Cognition 126 (2013) 441–458

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

A constructivist connectionist model of transitions on false-belief tasks

Vincent G. Berthiaume a,ÿ , Thomas R. Shultz a,b , Kristine H. Onishi

aDepartment of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, QC, Canada H3A 1B1
B School of Computer Science, McGill University, 3480 University, Montreal, QC, Canada H3A 2A7

article info

Article history:
Received 8 December 2009
Revised 24 October 2012
Accepted 15 November 2012
Available online 5 January 2013

Keywords:
theory of mind
False-belief tasks
computational models
connectionism
Sibling-descendant cascade-correlation

abstract

How do children come to understand that others have mental representations, eg, of an object's location? Preschoolers go through two transitions on verbal false-belief tasks, in which they have to predict where an agent will search for an object that was moved in her absence. First, while three-and-a-half-year-olds usually fail at approach tasks, in which the agent wants to find the object, children just under four succeed. Second, only after four do children succeed at tasks in which the agent wants to avoid the object. We present a constructivist connectionist model that autonomously reproduces the two transitions and suggests that the transitions are due to increases in general processing abilities enabling children to (1) overcome a default true-belief attribution by distinguishing false from true-belief situations, and to (2) predict search in avoidance situations, where there is often more than one correct, empty search location. Constructivist connectionist models are rigorous, flexible and powerful tools that can be analyzed before and after transitions to uncover novel and emergent mechanisms of cognitive development.

1.Introduction

In most social interactions we rely on our understanding ing of other people's mental states, an understanding usu ally referred to as a Theory of Mind (ToM; Premack & Woodruff, 1978). For instance, if Anne has a theory of mind and sees Sally pointing at some marbles, she might under stand that Sally has the mental state "desire to play with marbles", and offer to play with Sally, or she might hide the marbles to trick Sally. Indeed, we spontaneously attribute mental states to interpret the behavior of other people and animals. How do we arrive at this key under standing that others have mental states?

A standard view is that during preschool years there is a fundamental change in children's understanding of mental states, where they would learn that beliefs underlie the behavior of others (eg, Flavell, Green, Flavell, Harris, & Astington, 1995; Gopnik, 1996; Hedger & Fabricius, 2011; Perner, Rendl, & Garnham, 2007; Sobel, Buchanan, Butterfield, & Jenkins, 2010; Wellman & Cross, 2001; Wellman, Cross, & Watson, 2001; Wellman & Woolley, 1990). This view arose with the discovery of developmental transitions on tasks in which preschoolers are asked to predict the actions of agents that have false, out-dated mental representations. In a standard, approach version of this task (eg, Baron-Cohen, Leslie, & Frith, 1985), partici pants are asked to say where an agent will search for an attractive object that was moved from location A to location B in her absence. Transition 1 occurs when children change from incorrectly predicting that the agent will search in B, to correctly predicting that she will search in A (omniscient to-representational transition). Transition 2 occurs when children who succeed at approach tasks change from failure to success at avoidance tasks, in which the agent wants to avoid the object (approach-to-avoidance transition).

An alternative view is that the transitions are due to abilities distinct from the understanding of beliefs (eg,

ÿ Corresponding author. Present address: Department of Psychology, 1228 Tolman Hall, University of California, Berkeley, CA 94720-1650, United States. Phone: +1 510 859 8557.

Email addresses: vincent.berthiaume@mail.mcgill.ca (VG Berthiaume), thomas.shultz@mcgill.ca (TR Shultz), kris.onishi@mcgill.ca (KH Onishi).

0010-0277/$ - see front matter 2012 Elsevier BV All rights reserved. http://dx.doi.org/10.1016/j.cognition.2012.11.005

Bull, Phillips, & Conway, 2008; Carlson, Mandell, &Williams, 2004; Carpenter, Call, & Tomasello, 2002; de Villiers, 2007; Flynn, 2007; Flynn, O'Malley, & Wood, 2004; Frye, Zelazo, & Burack, 1998; Gordon & Olson, 1998; Hughes, 1998; Leslie, Friedman, & German, 2004; Lohmann & Tomasello, 2003; Pellicano, 2007; Riggs, Peterson, Robinson, & Mitchell, 1998; Roth & Leslie, 1998; Russell, 2007; Sabbagh, Xu, Carlson, Moses, & Lee, 2006; Zaitchik, 1990). Theoretical and computational models have attempted to adjudicate between the two views (Goodman et al., 2006; Leslie, German, & Polizzi, 2005; O'Loughlin & Thagard, 2000; Triona, Masnick, & Morris, 2002). However, our understanding of the transitions remains limited in that previous models built in assumptions and/or structures and transitions that, ideally, a more autonomous model would discover on its own.

To enhance our understanding of themechanisms underlying the transitions in children, we implemented a computational model of false-belief tasks using constructive neural networks. Among the advantages of computational models is that they can provide novel, emergent insights into mech anisms of change (see Harvey, Paolo, Wood, Quinn, & Tuci, 2005). Specifically, longitudinal testing can be done (without attrition, unlike humans) through testing at different points throughout training, and it is possible to analyze the model's internal structure to better understand the mecha nisms underlying task performance.

Our model, which avoids some of the assumptions built into prior models, succeeds at the false-belief tasks and suggests that success at these tasks requires more than using only simple associations (cf. Perner & Ruffman, 2005). As the model recruits hidden units, it autonomously reproduces the two transitions, and analyzes before and after the transitions provide novel insights into possible develop mental mechanisms. Our model suggests that children's de fault true-belief attribution may be due to observing more true-than false-belief search situations, and that the omni scientific-to-representational transition may be due to children overcoming this default true-belief attribution by distin guishing between true- and false-belief situations. Our model suggests that the approach-to-avoidance transition may be due to avoidant behaviors being harder to predict than approach behaviors because they are more variable.

1.1. The false-belief task transitions

Two transitions have been found in standard false-be lief tasks. Transition 1, the omniscient-to-representational transition, is consistent with changing from an omniscient ToM (others always know the true state of the world), to a representational ToM (others rely on representations that may represent the world accurately or not). This transition is robustly observed (see meta-analysis by Wellman et al., 2001 on the standard verbal false-belief task (Baron-Cohen et al., 1985), in which participants see a puppet, Sally, put a marble in a basket. Sally then leaves, and while she is gone puppet Anne moves the marble into a box, causing Sally to falsely believe it is still in the basket. When asked where Sally will search for the marble, children under 3 years and 8 months typically say she will search in the box (omniscient prediction) while older children predict search in the basket (representational prediction).

Transition 2, the approach-to-avoidance transition, is a change from succeeding only at tasks involving a desire to approach an object to succeeding at tasks that involve desires to either approach or avoid an object. For example in an avoidance task, children were told a story about Sally wanting to avoid putting a fish in the box containing a sick kitten (Leslie et al., 2005). While Sally is gone, the kitten moves from one box to another, leading Sally to hold a false belief about its location. Participants were then asked to predict where Sally would put the fish. Four-year-olds, who previously succeeded at a standard approach task, generally performed at chance level or lower. After 4 years, children transition from being able to solve only approach tasks to solving both approach and avoidance tasks (Cassidy, 1998; Friedman & Leslie, 2004a, 2004b, 2005; Leslie & Polizzi, 1998).

1.2. Previous models of false-belief task transitions

False-belief task transitions have been modeled in 4 theoretical or computational models.[1] Leslie et al. (2005) schematic description or theoretical model includes two processes: the Theory of Mind Mechanism (ToMM) provides children with all plausible beliefs (eg, marble is in the box, and the marble is in the basket), while the Selection Processor (SP) selects the particular belief to attribute to others. Based on the argument that everyday beliefs are generally true, attribution of true beliefs was assumed to be a default. The omniscient-to-representational transition was said to arise from the SP changing from being unable to inhibit, to being able to do one inhibition (ie, of the default true belief). The approach-to-avoidance transition was said to arise from the SP becoming able to perform two simultaneous inhibitions (ie, additionally inhibit the believed location of the object), since avoidance-task success was said to depend on first identifying the believed location of the object then inhibiting that location in favor of another. In sum, this theoretical model explains both task transitions, but because it is not implemented it may reflect the authors' preconceptions more than if the proposed processes emerged autonomously within an implemented model.

Two implemented, computational models reproduced the first transition through manipulation of a specific value in the model. First, O'Loughlin and Thagard (2000) built a constraint-satisfaction network, in which nodes represented propositions related to the false-belief task. Propositions that cohered (eg, "Sally puts marble in basket", "Sally thinks marble is in basket") were connected with positive/excitatory weights while the two propositions that did not cohere (ie, "Sally searches in basket", "Sally searches in box"), were connected with a negative/inhibitory weight. When connection weights controlling the false-belief search location were low (default), the true-belief search (eg, search in basket) was activated. However, when those weights were increased, the false-belief search location became activated and the true-belief location

inhibited since the two search-location propositions were

[1] Van Overwalle's (2010) auto-associative neural network and Wahl and Spada's (2000) symbolic inference model simulate success on false-belief tasks, but no transitions.

inconsistent, leading to a successful transition. Second, Triona et al. (2002) implemented an ACT-R production system model (eg, Anderson et al., 2004). The model was given facts (eg, "Sally is an agent" and "Sally puts marble in bas ket") and rules (eg, "Agents usually search for objects where they are"), and new facts were produced as output (eg, "Sally will search in the basket"). One rule implemented mented omniscient predictions and another rule implemented mented representational predictions, and the transition was created "by manipulating ... the probability that the production would achieve the goal" (p. 1045). That is, when the probability parameter was low, the output tended to be wrong (omniscient), but when the parameter was high, the output tended to be correct (representational). Therefore, while both models reproduced Transition 1, development was not autonomous because the transition was due to direct experimenter manipulation of either weights or a probability parameter.

A third implemented model autonomously reproduced Transition 1, but within a limited search space (Goodman et al., 2006). Two Bayesian networks were constructed. In the omniscient network, Sally's belief depended only on the marble's location, whereas in the representational net work it also depended on Sally's visual access to the marble's displacement. The model initially favored the omniscient network because it was more parsimonious, but later favored the representational network, because it was consistent with more search data. The transition was thus autonomous, but given that there were two networks, the search space was limited to two transitions (omniscient to-representational or representational-to-omniscient).

Thus, the existing models of false-belief task transitions reproduced the first transition, but not the second. In all cases, experimenters implemented specific false-belief task information while transitions were accomplished through stipulation (Leslie et al., 2005), direct manipulation of a value (O'Loughlin & Thagard, 2000; Triona et al., 2002), or selection from a limited set of pre-determined options (Goodman et al., 2006). In Experiment 1, we present a fully implemented model that succeeds at a false-belief task after autonomously passing through Transition 1.

2. Experiment 1: omniscient-to-representational transition

To better understand the mechanisms underlying the omniscient-to-representational transition, we imple mented in Experiment 1, a constructive neural network of a false-belief task with approach desires.

