer occurs when two tasks share the same procedures. However, the question remained open what the most minimal procedure looks like. Recently Niels Taatgen [6] proposed the primitive elements (PRIM). In the PRIM theory, cognitive procedures are broken down into two basic elements of cognition: the movement of information and the comparison of information. When learning a task, sequentially executed PRIMs evolve and this results in task-specific and general elements. When two tasks use the same sequence of PRIMs, transfer occurs. Here, we model the two abovementioned tasks by using the PRIM theory to identify whether the marble drop task and the false-belief task require the same underlying theory of mind strategies.

2 Methods

In this study, an extension of the cognitive architecture ACT-R [7] is used that implements the PRIM theory (http://www.ai.rug.nl/~niels/actransfer.html). By using such an architecture, the model inherits the cognitive constrains that the architecture implements on for example working memory and declarative memory.

The model of the marble drop game is based on a forward-reasoning plus backtracking algorithm [3,8,9]. The model first searches for the highest payoff (forward reasoning), and then tries to determine whether the payoff is attainable (backtracking). If a trapdoor of the opponent is encountered during the backtracking phase, the model switches perspective and then recursively uses the algorithm to determine the best choice for the other player and so forth.

The model for the false belief task is inspired on the model of Arslan, Taatgen, and Verbrugge [10]. Whereas their model has all the story facts in its declarative memory, the current model builds an internal representation as the story is presented. At the

end of the story, it tries to backtrack the actions and observations of each of the characters. First, the model will come up with a zeroth-order answer, the current location of the chocolate. Next, the model backtracks the movements of the chocolate and checks who performed or observed these actions. The model then infers the beliefs of the characters and answers the question.

In total, the model was run 20 times in four conditions that each had three blocks of 40 trials. The first two control conditions consisted of three blocks of the marble drop game (*md-md-md*) and three blocks of the false belief task (*fb-fb-fb*). The last two transfer conditions consisted of two blocks of the marble game with one block of the marble drop game in between and vice versa (*fb-md-fb* and *md-fb-md*, respectively). To examine transfer, only log-transformed reaction times of correct trials are considered. If the difference of the experimental block 2 and control block 3 is divided by the difference of the control bock 2 and control block 3, the transfer between the two tasks can be calculated.

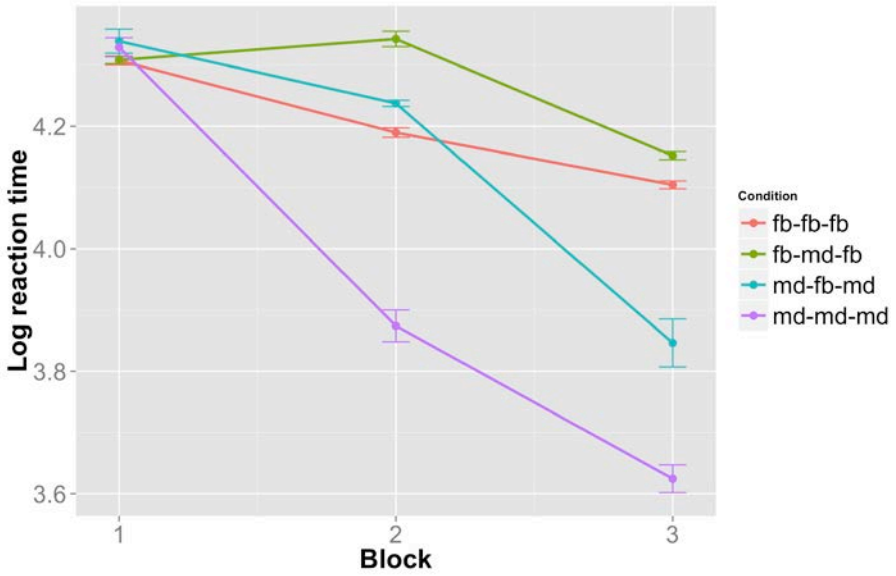

Fig. 1. The log-transformed reaction times for each condition. The standard error is depicted in the figure as error bars.