For simplicity, we modeled a version of the task (Onishi & Baillargeon, 2005) which was non-verbal,2 featured a single protagonist rather than two, and used a continuous looking time measure rather than a dichotomous verbal response, as the former is more naturally modeled by neural network output. In that non-verbal task, participants (15-month-old infants) watched an agent hide an object in one of two boxes (green, yellow). Next, they saw one of four belief-induction trials, which led the agent to hold a true or false belief that the object was in the green or yellow box. For instance, some participants saw the agent watch the object move from green to yellow (true belief that object is in yellow), while others saw that the agent was absent as the object moved (false belief that object is in green). Similarly, the other two belief-induction trials induced a true belief that the object was in yellow and a false belief that it was in green. Finally, each participant saw one of two test trials in which the agent searched in either green or yellow. Part ipants looked reliably longer when the agent did not search according to her belief, whether true or false. When the agent had a true belief, infants looked longer for search in the empty box than in the box containing the object. When the agent had a false belief, participants showed the reverse pattern: looking longer when she searched in the box containing the object than the empty one. This is the pattern we will look for as a marker of success in our approach to false-belief task.

2.1. Method

Our simulations used Sibling-Descendant Cascade Cor relation (SDCC; Baluja & Fahlman, 1994), a constructive neural network algorithm with supervised learning. Compared to backpropagation networks (eg, Rumelhart, McClelland, & PDP Research Group, 1986), constructive networks involve less experimenter design (because there is no need to specify the internal network topology) and tend to cover developmental changes better (because hid den unit recruitment produces qualitative performance changes that can produce developmental transitions, eg, Shultz, 2003, 2006; Shultz & Cohen, 2004; Shultz, Mareschal, & Schmidt, 1994). In fact, even though back propagation is a robust algorithm that has been used to simulate many developmental phenomena (eg, Elman et al., 1996; Thomas & Karmiloff-Smith, 2003), backpropagation networks did not learn to either produce omniscient or representational ToM predictions nor transitions when trained on the training sets of either Experiment 1 or 2, as predictions stagnated quickly near the beginning of training (see Appendix A for details).

Our networks' inputs and outputs can respectively be thought of as simulating perception and prediction (mea sured through looking time). Because it is unlikely that children learn about search behavior during false-belief tasks, network training simulated pre-experiment, every day experience with search behavior, while network testing ing simulated performance on the false-belief task. The basic structure of a network is shown in Fig. 1.

2.1.1. network input

Input represented information useful for predicting search and included 4 start locations and 4 end locations since in everyday life there are almost always more than two locations for objects (although to simulate perfor mance in the false-belief task, network testing included only two locations) . Object location was encoded by activation of 1.0 (with 0.0 for other locations). The agent

[2] The fact that 3-year-olds perform better on verbal false-belief tasks if the word "first" is added in the false-belief question, as in "Where will Sally look first for her marble?" (Siegal & Beattie, 1991; Surian & Leslie, 1999) suggests that children might not fully understand the linguistic distinction between "look for" and "find" (Bloom & German, 2000).

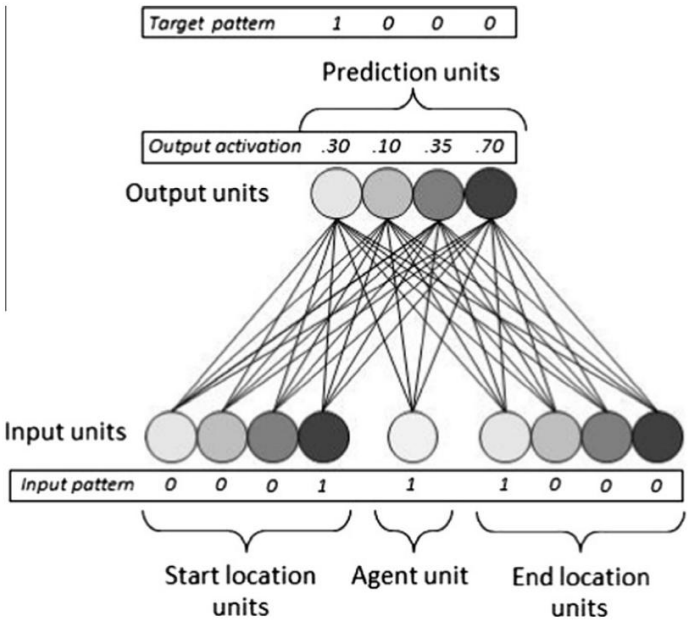

Fig. 1. Initial model and one training pattern (combination of input and target patterns). The initial network had 9 input, 4 output and 0 hidden units. In the displayed input pattern, the agent is watching as the object moves from green (location 4) to yellow (location 1). The target pattern (correct in this case) is search in yellow, while the network's output activation predicts search in green.

input unit encoded whether the agent was watching (activation 1.0) or not (activation 0.0) as the object moved or not.

2.1.2. Network output

Four output units represented a weighted prediction of search, ie, how strongly the model predicted search in each of the four locations. For instance, a network having the out put activations: yellow = 0.30, red = 0.10, blue = 0.35, and green = 0.70, predicted search most strongly in the green box, somewhat intermediately in yellow or blue, and most weakly in red.

2.1.3. Network training

Before training, networks contained only input and out put units fully interconnected with random weights (which could be positive, ie, excitatory, or negative, ie, inhibiting); thus initially, every input pattern resulted in random output activation. During training, networks were given sets of training patterns (combinations of input patterns and tar get patterns). After each epoch (one pass through all training patterns) of training, connection weights were modified to reduce output error(the discrepancy between output activations produced by the network and target patterns). When output error stagnated (failed to reduce by 1% over 8 epochs), a hidden unit was recruited from a pool of 8 ran domly initialized candidates. Training continued with the newly incorporated hidden unit until error reduction stagnated even when recruiting additional hidden units.

2.1.3.1. Training patterns. There were 32 possible input pat terms, obtained by crossing all the input values (4 start locations, 4 end locations, and 2 agent-watching or not). Including twice as many true- as false-belief training patterns,[3] added another 16 possible patterns (4 start and 4 end locations, 1 agent watching), yielding 48 input pattern instances. Since in everyday life, people do not always search for objects correctly (eg, due to forgetting, distraction), network training was stochastic. That is, each input pattern occurred most often with target patterns for correct (representational) search and sometimes with target pat terns for incorrect search. Specifically, each input pattern oc curred with the correct target pattern 18 out of 21 times (simulating correct search 85.7% of the time) and with incorrect target patterns 3 out of 21 times (search 4.7% of the time in each of the 3 incorrect locations) resulting in 1008 training pattern instances.

2.1.3.2. Output of training. Because output units encoded weighted predictions of search, after successful training, output activations should match training frequencies. that

[3] When training included equal numbers of true- and false-belief patterns (respectively defined as patterns in which the agent input was activated or not activated), with 0 hidden units, networks failed to show omniscient predictions: Error in the true-belief condition was lower for search in the object, Mobject = .13, SD = .13, than in the empty location, Mempty = .56, SD = .31, F(1, 26) = 22, p < .001, g2 = . 46, but in the false-belief condition error for search in the object, Mobject = .25, SD = .01, and empty locations, Mempty = .25, SD = .01, did not differ, F(1, 26) < 1. After recruiting 1 hidden unit, the networks showed representational predictions (true-belief condition: Mobject = .08, SD = .10, Mempty = .66, SD = .29, F(1, 26) = 50, p < .001, g2 = .66; false-belief condition: Mobject = .42, SD = .08, Mempty = .12, SD = .06, F(1, 26) = 118, p < .001, g2 = .82). Networks thus succeeded at the approach task but failed to cover Transition 1. For this reason, training for Experiments 1 and 2 included twice as many true- as false-belief training patterns, consistent with a default true-belief assumption (Leslie et al. , 2005), which seems plausible in that people's beliefs about the location of objects are generally true (see also Fodor, 1992).

is, although during training networks were presented with individual search targets, ie, 1.0 for the search location and 0.0 for the other locations, after training, networks were expected to produce, for each input pattern, activations of approximately 0.85 at the correct search location and 0.05 at the three other output units.

2.1.4. Network testing

To determine whether networks and developmen such changes, theywere assessed at multiple test points, ie, after each hidden unit was recruited. Paralleling the infant experiment (Onishi & Baillargeon, 2005), (1) each network was tested on one test pattern that corresponded to one of the 8 infant task conditions (2 belief conditions by 2 belief locations by 2 search locations), (2) 7 networks were tested in each condition (for a total of 56 networks in Experiment 1), and (3) only two start, end, and search locations were used (although training had four locations). To prevent learning from testing patterns, connection weights were frozen during test.

2.1.4.1. Calculation of output error.

Output error at test was an indication of distance between the network's search prediction and a specific test pattern, and calculated using the following formula:

½ðOg Tg Þ¯ þ ðOy Tyþ ¯ =2¼ ½ð:8571Þ¯ þ ð:0480Þ ¯ =2¼:011 ð1Þ

That is, output error is the mean (across the two output units) of the squared difference (for each unit) between out put activation (O; the network's prediction of search) and the target pattern (T; the actual location of search in test).
The subscripts g and y represent the green and yellow boxes respectively.

For networks in the ''search in green'' test conditions, we measured output error relative to a search-in-green target pattern (target output activation: green = 1.0, yellow low = 0.0), while for ''search in yellow '' conditions, output error was calculated relative to a search-in-yellow target pattern (target output activation: green = 0.0, yellow low = 1.0). With the example post-training output activations shown in Eq. (1), search-in-green yields output error of .011, while search-in-yellow yields .820. Thus, with these example values, output error (distance from prediction) would be greater for search in yellow than for search in green, consistent with predicting search in green.

2.2. Results

2.2.1. Analysis plan

Analyzes of variance ANOVAs were performed on out put error at each test point with the factors of belief condition (true, false) and search location (object, empty).
Search location was a factor that collapsed over box color (green, yellow), unimportant here, to obtain object search (search in green when object is in green and in yellow when it is in yellow) and empty search (search in yellow when object is in green and in green when it is in yellow).
For conciseness, we report results only for test points that show a change in the pattern of predictions compared to the previous test point.

2.2.2. Predictions

If networks had predictions consistent with a representational ToM, they should predict search in different locations depending on whether the agent watched the object interaction move or not, thus producing a significant belief condition by search location in which error is lower in the location in which the model predicts search is more likely. Interactions were explored using planned comparison sounds, and because of a lack of homogeneity of variance, were confirmed using Mann–Whitney U non-parametric tests throughout. For a representational ToM, in the true-belief condition, error should be lower for search in the object than the empty location, but in the false-belief condition, the pat tern should be reversed: lower error for search in the empty than the object location, because the agent would not know that the object had moved. For an omniscient ToM, in both belief conditions, error should be lower for search in the object than in the empty location.

2.2.3. Zero hidden units

With 0 hidden units, networks had omniscient predictions (Fig. 2). Although the belief condition by search location interaction was significant, $F(1,52) = 19$, $p < .001$, planned comparisons showed lower error for search in the object than in the empty location for both true-, Mobject = .07 , SD = .07, Mempty = .64, SD = .23, $F(1,26) = 77$, $p < .001$, $g2 = .75$, and false-belief conditions, Mobject = .13, SD = .01 , Mempty = .42, SD = .01, $F(1,26) = 3606$, $p < .001$, $g2 = .99$. Therefore, having more true-than false-belief search in training enabled networks to predict true-belief search by default, ie, to show predictions consistent with an omniscient ToM.

2.2.4. One hidden

unit On average,4 after recruiting 1 hidden unit, networks showed predictions consistent with representational ToM (Fig. 3), a pattern of results reliably different from predictions with 0 hidden units, as indicated by a significant three-way interaction between number of hidden units, be lief, and search, $F(1,52) = 74$, $p < .001$. Specifically, with 1 hidden unit, the belief by search interaction was significant, $F(1,52) = 154$, $p < .001$. Error in the true-belief condition was lower for search in the object, Mobject = .04, SD = .03, than the empty location, Mempty = .76, SD = .16, $F(1, 26) = 285$, $p < .001$, $g2 = .92$, but in the false-belief condition was lower for search in the empty, Mempty = .26, SD = .18, than the object location, Mobject = .40, SD = .10, $F (1.26) = 6.7$, $p < .05$, $g2 = .20$. Thus, by recruiting a hidden unit, networks reproduced the omniscient-to-representational transition.

[4] Although the predictions of networks with 1 hidden unit were consistent with task success on average, network predictions taken individually showed a more gradual pattern of success over the duration of training (which, based on stagnation of the reduction in output error, continued until M = 7.14, SD = .75, range = 6–9 hidden units had been recruited, and took M = 744.64, SD = 76.93, range = 588–905 epochs). That is, no individual network succeeded at all test patterns with 1 hidden unit, but rather succeeded only on some of them, albeit enough to show average success at the task. The proportion of successfully learned patterns increased gradually until the end of training, at which point all networks successfully learned all training patterns.

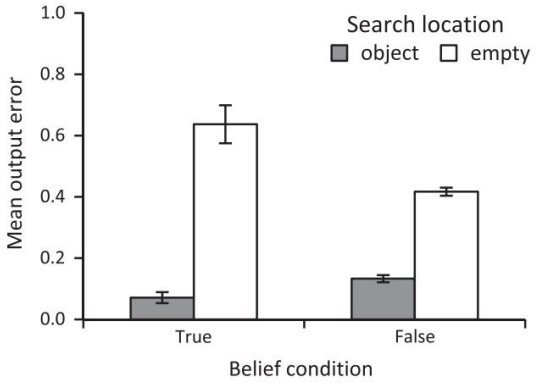

Fig. 2. Mean output error and standard error (SE) bars for networks with 0 hidden units, consistent with omniscient predictions.

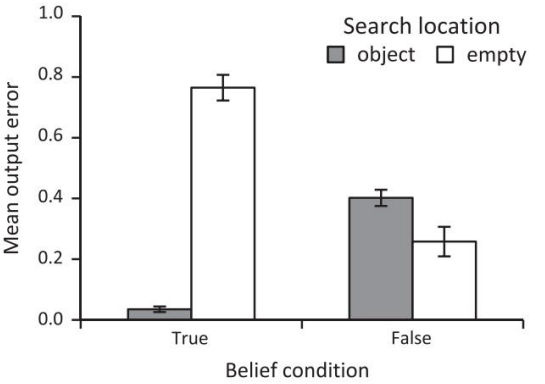

Fig. 3. Mean output error and SE bars for networks with 1 hidden unit, consistent with success at the approach task.

2.2.5. Output contribution analyzes

To explore the developmental mechanisms underlying Transition 1, we analyzed networks' internal structure be fore and after the transition using Principal Component Analyzes (PCAs) of output contributions (following Shultz & Elman, 1994; Shultz, Oshima-Takane, & Takane , 1995).

An output contribution is defined as the product of a sending unit's activation and the weight connecting that unit to an output unit. For each network, at each test point, a matrix of output contributions can be obtained with rows for each training pattern, columns for each connection entering the network's outputs, and the output contributions in the cells of the matrix. We used PCA on this matrix to reduce the number of column variables (in our case, out put connection weights) by computing components that capitalize on correlations between variables (Cattell, 1966). PCA provides two measures that help to interpret the structure of the matrix and thus the internal structure of individual networks: (1) component scores are given to each training pattern and show how patterns are categorized by the principal components, and (2) component load ings are given to each output weight and show which weights most strongly contribute to each component.

PCA was done on five networks (selected from the 56), both with 0 and 1 hidden unit. At 0 hidden units, Component 1 for each network categorized training patterns based on end location, eg, patterns where the object ended in yellow all had negative Component 1 scores, while end-in-green patterns had positive scores. The out put weights that most contributed to Component 1 were those for end locations, eg, weights from end-in-yellow and end-in-green. Thus, before recruiting a hidden unit, networks relied on the end locations to make search predictions, which makes sense given that this heuristic makes the right prediction for 2/3 of the training patterns (true-belief patterns). While networks varied as to the specific locations that Component 1 categorized, all five networks with 0 hidden units categorized patterns by their end locations.

In contrast, when networks had recruited 1 hidden unit, Component 1 categorized patterns based on belief, eg, false-belief patterns had negative scores while true-belief patterns had positive ones. For each network, the output weights from the agent and hidden unit most contributed to Component 1, indicating that the categorization of the patterns relied on a combination of these units. All five analyzed networks assigned different values to the true- and false-belief patterns, although the specific values differed (eg, positive vs. negative scores). Additional computational details are provided in Appendix B.

Output contribution analyzes thus showed that net works producing omniscient predictions categorized training patterns based on the end location of the object. This default attribution was due to more training on true-than false-belief situations, and this default was overcome using the agent unit and a recruited hidden unit to distinguish between true- and false-belief patterns, enabling representational ToM predictions.

2.3. Discussion

previous computational models of false-belief tasks have either not covered the transitions (Van Overwalle, 2010; Wahl & Spada, 2000) or did so through parameter manipulation (O'Loughlin & Thagard, 2000; Triona et al., 2002) or by using experimenter-designed topologies (Goodman et al., 2006), our model is the first to autonomously build structures to produce and transition between predictions consistent with omniscient and representational ToM on an approach false-belief task.

Analysis of output contributions revealed that when making predictions consistent with omniscient ToM, net works categorized training patterns based on their end location and initially produced omniscient predictions by relying on end-location connections. Since correct true belief search is in the end location, it makes sense that training on more true-than false-belief situations caused networks to make predictions relying mostly on end locations.

With additional training, networks transitioned to representational ToM predictions. Output contribution analyzes indicated that in making these predictions, networks relied on the agent and hidden units to distinguish all true from false-belief situations (even though 15% of training patterns involved an incorrect search outcome), an important conceptual distinction for making predictions

consistent with a representational ToM. In theory, this distinction was not necessary; networks could have learned to predict search on a pattern-by-pattern basis, without categorizing patterns in this (or any) way but rather rote memorizing the outcome of each pattern. Our networks thus suggest a novel mechanism underlying Transition 1, specifically, the transition might be supported by learning to distinguish between false- and true-belief situations.

Leslie et al.'s (2005) theoretical model assumed that, at the functional level, inhibitory abilities would play a critical role in the false-belief-task transitions. At a functional level, any model that covers a change from default predictions to predictions based on the default and other factors may be argued to implement inhibition, thus our model might be seen as implementing functional inhibition of the default true-belief attribution. However, at the implementation level, we did not introduce inhibition to cover the transition. Instead, each network implemented unique combinations of inhibitory and excitatory connection weights from the beginning and throughout development.

Experiment 1 showed that our autonomously developing model covered the omniscient-to-representational transition, and suggests that a mechanism for the transition is discovering a distinction between true- and false belief situations. Since children go through two transitions, we next added an avoidance task to our model, to determine mine whether a single model could reproduce both transitions and in the expected order.

3. Experiment 2: approach-to-avoidance transition

3.1. Method

Experiment 2 used the same method as Experiment 1, except that avoidance search patterns were included in training and testing. A total of 112 networks were trained, 56 in the approach task and the other 56 in the avoidance task.

3.1.1. Network input

The object to be avoided was encoded as 1.0 to distinguish guish it from the approach object (still encoded as 1.0).

3.1.2. Network output

Network output again represented the networks' weighted prediction of search in each location. For ap proach objects, search should tend to be in the believed location of the object, while for avoidance objects, search should tend to be anywhere but the believed location of the object.

3.1.3. Network training

Initialization and training of networks was as in Expert ment 1, except as noted.

3.1.3.1. Training patterns. As before, training included twice as many true- as false-belief patterns, and was stochastic, with each input pattern being matched with 18 correct (85.7%) and 3 incorrect (14.3%) search outcomes. There were a total of 1008 approach and 1008 avoidance training pattern instances. In avoidance, the 18 correct searches were equally divided among the 3 correct avoidance locations (because to avoid an object, one can correctly search in all locations that do not contain the object), while the 3 incorrect searches were in the actual location of the object. The distribution of search across the 4 locations was thus different for approach and avoidance: in approach, the 18 correct searches were all in one location, whereas in avoidance, the 18 correct searches were spread across 3 locations.

3.1.3.2. Output of training. In the approach task, expected activations were as in Experiment 1. In the avoidance task, activations of approximately 0.286 (= .857/3) at the three correct locations and 0.143 at the incorrect output unit were expected.

3.1.4. Network testing

Seven networks were each tested under one of 16 conditions (2 belief conditions by 2 belief locations by 2 search locations by 2 tasks) with two locations.

3.1.4.1. Calculation of network error. Mean output error was calculated using Eq. (1). Equity (2) shows an example error calculation for a post-training network tested on the input pattern avoidance object starts in green and ends in yellow with agent not watching, with a correct target search (yel low location).

½ðOg Tg Þ = þ ðOy TyÞ =2 ¼ ½ð:143 0Þ ¼ þ ð:286 1Þ =2 :265 ð2Þ

Using the same input example, but for an incorrect search in green, yields the higher error of .408.

3.2. Results

3.2.1. analysis plan

ANOVAs with the factors of belief condition (true, false) and search location (object, empty) were performed inde pendently for approach and avoidance tasks at each test point and again we report results for test points that showed changes in prediction patterns.

3.2.2. Predictions

In each task, representational ToM predictions should yield different interactions between belief condition and search location. For approach, network error in the true belief condition should be lower for search in the object location, and in the false-belief condition it should be lower in the empty location. For avoidance, the reverse patterns should hold: error in the true-belief condition should be lower for search in the empty location, and in the false-belief condition it should be lower in the object location.

3.2.3. Zero hidden units

For both tasks, with 0 hidden units, networks had omni scient ToM predictions (Fig. 4). Network predictions reliably differed between tasks, as the three-way interaction between task, belief condition and search location was significant, $F(1,104) = 46$, $p < .001$. For the approach task, the

Fig. 4. Mean output error and SE bars for networks with 0 hidden units, indicating omniscient predictions in both approach and avoidance tasks.