3 Simulation results

The transfer for the false-belief task on the marble drop game is 11.14%. The transfer for the marble drop game on the false-belief task is 44.2%. As can be seen in Fig. 1, the last block of the *fb-md-fb* condition (green line) lies between the second and third

block of the *fb-fb-fb* condition (red line). A t-test shows that the reaction times indeed differ for the second control block and the second experimental block ($t = -3.67$, $p > 0.002$). In contrast, the error bars of last block in the *md-fb-md* condition (blue line) do overlap with the error bars of the second block of the control condition *md-md-md* (purple line). Also, a t-test does not reveal any significant differences ($t = -0.59$, $p = 0.563$).

4 Discussion and outlook

Transfer is found for the marble-drop task onto the false-belief task, but not the other way around. This finding could be explained by the complexity of both tasks. The marble drop task is a more complex task and requires more use of the working-memory than the false-belief task—thus the model gets more training in specific working-memory strategies when doing marble drop. For developmental studies, the difference in complexity might mask transfer-effects. A recent study showed indeed that there is no transfer from the false-belief task onto the marble-drop game [11]. Whether there is transfer the other way around still remains an open question.

Further eye-tracking studies could help validate or invalidate proposed models. Furthermore, learning effects that occur in the last block could potentially inflate the amount of transfer found. Modelling a non-related task of the similar complexity that also demand inhibition and recursion could control these issues.

5 References

1. Kloo, D. & Penner, J.: Training Transfer Between Card Sorting and False Belief Understanding: Helping Children Apply Conflicting Descriptions, Child Dev. 74, 1823-1839 (2003)
2. Santiesteban, I., White, S., Cook, J., Gilbert, J.S., Heyes, C., Bird, G.: Training Social Cognition: From Imitation to Theory of Mind. Cognition 122, 228-235 (2012)
3. 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, e45961 (2012)
4. Woodworth, R.S., & Thorndike, E.L.: The Influence of Improvement in One Mental Function Upon the Efficiency of Other Functions, Psych. Rev. 8, 247-261 (1901)
5. Singley, M.K., Anderson, J.R., The Transfer of Text-editing Skill. Int. J. Man. Mac. Stud. 22, 403-423 (1985)
6. Taatgen, N.A.: The Nature and Transfer of Cognitive Skills. Psychol. Rev. 120, 439-571 (2013)
7. Anderson, J.R.: How can the Human Mind Occur in the Physical Universe? Oxford university press, New York (2007)
8. Bergwerff, G., Meijering., B., Szymanik, J., Verbrugge, R., Wierda, S.M.: Computational and Algorithmic Models of Strategies in Turn-based Games. In P. Bello et al. (Eds) Proceedings of the 36th Annual Meeting of the Cognitive Science Society (2014).
9. Szymanik, J., Meijering, B., Verbrugge, R.: Using Intrinsic Complexity of Turn-taking Games to Predict Participants' Reaction Times. In R. West & Stewart (Eds.), Proceedings

of the 12th International Conference on Cognitive Modeling, pp. 1426-1431. Carleton University, Ottawa (2013)
10. Arslan, B., Taatgen, N.A., Verbrugge, R.: Modeling Developmental Transitions in Reasoning about False Beliefs of Others. In R. West & Stewart (Eds.), Proceedings of the 12th International Conference on Cognitive Modeling, pp. 77-82. Carleton University, Ottawa (2013)
11. Arslan, B., Verbrugge, R., Taatgen, N.A., Hollebrandse, B.: Teaching Children to Attribute Second-order False Belief: A Training Study. In J. Szymanik & R Verbrugge (Eds.), Proceedings Second Workshop 'Reasoning about Other Minds: Logical and Cognitive Perspectives': CEUR proceedings. (forthcoming)