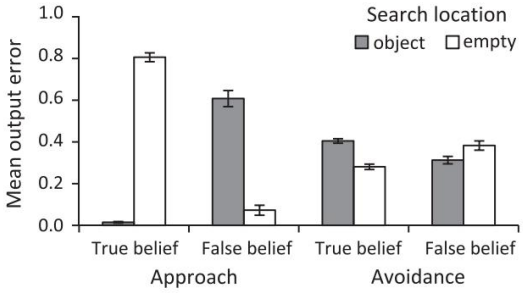

Fig. 6. Mean output error and SE bars for networks with 6 hidden units, consistent with representational predictions for both approach and avoidance tasks.

belief condition by search location interaction was significant, $F(1,52) = 24$, $p < .001$, but planned comparisons showed lower error for search in the object than the empty location for both true-, Mobject = .09, SD = .05, Mempty = .55, SD = .13, $F(1,26) = 142$, $p < .001$, $g2 = .85$, and false-belief conditions, Mobject = .14, SD = .01, Mempty = .41, SD = .01, $F(1,26) = 4637$, $p < .001$, $g2 = .99$.

For the avoidance task, the belief condition by search location interaction was also significant, $F(1,52) = 26$, $p < .001$, but error was now lower for search in the empty than the object location for both true-, Mempty = .29, SD = .07, Mobject = .50, SD = .04, $F(1,26) = 89$, $p < .001$, $g2 = .77$, and false-belief conditions, Mempty = .36, SD = .004, Mobject = .46, SD = .004, $F(1,26) = 3878$, $p < .001$, $g2 = .98$.
The model could thus handle simultaneous approach and avoidance tasks, showing omniscient predictions for both.

3.2.4. Three hidden units

After recruiting 3 hidden units, networks showed predictions consistent with representational ToM for the ap proach, but not the avoidance task (Fig. 5), a pattern of results reliably different from predictions with 0 hidden units, as indicated by a significant three- way interaction between number of hidden units, belief, and search, $F(1,108) = 30$, $p < .001$. Network predictions reliably differed between tasks, as the three-way interaction between task, belief condition and search location was significant, $F(1,104) = 159$, $p < .001$. For approach, the belief by search interaction was significant, $F(1,52) = 141$, $p < .001$, and in the true-belief condition error was lower for search in

the object, Mobject = .03, SD = .03, than the empty location,

Fig. 5. Mean output error and SE bars for networks with 3 hidden units, consistent with representational predictions for the approach, but not the avoidance task.

Mempty = .70, SD = .11, $F(1,26) = 480$, $p < .001$, $g2 = .95$, but in the false-belief condition, the reverse pattern was found, Mempty = .19, SD = .14, Mobject = .41, SD = .21, $F(1,26) = 10$, $p < .005$, $g2 = .28$. For the avoidance task, the belief by search interaction was also significant, $F(1,52) = 20$, $p < .001$, and in the true-belief condition error was lower for search in the empty, Mempty = .29, SD = .06, than the object location, Mobject = .42, SD = .05, $F(1,26) = 36$, $p < .001$, $g2 = .58$, but in the false-belief condition the locations did not differ, Mempty = .39, SD = .08, Mobject = .36, SD = .07, $F(1,26) < 1$.

3.2.5. Six hidden

units After recruiting 6 hidden units, networks showed predictions consistent with representational ToM for both tasks[5] (Fig. 6), a pattern of results reliably different from predictions with 3 hidden units, as indicated by a significant three-way interaction between number of hidden units, be lief, and search, $F(1,108) = 26$, $p < .001$. Network predictions reliably differed between tasks, as the three-way interaction between task, belief condition and search location was significant, $F(1,104) = 637$, $p < .001$. For the approach task, the belief by search interaction was significant, $F(1,52) = 693$, $p < .001$, and error was lower in the true-belief condition for search in the object, Mobject = .02, SD = .01, than the empty location, Mempty = .81, SD = .08, $F(1,26) = 1355$, $p < .001$, $g2 = .98$, but in the false-belief condition, the re verse pattern was found, Mempty = .07, SD = .09, Mobject = .61, SD = .14, $F(1, 26) = 138$, $p < .001$, $g2 = .84$. For the avoidance task, the belief by search interaction was also significant, $F(1, 52) = 35$, $p < .001$, and error was lower in the true-belief condition for search in the empty, Mempty = .28, SD = .05, than the object location, Mobject = .40, SD = .04, $F(1, 26) = 55$, $p < .001$, $g2 = .68$, but now in the false-belief condition, the reverse pattern was found, Mobject = .31, SD = .07, Mempty = .38, SD = .08, $F(1,26) = 6.1$, $p < .05$, $g2 = .1$

3.2.6. Output contribution analyzes

As for Experiment 1, we performed PCAs on the output contributions of 5 networks from Experiment 2. We

[5] Training continued until M = 7.82, SD = 1.16, range = 5–11 hidden units had been recruited and took between M = 930.51, SD = 128.49, range = 608– 1275 epochs. Again, the proportion of successfully learned training patterns increased gradually during training, until all networks successfully learned all training patterns by the end of training.

analyzed each network with 0, 3 and 6 hidden units. At all 3 test points, each network's Component 1 used all location units to categorize training patterns based on task (ap proach, avoidance), which makes sense given that object location was encoded at the input as +1 and 1 in the ap proach and avoidance tasks, respectively. With 3 and 6 hid den units, but not with 0, PCA revealed a Component 2 which always categorized patterns based on belief (true, false), as had Component 1 in Experiment 1. Thus, output contribution analyzes revealed that networks first distin guished between the approach and avoidance tasks, and next distinguished between true- and false-belief situations, but these analyzes did not distinguish between the two transitions per se (additional computational details are provided in Appendix C). To further explore the mech anisms underlying Transition 2, we performed 2 supple mental simulations.

3.2.7. Exploring Transition 2

There are at least 2 possible explanations for Transition 2. First, perhaps predicting avoidance search is harder than predicting approach behaviors because avoidant behaviors are more variable. Second, analogous to the explanation for Transition 1, there could be a default assumption of ap proach, a default which is overridden with experience.

3.2.7.1. Variability of avoidance search. Whenever there are more than 2 locations and a single object, correct avoid ance search is likely to be more variable than correct ap proach search, since in avoidance search there are more correct search locations (all the empty locations) than in approach search (the 1 location containing the object). To test whether the difference in number of correct-search possibilities were involved in Transition 2, we trained and tested another group of networks with only 2 locations. With 2 search locations, there is 1 correct response for both avoidance and approach tasks.

Networks trained at two locations produced omniscient predictions for both approach and avoidance tasks with 0 hidden units and representational predictions for both tasks with 1 hidden unit.

Equating the variability of correct search in the two tasks led to the two tasks being solved successfully with the same computational power (number of hidden units). This sug gestures that success at approach before avoidance when there were four locations during training, was due to the greater variability of avoidance than approach search.

3.2.7.2. Default assumption of approach. Since including more true-than false-belief training patterns permitted networks to form a default prediction of true-belief search which was then overcome by training, we tested the idea of a default approach bias by including twice as many ap proach as avoidance training patterns (although more experience with approach than avoidance has been argued to be unlikely, eg, Leslie & Polizzi, 1998). Only two locations were used in training, to remove the effect of avoiding ance being more variable than approach.

Networks succeeded at approach and avoidance simultaneously with 1 hidden unit,7 thus failing to capture the second transition and failing to support the idea that this transition may be due to a default assumption of approach search that must be overridden.

3.3. Discussion

Networks in Experiment 2 autonomously reproduced both false-belief task transitions. Our model is the first to cover these two transitions in a unified, implemented model, as previous attempts did not cover both transitions (Goodman et al., 2006; O'Loughlin & Thagard, 2000; Triona et al., 2002), required several networks (Goodman et al., 2006), or were not implemented (Leslie et al., 2005).

Networks learned to categorize patterns first by task (approach, avoidance) and then by beliefs. Evidence for this claim is twofold. First, early in training, networks treated approach and avoidance tasks differently, producing different (albeit omniscient) predictions for each task, while later in training, networks treated belief situations differently to produce correct true- and false-belief predictions in both tasks . Second, PCA of output contributions

[6] Networks trained with two locations showed omniscient predictions in both tasks with 0 hidden units. For approach, error was lower for search in the object location for both true- and false-belief conditions (belief by search interaction: F(1, 52) = 21, p < .001, true-belief: Mobject = .08, SD = .06, Mempty = .56, SD = .18, false-belief: Mobject = .15, SD = .01, Mempty = .39, SD = .02, both Fs(1, 26) > 85, ps < .001, g2 s > .77), while for avoidance, error was lower for search in the empty location for both true- and false-belief conditions (belief by search interaction: F(1, 52) = 20, p < .001, true-belief: Mempty = .08, SD = .06, Mobject = .56, SD = .19, false-belief: Mempty = .15, SD = .01, Mobject = .38, SD = .01 , both Fs(1, 26) > 79, ps < .001, g2 s > .75). Networks showed representational predictions in both tasks with 1 hidden unit. For approach, error was lower for search in the object than the empty location for the true-belief condition, but lower in the empty than the object location for the false-belief condition (belief by search interaction: F(1, 52) = 876, p < .001, true-belief: Mobject = .03, SD = .02, Mempty = .70, SD = .10, false-belief: Mempty = .07, SD = .03, Mobject = .55, SD = .09, both Fs(1, 26) > 321, ps < .001, g2 s > .93), while for avoidance, error was lower for search in the empty than the object location for the true-belief condition, but lower in the object than the empty location for the false belief condition (belief by search interaction: F(1, 52) = 873, p < .001, true belief: Mempty = .03, SD = .02, Mobject = . 69, SD = .09, false-belief: Mobject = .05, SD = .02, Mempty = .56, SD = .12, both Fs(1, 26) > 268, ps < .001, g2 s > . 91).

[7] Networks trained with twice as many approaches as avoidance patterns showed omniscient predictions in both tasks with 0 hidden units. For approach, error was lower for search in the object location for both true and false-belief conditions (belief by search interaction: F(1, 52) = 29, p < .001, true-belief: Mobject = .08, SD = .06, Mempty = .57, SD = .18, false belief: Mobject = .15, SD = .03, Mempty = .35, SD = .05, both Fs(1, 26) > 96, ps < . 001, g2 s > .79), while for avoidance, error was lower for search in the empty location for both true- and false-belief conditions (belief by search interaction: F(1, 52) = 18, p < .001 , true-belief: Mempty = .07, SD = .06, Mobject = .55, SD = .18, false-belief: Mempty = .15, SD = .06, Mobject = .39, SD = .01, both Fs(1, 26) > 85, ps < .001, g2 s > .77). Networks showed representational predictions in both tasks with 1 hidden unit. For approach, error was lower for search in the object than the empty location for the true belief condition, but lower in the empty than the object location for the false-belief condition (belief by search interaction: F(1, 52) = 688 , p < .001, true-belief: Mobject = .03, SD = .01, Mempty = .70, SD = .14, false-belief: Mempty = .07, SD = .03, Mobject = .57, SD = .09, both Fs(1, 26) > 328, ps < .001, g2 s > .93), while for avoidance, error was lower for search in the empty than the object location for the true-belief condition, but lower in the object than the empty location for the false-belief condition (belief by search interaction: F(1, 52) = 567, p < .001, true-belief: Mempty = .03, SD = .02, Mobject = .72, SD = .15, false-belief: Mobject = .07, SD = .03, Mempty = .53, SD = .10, both Fs(1, 26) > 262, ps < .001, g2 s > .91).

revealed that early in training, networks only categorized training patterns based on task, while later in training they also categorized them based on belief. This pattern of cat egorization suggests that when predicting others' search behavior, determining the searcher's desire would have the most predictive power, with determining their belief coming later. The model therefore suggests that a heuristic a learner might use is to initially identify others' desires while ignoring variation in belief and later learn to incorporate information about variation in beliefs.

Additionally, the model suggests a new explanation for the approach-to-avoidance transition. In contrast to the hypothesis that this transition arises when children are able to do two simultaneous inhibitions (Leslie et al., 2005), our results suggest that avoidance search is generally harder to predict than approach search because in avoidance there are generally more correct locations in which one can predict search.

4. General discussion

Our implemented, constructivist model of false-belief tasks is the first to successfully cover the two false-belief transitions observed with preschoolers: (1) from omni scientific predictions to representational predictions only in approach, and (2) from failure in avoidance to represent tional predictions in both approach and avoidance.

In Experiment 1, network training included only ap proach search situations, with twice as many true- as false-belief search patterns. Networks initially produced default omniscient predictions by relying on end location input units. Then with additional training, networks used the agent and hidden units to distinguish false- from true-belief situations and overcome the default true-belief attribution, thus covering Transition 1 to representational predictions. Although it had been previously hypothesized that most beliefs were likely to be true and thus that search behavior would more often be consistent with true rather than false beliefs (eg, Leslie et al., 2005), our model is the first to support this idea computationally.

In Experiment 2, training included both approach and avoidance search situations, still with twice as many true- as false-belief patterns. Networks first used location units to cat egorize training patterns by task, producing outcomes consistent with omniscient predictions for both approach and avoidance tasks. With additional training, networks used the actor and hidden units to categorize patterns by belief, while producing outcomes consistent with representational predictions for both tasks, thus covering Transition 2. Ourmodel sug gests that Transition 2 was due to avoidance search being harder to predict than approach search, thus requiring more computational power to solve correctly. When there are numerous locations for one object, there are many empty locations (and hence possibly many correct avoidance locations) but only one containing the object (hence only one correct ap proach location). Indeed, when the number of correct locations for approach and avoidance search were the same, Transition 2 was not observed. Themodel also provides evidence that Transition 2 is not due to overcoming a default approach attribution, because when the ratio of approach to avoidance training was increased (analogous to increasing the ratio of true- to false-belief search which established a true-belief de fault), Transition 2 was not observed.

Although there have been previous models of false-belief tasks (Goodman et al., 2006; Leslie et al., 2005; O'Loughlin & Thagard, 2000; Triona et al., 2002; Van Overwalle, 2010; Wahl & Spada, 2000), ours is the first to use a structure that was not entirely determined by the experimenters (such as specific initial propositions and specification of connection weights as in O'Loughlin & Thagard, 2000; or probability distributions as in Goodman et al. , 2006) and to learn from stochastic tra Further, our model's success emerges from generic algorithms not customized for false-belief tasks. For example, changes in capacity to inhibit have been raised as an explanation for changes in behavior on false-belief tasks (eg, Leslie et al., 2005), but in our model functional inhibition of the default true-belief location emerged from implemented generic with putational operations that included both inhibitory and excit atory activations from the beginning and throughout network development. Finally, our model was the first to autonomously demonstrate task transitions by learning from relevant experience, as previous models did not implement transitions (Van Overwalle, 2010; Wahl & Spada, 2000), selected from a highly limited set of potential transitions (Goodman et al ., 2006), or had transitions that were implemented through direct manipulation of particular parameter values (O'Loughlin & Thagard, 2000; Triona et al., 2002). Our model thus illustrates the power of constructive neural networks to simulate cognitive development.

Results of our model have implications for the debate about whether infants have an understanding of beliefs or if they succeed at non-verbal tasks by using 3-way associations (Perner & Ruffman, 2005). The idea is that during familiarization (eg, Onishi & Baillargeon, 2005), infants would form an association between the agent, the object, and the object's location, and in test they would simply ex pect the same agent/object/object location association to reoccur. Thus to explain success at the task, the 3-way association account does not require children to distinguish between true- or false-belief situations, but rather to simply expect previous associations to reoccur.

Ourmodel provides 2 arguments suggesting that success at the implicit task requires more than only using simple associations. First, non-linearly separable problems (problems for which different output values cannot be separated by a single line when plotted in a two-dimensional space, or by a hyper plane in higher dimensions) cannot be solved with association tionist neural networks, which have no hidden units but only input and output units (eg, perceptrons; Minsky & Papert, 1969). A standard example of a non-linearly separable problem is exclusive or (XOR, eg, Rumelhart et al., 1986) in which there are two inputs, 0 or 1, and one output, which is 1 if exactly one of the inputs is 1, and 0 otherwise—and in fact the two-location version of the false-belief task can be con sidered isomorphic to the XOR problem.[8] Further, the fact

[8] With the two inputs being end location (0 for green, 1 for yellow) and agent watching (0 for watching the object moves to the end location, 1 for not watching), and the output being the correct search location (0 for search in green, 1 for search in yellow), the output will be 1 if exactly one of the inputs is 1 and 0 otherwise.

that networks required hidden units to succeed in our implementation of the task suggests that the false-belief task is not a linearly-separable problem and cannot be solved by simple, linear associations.

Second, the finding that network success was consistently supported by a categorization of training patterns based on belief suggests knowledge about belief situations over-and-above individual 3-way associations. Although at some level, producing different behaviors in different situ actions (eg, true- and false-belief situations) suggests being able to distinguish between situations, networks were not required to categorize search situations systematically (eg, networks might have memorized individual input– output patterns, which would be more similar to expecting a given search situation to reoccur). These results suggest that a distinction between true- and false-belief situations is required to succeed at the task, unlike simple association explanations. In short, because networks required hidden units to categorize true- and false-belief situations and to succeed at the task, our model suggests that succeeding at non-verbal false-belief tasks requires more than mere associations.

The model also has implications for our understanding of the two transitions observed with preschoolers. For the omniscient-to-representational transition, one view is that children genuinely construct new theories about other ers' behavior as they gather experience about the world (Gopnik, 1996; Hedger & Fabricius, 2011), and that they would thus initially understand others ' actions only in terms of simple desires but would later also develop an understanding of beliefs (Wellman & Cross, 2001; Wellman et al., 2001; Wellman & Woolley, 1990). Also arguing for a change in the understanding of mental states, it has been suggested that children develop either the understanding of how mental events relate to each other and to real-world objects and desires (Perner, Mauer, & Hildenbrand, 2011; Perner et al., 2007), or develop the understanding that mental events are causally related to behavior (Flavell et al., 1995; Sobel et al., 2010). Another view is that the source of the transition is not development in the understanding of beliefs, but changes in auxiliary skills such as: executive function (eg, Bull et al., 2008; Carlson et al., 2004; Carpenter et al., 2002; Flynn, 2007; Flynn et al., 2004; Frye, Zelazo, Brooks, & Samuels, 1996; Frye et al., 1998; Hughes, 1998; Leslie et al., 2004; Pellicano, 2007; Russell, 2007; Sabbagh, Xu, et al., 2006), understanding and using representations (eg, Riggs et al., 1998; Roth & Leslie, 1998; Zaitchik, 1990), working memory (eg, Gordon & Olson, 1998), or language (eg, de Villiers, 2007; Lohmann & Tomasello, 2003).

Our model is more in line with this latter view. First, while each network categorized training patterns on the basis of belief (true, false), the ability to make this distinction is separate from an understanding that others have beliefs. Children may understand that others have beliefs—for instance true beliefs—without distinguishing between true and false beliefs, and without succeeding at false-belief tasks. Children may initially have a bias to attribute true beliefs to others because beliefs tend to represent true states of affairs (Bloom & German, 2000; Fodor, 1992; Leslie et al., 2005; Sabbagh, Moses, & Shiverick, 2006) or, as in the model, due to observing more true-than false-belief situations.

Second, networks went through Transition 1 by recruiting generic computational power (ie, unspecialized hid den units), suggesting that an increase in children's general processing capacity may contribute to success at false belief tasks. In children this might represent the development of executive functioning (eg, Bull et al., 2008; Carlson et al., 2004; Carpenter et al., 2002; Flynn, 2007; Flynn et al., 2004; Frye et al., 1998; Hughes, 1998; Leslie et al., 2004; Pellicano, 2007; Russell, 2007; Sabbagh, Xu, et al., 2006) or long-term memory (eg, Jones, Gobet, & Pine, 2008). Our results can thus be considered in line with Bloom and German (2000), in that the false-belief task may not be the best task to evaluate the understanding that others have beliefs. Indeed, tasks with a structure similar to false-belief tasks, such as the false-photograph task, may be problematic for children, regardless of the belief component (Bloom & German, 2000; Roth & Leslie, 1998).
In the false-photograph task (eg, Zaitchik, 1990), a photo graph is first taken of an object in location A. The object is then moved to location B, and children are asked to say where the object is in the photograph. Preschoolers go through similar transitions (from saying "location B" to saying "location A") on both false-belief and false-photo graph tasks, though only the former depends on under standing beliefs (Bloom & German, 2000 ; Davis & Pratt, 1995; Leekam & Perner, 1991; Leslie & Thaiss, 1992; Zaitchik, 1990; but also see Sabbagh, Moses, et al., 2006; Slaughter, 1998). In contrast, autistic children of similar mental age usually fail at the belief task while succeeding at the photograph task. These results are consistent with autistic children having specific problems with beliefs (although see Klin, 2000; Sabbagh, Moses, et al., 2006), and with normally-developing children having specific problems with the structure of the tasks (eg, Bloom & German, 2000)—either with processing something that is false, or being unable to represent two locations (current, previous) for an object simultaneously (Riggs et al., 1998; Roth & Leslie, 1998; Zaitchik, 1990).

Our view is further supported by the mounting evi dence that infants understand something about, and are influenced by, the beliefs and knowledge of others (Baillargeon, Scott, & He, 2010; Luo, 2011; Luo & Baillargeon, 2010; Moll, Carpenter , & Tomasello, 2007; Moll, Koring, Carpenter, & Tomasello, 2006; Moll, Richter, Carpenter, & Tomasello, 2008; Moll & Tomasello, 2007; Onishi & Baillargeon, 2005; Poulin-Dubois, Sodian, Metz, Tilden, & Schoeppner, 2007; Scott & Baillargeon, 2009; Scott, Baillargeon, Song, & Leslie, 2010; Sodian, Thoermer, & Metz, 2007; Song, Onishi, Baillargeon, & Fisher, 2008; Southgate, Chevallier, & Csibra, 2010 ; Southgate, Senju, & Csibra, 2007; Surian, Caldi, & Sperber, 2007; Tomasello & Haberl, 2003; Träuble, Marinovic ´, & Pauen, 2010), perhaps as early as by 7 months (Kovács, Téglás, & Endress , 2010), and that this knowledge can even influence an infant's overt helping behavior (Buttelmann, Carpenter, & Tomasello, 2009).

Thus, our results suggest that false-belief tasks cannot be solved by mere associations and that the omniscient to-representational ToM transition may arise from over coming a default true-belief attribution by categorization

true- and false-belief situations. For the approach-to avoidance transition, the model suggests that it is due to avoidance search being less consistent than approach search, a novel explanation that is different but not contradictory to the claim that avoidance tasks require additional inhibition skills (Friedman & Leslie, 2004a, 2004b, 2005; Leslie et al., 2004, 2005). Analysis of the internal structure of the networks showed categorization of the training pat terns first by task then by belief. Our model thus suggests that in order to predict where others will search in false belief tasks, determining whether they want to approach or to avoid the object has the most predictive power, and then determining whether they have a true or a false belief about the object's location would come next, possibly be cause, as a first-pass estimation, it is more reasonable to assume true beliefs than to assume desires to approach. Future empirical studies may explore the role of search behavior consistency in false-belief task transitions, and the potential use of heuristics identifying others' desires first, and others' beliefs next.

In short, our model of false-belief tasks is the first computational model to autonomously construct and transition between structures and to cover the two major false-belief task transitions. The model suggests that observing more true-than false-belief behavior produces an initial bias to attribute true beliefs and that categorizing true- and false-belief situations supports a transition to representational ToM predictions. It also suggests that the relative consistency of approach compared to avoidance search behavior contributes to the pattern of initial success only at approach tasks and later success at both ap proach and avoidance tasks. While models, like theories, are inevitably prone to simplifications, computational models are rigorous, flexible and powerful tools for studying potential computational factors in development. The enduring debate about the mechanisms underlying devel opmental transitions on false-belief tasks can only benefit from multi-disciplinary approaches exploring the emergence of psychological phenomena.

Acknowledgments

We are grateful to Yoshio Takane and Vanessa Evans for helpful comments on earlier drafts of this paper, and to Debra Titone for insightful questioning that led to the simulations presented in Section 3.2.7.1 Variability of avoidance search. This research was supported by a scholarship to VGB from the Fonds Québécois de la Recherche sur la Nature et les Technologies, a grant to KHO from the Fonds Québécois de la Recherche sur la Société et la Culture, as well as a scholarship to VGB and grants to TRS and KHO from the Natural Sciences and Engineering Research Council of Canada.

Appendix A. Results from backpropagation networks

The networks using the backpropagation algorithm had the same 9 input and 4 output nodes as the Sibling Descendant Cascade Correlation (SDCC) networks, and each had one fully interconnected hidden layer containing a number of hidden units equal to the square root of the total number of training patterns (following Dandurand, Grainger, & Dufau, 2010). Training included either the full training sets of Experiments 1 (1008 patterns, 32 hidden units) or 2 (2016 patterns, 45 hidden units) and lasted 1500 epochs (to ensure backpropagation networks had substantial time to learn, we trained them for approximately twice the number of epochs that SDCC networks required in Experiments 1 and 2, which was respectively M = 744.64, SD = 76.93, and M = 930.51, SD = 128.49 epochs). Testing occurred every 50 epochs. For each Experiment, we trained and tested 56 (Experiment 1) or 112 (Experiment 2) networks in each of 6 versions of the mod el, using all possible combinations of 3 values for the learning rate parameter (.25, .50 , .75) and two values for the momentum parameter (.45, .90). For each network in each version in both Experiments, error stagnated within the first 100 epochs, when predictions did not match either omniscient or representational ToM predictions.

Because our training was stochastic, which can be problematic for deterministic neural networks (such as backpropagation and SDCC), we verified that our back propagation networks could at least learn from some sorts of stochastic training. We trained 112 backpropagation networks with the same structure as above on a simplified training set based on Experiment 2. The training set contained 10 patterns, obtained by combining one input pattern (agent watching avoidance object moves from green to yellow) with 9 correct target searches (three in each red, blue and green), and 1 incorrect target (search in yellow). Networks learned to predict search in the three empty locations and not in the object location.

In sum, our backpropagation networks could learn from simple stochastic training, but failed to produce either omniscient or representational ToM predictions from the training sets for either Experiment 1 or 2, and they did not produce any transitions.

Appendix B. Contribution analyzes for Experiment 1

To analyze the internal structure of our model, we conduct ducted contribution analyzes on 5 of the 56 networks from Experiment 1 (following Shultz & Elman, 1994; Shultz et al., 1995). For each network, we constructed a contribution matrix which contained rows for each training pattern, columns for each output connection weight, and output contributions in the cells. The variance–covariance matrix of the contribution matrix was subjected to a Principal Components Analysis (PCA), using 1.0 as the mini mum eigenvalue for retention and varimax rotation to improve interpretability of the solutions. Only components with eigenvalues greater than the mean eigenvalue were kept, and in cases where components were eliminated, the analyzes were repeated with only the number of retained components. The PCA produced component scores for each training pattern, indicating how each component weighted or categorized each pattern. To better visualize the relationship between the components and the individual training patterns, we plotted each training pattern as a function of its component scores for the retained

components. The PCA also produced loadings indicating how each connection weight contributed to each component. Because all 5 networks showed similar patterns, we report detailed results for 1 network at two test points, ie, when it produced omniscient predictions (0 hidden units) and when it produced representational ToM predictions (1 hidden unit).

B.1. Contribution analyzes of a network with 0 hidden units

At 0 hidden units, we retained three components, which overall accounted for 73.7% of the contribution variance. Fig. 7 shows component scores for each of the 1008 train ing patterns plotted along Components 1 and 2. There ap pear to be only 4 points because the patterns share the same value along each Component.

Component 1 (x-axis), which explained 29.5% of the variance, grouped training patterns based on whether the object ended in the yellow (circles at 1.6) or green (dia monds at 1.2) locations, with intermediate values for end-blue (squares) and end-red (triangles). Examination of the rotated component matrix revealed that the contri butions from the yellow and green end-location units loaded most heavily on Component 1. Thus Component 1 successfully distinguished end-yellow from end-green pat terns, while tending to lump together end-blue and end red patterns.

Component 2 (y-axis), which explained 25.1% of the variance, separated end-green patterns (diamonds at .8) from end-blue patterns (squares at 1.8) with end-red and end-yellow in between (at approximately .5). On the rotated component matrix, contributions from green and blue end-location units loaded most heavily on Component 2.

Component 3 (not plotted) explained 19.1% of the variance. Component 3 separated all end locations, and contri butions from the end-location units loaded on it most heavily.

The four other analyzed networks similarly presented components that grouped patterns according to their end locations, although the specific locations, eg, end in red versus end in yellow, etc., varied across networks. Overall, contribution analyzes of networks with omniscient expectations showed that networks were distinguishing patterns based on their end locations (in line with their omniscient predictions for both true- and false-belief conditions), but did not discriminate between true- and false-belief pat terns. Further, the agent unit was not loading on any of the retained components, suggesting that at this point net works were not relying on it to make search predictions.

B.2. Contribution analyzes of a network with 1 hidden unit

At 1 hidden unit, we retained three components, which overall accounted for 78.4% of the contribution variance. Fig. 8 shows component scores for each of the training patterns for Components 1 and 2.

Component 1, which explained 41.9% of the variance, grouped training patterns based on whether the patterns represented false-belief (negative) or true-belief situations (positive). In the rotated component matrix, contributions from the agent and the hidden unit loaded most heavily on Component 1, indicating that the network used a combination of the agent and hidden units to distinguish between true- and false-belief patterns.

Components 2 (y-axis) and 3 (not plotted), respectively explained 22.2% and 14.3% of the variance, and both grouped patterns according to the end location of the

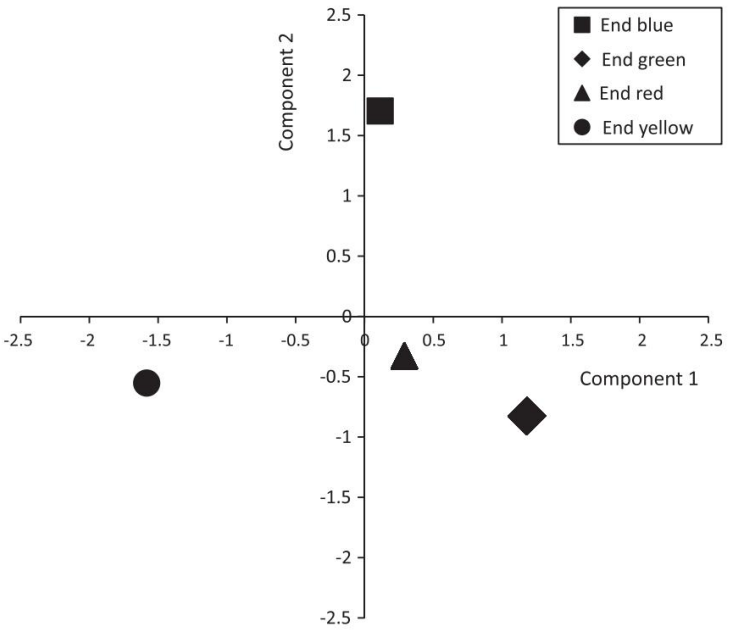

Fig. 7. Grouping of the 1008 training patterns by Components 1 and 2 of a single network from Experiment 1, which was producing omniscient predictions with 0 hidden units.

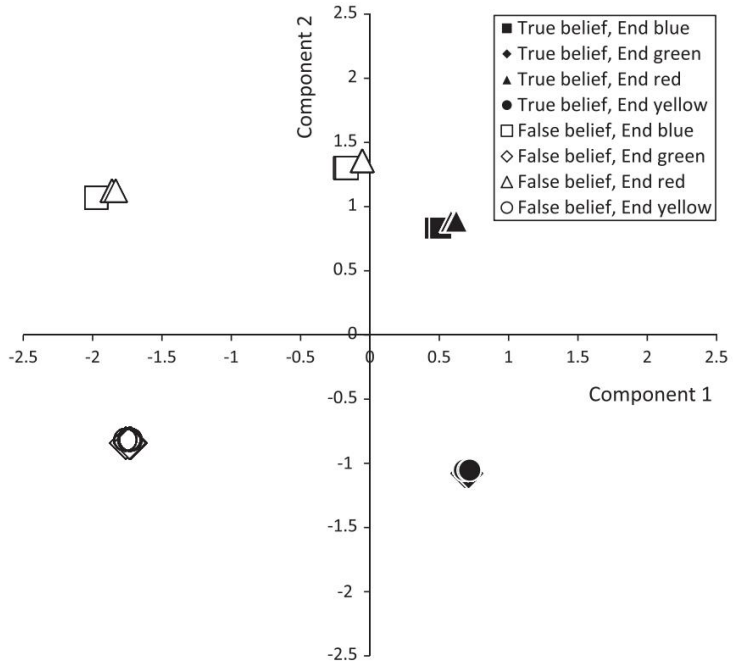

Fig. 8. Grouping of the 1008 training patterns by Components 1 and 2 of a single network from Experiment 1, which was producing representational ToM predictions with 1 hidden unit.

object. Component 2 separated end-red and end-blue (po sitive) from end-yellow and end-green (negative) patterns. Component 3 separated end-red from end-blue, with end yellow and end-green patterns in between. Reflecting this partition of scores, Component 2 had strong loadings from all end locations, and Component 3 has strongest loadings from end-red and end-blue connection weights.

In the four other analyzed networks, Component 1 sep arated true- from false-belief patterns, while Components 2 and 3 distinguished end locations, although the specific end locations varied across networks. Overall, contribution analyzes of networks with representational predictions showed that the networks used a combination of the agent and hidden units to separate true- from false-belief training patterns, while keeping track of end locations.

In sum, PCA of contributions for Experiment 1 showed that networks (1) initially used end locations to distinguish between patterns and produce omniscient predictions, and (2) learned to use agent and hidden units to distinguish be tween true- and false-belief patterns and produce representational ToM predictions.

Appendix C. Contribution analyzes for Experiment 2

Due to excessive variation in the output contributions, the standard PCA of output contributions was difficult to interpret, thus analyzes of Transition 2 differed in two ways from those of Transition 1. First, we averaged contri butions across output connection weights, based on their function in the network. This yielded contributions for 6 mean weight variables produced by collapsing contributions over the 36 initial weight columns (9 inputs by 4 outputs). The matching-start weight variable was obtained by collapsing across the 4 connection weights between matching input start and output locations (ie, collapsing over the weights connecting red-start to red-output, green-start to green-output, etc.). The matching-end vari able was similarly obtained by collapsing over matching input end and output locations. In contrast, non-match ing-start and non-matching-end variables were obtained by collapsing over weights connecting non-matching start or non-matching end input locations to outputs (ie, col lapsing over the weights connecting red-start to green output, to blue-output, to yellow-output, green-start to red-output, to blue-output, to yellow-output, etc.). The output weights of the agent unit were kept as they were.

Finally, for networks that had recruited hidden units, all contributions associated with weights from hidden units were averaged into a single hidden-unit variable.

Second, PCA was carried out using correlation matrices instead of variance–covariance matrices. Using the correlation matrix had the effect of standardizing contribution variables so that each had a mean of 0 and standard devi ation of 1, thus reducing variation due to connection weight size (Shultz et al., 1995). While there is a debate about whether to use correlation or covariance matrices when performing PCA (Shultz et al., 1995), reducing variation in this specific case uniquely enabled interpretable PCA results.

Contributions from 5 of the 112 networks were subjected to PCA, varimax rotation was applied (when more than one component was extracted), and only components with eigenvalues higher than 1 were kept. Training pat terns were plotted along the retained components, and loadings from the rotated component matrix were

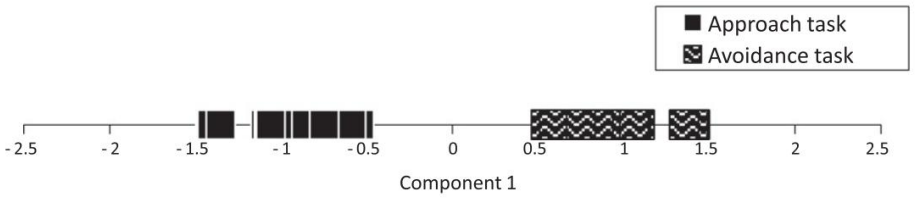

Fig. 9. Grouping of the 2016 training patterns by Component 1 of a single network from Experiment 2, which was producing omniscient predictions with 0 hidden units.

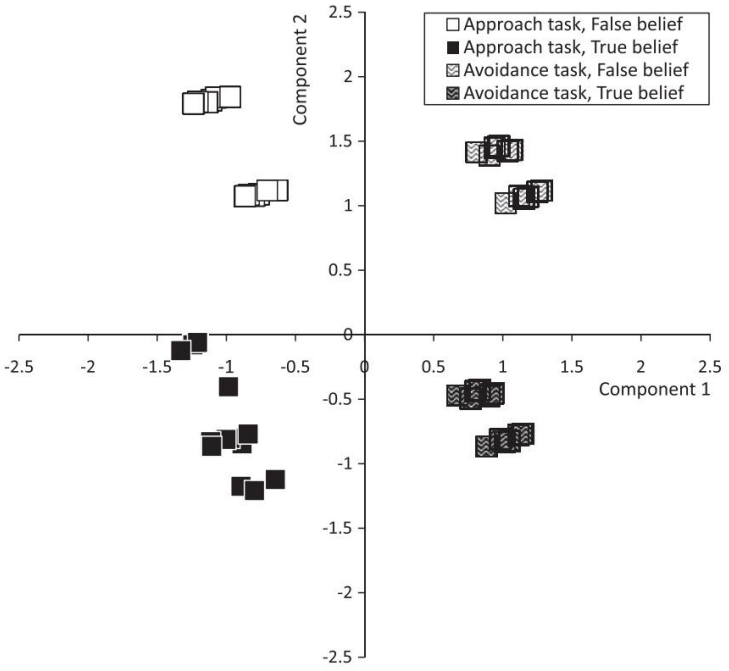

Fig. 10. Grouping of the 2016 training patterns by Components 1 and 2 of a single network from Experiment 2, which was producing representational ToM predictions only for the approach task with 3 hidden units.

analyzed. As all 5 networks showed similar patterns, we report in detail only the results for 1 network at three test points, ie, when it produced omniscient predictions (0 hidden units), when it succeeded at the approach task only (3 hidden units), and when it produced representational ToM predictions for both tasks (6 hidden units).

C.1. Contribution analyzes of a network with 0 hidden units

At 0 hidden units, we retained one component, which overall accounted for 75.6% of the contribution variance. Fig. 9 shows each of the 2016 training patterns plotted along Component 1.

Component 1 separated approach (negative) and avoidance (positive) training patterns. In the component matrix, contributions from all location variables (matching-start, matching-end, non-matching-start, non-matching-end) loaded heavily on Component 1, indicating that the net work used all locations to distinguish between approach and avoidance patterns. Use of the location variables to separate the tasks makes sense, as task was encoded at all locations with positive (approach) or negative (avoid ance) values.

The four other analyzed networks similarly yielded a single component that grouped patterns by task, though whether approach and avoidance patterns were assigned positive or negative scores varied across networks. Thus, contribution analyzes of networks with omniscient expectations show that networks distinguished between ap proach and avoidance desires, but did not distinguish between true- and false-belief patterns.

C.2. Contribution analyzes of a network with 3 hidden units

At 3 hidden units, we retained two components, which overall accounted for 89.5% of the contribution variance. Fig. 10 shows the training patterns plotted along Components 1 and 2.

As at the previous test point, Component 1 (x-axis), which explained 71.6% of the variance, divided patterns between approach (negative) and avoidance (positive) with contributions from the four location variables loading

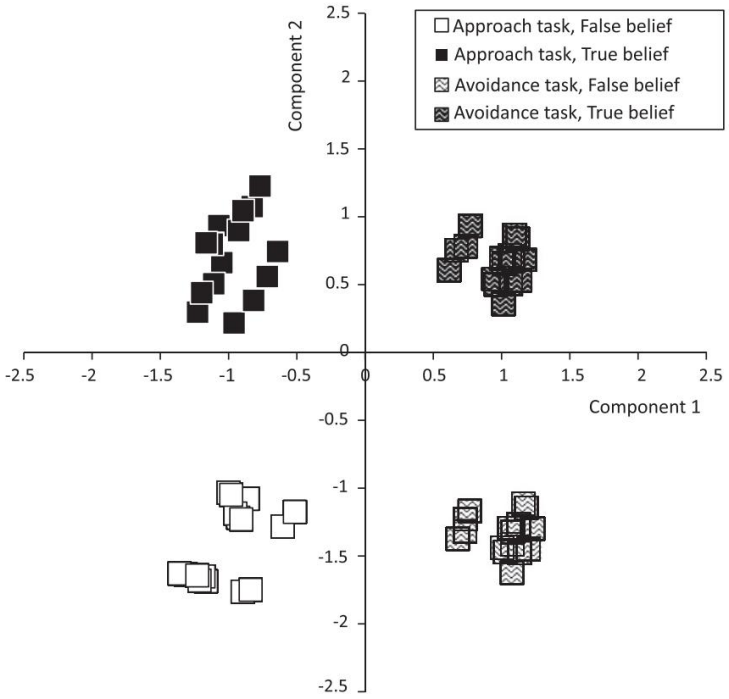

Fig. 11. Grouping of the 2016 training patterns by Components 1 and 2 of a single network from Experiment 2, which was producing representational ToM predictions in both tasks with 6 hidden units.

most heavily on this component. Component 2 (y-axis), which explained 17.9% of the variance, grouped true- (neg active) and false-belief (positive) patterns and the agent and hidden unit variables loaded most heavily on it.

In the four other analyzed networks, Components 1 and 2 divided the patterns similarly, showing that when making representational ToM predictions for approach but not the avoidance task, networks distinguished between approach and avoidance and learned to separate true and false-belief patterns using the agent and hidden units.

C.3. Contribution analyzes of a network with 6 hidden units

At 6 hidden units, we retained two components, which overall accounted for 87.5% of the contribution variance. Fig. 11 shows the training patterns plotted along Components 1 and 2.

Once again, Component 1 (x-axis), which explained 70.1% of the variance, divided approach (negative) and avoidance (positive) patterns and had the heaviest weighing tings from the location variables. Component 2 (y-axis) explained 17.4% of the variance, and separated true- (positive) from false-belief (negative) patterns for both approach and avoidance tasks while having the heaviest weightings from the agent and hidden units.

The four other analyzed networks yielded similar with ponents (although the sign of scores assigned to avoid ance/approach and true-belief/false-belief patterns varied). PCA for networks with 6 hidden units thus did not differ from PCA for networks with 3 hidden units: networks distinguished between approach and avoidance patterns (using locations) and between true- and false-be lief patterns (using the agent and hidden units).

In sum, PCA of contributions for Experiment 2 showed that networks with 0 hidden units (1) initially used locations to separate approach from avoidance task patterns to produce omniscient predictions in both tasks, and that once networks were succeeding at the approach task with 3 hidden units, and when they were succeeding at both ap proach and avoidance tasks with 6 hidden units, they (2) used the agent and hidden units to distinguish between true- and false-belief patterns. However, because PCA did not differ between networks succeeding only at the ap proach task and those succeeding at both the approach and avoidance tasks, this technique did not reveal the mechanisms underlying the approach to avoidance transition.

References

Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, & Qin Y. (2004). An integrated theory of the mind. Psychological Review, 111(4), 1036–1060.

Baillargeon, R., Scott, RM, & He, Z. (2010). False-belief understanding in infants. Trends in Cognitive Sciences, 14(3), 110–118.

Baluja, S., & Fahlman, SE (1994). Reducing network depth in the cascade correlation (No. CMU-CS-94-209). Pittsburgh: School of Computer Science, Carnegie Mellon University.

Baron-Cohen, S., Leslie, AM, & Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition, 21(1), 37–46.

Bloom, P., & German, TP (2000). Two reasons to abandon the false belief task as a test of theory of mind. Cognition, 77(1), B25–B31.

Bull, R., Phillips, LH, & Conway, CA (2008). The role of control functions in mentalizing: Dual-task studies of theory of mind and executive function. Cognition, 107(2), 663–672.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month old infants show false belief understanding in an active helping paradigm. Cognition, 112(2), 337–342.
Carlson, SM, Mandell, DJ, & Williams, L. (2004). Executive function and theory of mind: Stability and prediction from ages 2 to 3. Developmental Psychology, 40(6), 1105–1122.
Carpenter, M., Call, J., & Tomasello, M. (2002). A new false belief test for 36-month-olds. British Journal of Developmental Psychology, 20(3), 393–420.

Cassidy, KW (1998). Three- and four-year-old children's ability to use desire- and belief-based reasoning. Cognition, 66(1), B1–B11.
Cattell, RB (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245–276.
Dandurand, F., Grainger, J., & Dufau, S. (2010). Learning location-invariant orthographic representations for printed words. Connection Science, 22(1), 25–42.
Davis, HL, & Pratt, C. (1995). The development of children's theory of mind: The working memory explanation. Australian Journal of Psychology, 47(1), 25–31.

de Villiers, JG (2007). The interface of language and theory of mind. Language, 117(11), 1858–1878.
Elman, JL, Bates, EA, Johnson, MH, Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). Rethinking Innateness. Cambridge, MA: MIT press.
Flavell JH, Green FL, Flavell ER, Harris PL, & Astington JW (1995). Young children's knowledge about thinking. Monographs of the Society for Research in Child Development, 60(1), 1–113.
Flynn, E. (2007). The role of inhibitory control in false belief understanding. Infant and Child Development, 16(1), 53–69.
Flynn, E., O'Malley, C., & Wood, D. (2004). A longitudinal, microgenetic study of the emergence of false belief understanding and inhibition skills. Developmental Science, 7(1), 103–115.
Fodor, JA (1992). A theory of the child's theory of mind. Cognition, 44(3), 283–296.
Friedman, O., & Leslie, AM (2004a). Mechanisms of belief-desire reasoning: Inhibition and bias. Psychological Science, 15(8), 547–552.
Friedman, O., & Leslie, AM (2004b). A developmental shift in processes underlying successful belief-desire reasoning. Cognitive Science, 28(6), 963–977.

Friedman, O., & Leslie, AM (2005). Processing demands in belief-desire reasoning: Inhibition or general difficulty? Developmental Science, 8(3), 218–225.

Frye, D., Zelazo, PD, Brooks, PJ, & Samuels, MC (1996). Inference and action in early causal reasoning. Developmental Psychology, 32(1), 120–131.

Frye, D., Zelazo, PD, & Burack, JA (1998). Cognitive complexity and control. Current Directions in Psychological Science, 7(4), 116–121.
Goodman ND, Baker CL, Bonawitz EB, Mansinghka VK, Gopnik A, Wellman HM, et al. (2006). Intuitive theories of mind: A rational approach to false belief. Proceedings of the 28th annual conference of the cognitive science society (pp. 1382–1387). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gopnik, A. (1996). The scientist as child. Philosophy of Science, 63(4), 485–514.
Gordon, ACL, & Olson, DR (1998). The relation between acquisition of a theory of mind and the capacity to hold in mind. Journal of Experimental Child Psychology, 68(1), 70–83.
Harvey, I., Paolo, ED, Wood, R., Quinn, M., & Tuci, E. (2005). Evolutionary robotics: A new scientific tool for studying cognition. Artificial Life, 11(1–2), 79–98.

Hedger, JA, & Fabricius, WV (2011). True belief belies false belief: Recent findings of competence in infants and limitations in 5-year olds, and implications for theory of mind development. Review of Philosophy and Psychology, 2, 429–447.
Hughes, C. (1998). Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind? Developmental Psychology, 34(6), 1326–1339.
Jones, G., Gobet, F., & Pine, J. (2008). Computer simulations of developmental change: The contributions of working memory capacity and long-term knowledge. Cognitive Science, 32(7), 1148–1176.

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The social attribution task. The Journal of Child Psychology and Psychiatry, 41(7), 831–846.

Kovacs, A. M., Téglás, E., & Endress, AD (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. Science, 330(6012), 1830–1834.

Leekam, SR, & Perner, J. (1991). Does the autistic child have a metarepresentational deficit? Cognition, 40(3), 203–218.

Leslie, AM, Friedman, O., & German, TP (2004). Core mechanisms in "theory of mind". Trends in Cognitive Sciences, 8(12), 529–533.
Leslie, AM, German, TP, & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. Cognitive Psychology, 50(1), 45–85.
Leslie, AM, & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. Developmental Science, 1(2), 247-253.
Leslie, AM, & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. Cognition, 43(3), 225–251.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. Child Development, 74(4), 1130–1144.
Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? Cognition, 121(3), 289–298.
Luo, Y., & Baillargeon, R. (2010). Toward a mentalistic account of early psychological reasoning. Current Directions in Psychological Science, 19(5), 301–307.
Minsky, M., & Papert, S. (1969). perceptrons. Cambridge, MA: MIT Press.
Moll, H., Carpenter, M., & Tomasello, M. (2007). Fourteen-month-olds know what others experience only in joint engagement. Developmental Science, 10(6), 826-835.
Moll, H., Koring, C., Carpenter, M., & Tomasello, M. (2006). Infants determine others' focus of attention by pragmatics and exclusion. Journal of Cognition and Development, 7(3), 411–430.
Moll, H., Richter, N., Carpenter, M., & Tomasello, M. (2008). Fourteen month-olds know what "we" have shared in a special way. Infancy, 13(1), 90–101.

Moll, H., & Tomasello, M. (2007). How 14- and 18-month-olds know what others have experienced. Developmental Psychology, 43(2), 309–317.
O'Loughlin, C., & Thagard, P. (2000). Autism and coherence. A computational model. Mind & Language, 15(4), 375–392.
Onishi, KH, & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? Science, 308(5719), 255-258.
Pellicano, E. (2007). Links between theory of mind and executive function in young children with autism: Clues to developmental primacy. Developmental Psychology, 43(4), 974–990.
Perner, J., Mauer, MC, & Hildenbrand, M. (2011). Identity: Key to children's understanding of belief. Science, 333(6041), 474-477.

Perner, J., Rendl, B., & Garnham, A. (2007). Objects of desire, thought, and reality: Problems of anchoring discourse referents in development. Mind & Language, 22(5), 475–513.
Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? Science, 308(5719), 214-216.
Poulin-Dubois, D., Sodian, B., Metz, U., Tilden, J., & Schoeppner, B. (2007). Out of sight is not out of mind: Developmental changes in infants' understanding of visual perception during the second year. Journal of Cognition and Development, 8(4), 401–425.
Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind. Behavioral and Brain Sciences, 1(4), 515–526.
Riggs, KJ, Peterson, DM, Robinson, EJ, & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? Cognitive Development, 13(1), 73–90.
Roth, D., & Leslie, AM (1998). Solving belief problems: Toward a task analysis. Cognition, 66(1), 1–31.
Rumelhart, DE, McClelland, JL, & PDP Research Group (1986). Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press.
Russell, J. (2007). Controlling core knowledge: Conditions for the ascription of intentional states to self and others by children. Synthesis, 159(2), 167–196.
Sabbagh, MA, Moses, LJ, & Shiverick, S. (2006a). Executive functioning and preschoolers' understanding of false beliefs, false photographs, and false signs. Child Development, 77(4), 1034–1049.
Sabbagh MA, Xu F., Carlson SM, Moses LJ, & Lee K. (2006b). The development of executive functioning and theory of mind. Psychological Science, 17(1), 74–81.
Scott, RM, & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. Child Development, 80(4), 1172–1196.

Scott RM, Baillargeon R, Song H, & Leslie AM (2010). Attributing false beliefs about non-obvious properties at 18 months. Cognitive Psychology, 61(4), 366–395.
Shultz, TR (2003). Computational developmental psychology. Cambridge, MA: MIT Press.
Shultz, TR (2006). Constructive learning in the modeling of psychological development. Processes of change in brain and cognitive development: Attention and performance XXI (pp. 61–86). Oxford: Oxford University Press.

Shultz, TR, & Cohen, LB (2004). Modeling age differences in infant category learning. Infancy, 5(2), 153–171.

Shultz, TR, & Elman, JL (1994). Analyzing cross connected networks. Advances in neural information processing systems (Vol. 6, pp. 1117–1124). San Francisco, CA: Morgan Kaufmann.

Shultz, TR, Mareschal, D., & Schmidt, WC (1994). Modeling cognitive development on balance scale phenomena. Machine Learning, 16, 57–86.

Shultz, TR, Oshima-Takane, Y., & Takane, Y. (1995). Analysis of unstandardized contributions in cross connected networks. Advances in neural information processing systems (Vol. 7, pp. 601–608). Cambridge, MA: MIT Press.

Siegal, M., & Beattie, K. (1991). Where to look first for children's knowledge of false beliefs. Cognition, 38(1), 1–12.

Slaughter, V. (1998). Children's understanding of pictorial and mental representations. Child Development, 69(2), 321–332.

Sobel DM, Buchanan DW, Butterfield J, & Jenkins OC (2010). Interactions between causal models, theories, and social cognitive development. Neural Networks, 23(8–9), 1060–1071.

Sodian, B., Thoermer, C., & Metz, U. (2007). Now I see it but you don't: 14-month-olds can represent another person's visual perspective. Developmental Science, 10(2), 199–204.

Song, H., Onishi, KH, Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. Cognition, 109(3), 295–315.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. Developmental Science, 13(6), 907-912.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. Psychological Science, 18(7), 587–592.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. Psychological Science, 18(7), 580–586.

Surian, L., & Leslie, AM (1999). Competence and performance in false belief understanding: A comparison of autistic and normal 3-year-old children. British Journal of Developmental Psychology, 17(1), 141–155.

Thomas, MSC, & Karmiloff-Smith, A. (2003). Connectionist models of development, developmental disorders and individual differences. In R. Sternberg, J. Lautrey, & T. Lubart (Eds.), Models of intelligence. International perspectives (pp. 133–150). Washington DC: American Psychological Association.

Tomasello, M., & Haberl, K. (2003). Understanding attention: 12- and 18-month-olds know what is new for other persons. Developmental Psychology, 39(5), 906–912.

Träuble, B., Marinovic´, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs? Infancy, 15(4), 434–444.

Triona LM, Masnick AM, & Morris BJ (2002). What does it take to pass the false belief task? An ACT-R model. Proceedings of the 24th annual conference of the cognitive science society (p. 1045). Mahwah, NJ: Lawrence Erlbaum Associates, Inc., p. 1045.

Van Overwalle, F. (2010). Infants' teleological and belief inference. A recurrent connectionist approach to their minimal representational and computational requirements. NeuroImage, 52(3), 1095–1108.

Wahl, S., & Spada, H. (2000). Children's reasoning about intentions, beliefs and behavior. Cognitive Science Quarterly, 1(1), 5–34.

Wellman, HM, & Cross, D. (2001). Theory of mind and conceptual change. Child Development, 72(3), 702–707.

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

Wellman, HM, & Woolley, JD (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. Cognition, 35(3), 245–275.

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. Cognition, 35(1), 41–68.