Cognitive Architecture of Belief Reasoning in Children and Adults: A Two-Systems Account Primer

Jason Low, Ian Apperly, Hannes Rakoczy and Stephen Butterfill

Forthcoming in Child Development Perspective

ABSTRACT—Characterizing the cognitive architecture of human mindreading forces us to address two puzzles in people's attributions of belief: why children show inconsistent expectations about others' belief-based actions, and why adults' belief reasoning is sometimes automatic and sometimes not. The seemingly puzzling data suggest humans have multiple mindreading systems that use different models of the mental. The efficient system is shared by infants, children and adults, and uses a minimal model of mind, which enables belief-like states to be tracked. The flexible system is late-developing and uses a canonical model, which incorporates propositional attitudes. A given model's operation has signature limits that produce performance contrasts, in children as well as adults, between certain types of mindreading tasks.

Our everyday mindreading ability helps us reason about how beliefs might influence people's actions, inter-personal communications and conduct. There are difficult puzzles surrounding the nature of human beings' belief attribution: (I) children show an apparently contradictory pattern of success and failure in their responses to scenarios involving others' belief-based actions, and (II) belief reasoning is both non-automatic and automatic. To solve these puzzles, we highlight evidence from cognitive studies of children and adults to survey an exciting approach to the architecture of mindreading suggesting that human beings can be in (at least) two minds about the ways in which others' beliefs cause and rationalize behavior (1, 2). We discuss how 'signature limits' on low-level processes make it possible to differentiate between efficient versus flexible instances of mindreading. We then evaluate a contrasting account suggesting that human beings have a unitary and abstract psychological reasoning system from early in life.

PUZZLES IN PEOPLE'S ATTRIBUTION OF BELIEF

Puzzle I: Infants pass false-belief tasks but 3-year-olds fail?

A measure of the development of our mindreading ability is the false-belief task. Wimmer and Perner (3) showed preschoolers a story where Maxi witnesses a target placed at location-X. In Maxi's absence, the target is moved to location-Y. Children are asked to predict where Maxi would look for the target. Most 3-year-olds answered Maxi would look in Y, as if false-belief were impossible; by contrast, many 4-year-olds answered X, indicating they recognized Maxi's false-belief. The incorporation of belief into children's understanding of minds from about age 4-years onwards is a well-replicated and robust experimental finding (4). Once children master verbal

false-belief tasks, they do so systematically and coherently for a large variety of topics and task formats. Importantly, 4-year-olds' grasp of beliefs includes appreciating that beliefs are essentially aspectual; that is beliefs represent a given object under some guises but not others. Rakoczy, Bergfeld, Schwarz, and Fizke (5) found that when 4-year-olds pass standard false-belief tasks, they begin to understand that an agent, depending on how he or she represents something, can mistakenly believe that there are two objects present when, in fact, there is only one.

The findings from explicit verbal tasks contrast with results from non-verbal measures. Whereas 3-year-olds' verbal predictions indicate that they reason as if false-belief were impossible, their gaze anticipations to the same situation indicate that they can track others' false-beliefs (6, 7, 8). The dissociation is supported by violation-of-expectation studies contrasting looking-times to scenarios that are either consistent or inconsistent with an agent's belief. Onishi and Baillargeon (9) showed 15-month-olds scenarios of an agent forming either a true- or false-belief about an object's location. The agent searched in the belief-compatible or the belief-incompatible location. Infants looked longer when the agent searched in the belief-incompatible location. Longer looking is interpreted as infants expecting agents to act according to their beliefs. Other studies suggest that 7- to 18-month-olds can track false-beliefs about contents and types of objects, and tailor their helping and communication to others' false-belief about object-location (10). The first puzzle is thus: How can infants and toddlers display sensitivity to others' false-beliefs when responding in some ways while they treat false-belief as impossible when responding to the very same situation in other ways?

Puzzle II: Belief reasoning is both non-automatic and automatic

Studies of adult humans also point to seemingly incompatible sets of findings regarding the automaticity of mindreading inferences. A mindreading process is automatic if its occurrence is to a significant degree independent of its relevance to participants' tasks and motives. Apperly and colleagues (11) found that false-beliefs are not ascribed automatically: adults with no specific motivation to attend to a character's beliefs were slower to respond to unpredictable probe questions about an agent's false-belief of an object's whereabouts than to matched control probes. The case for non-automaticity is also supported by research showing that belief tracking frequently depends on attention and working memory resources in fully competent adults and, further, that even merely holding in mind someone else's belief incurs significant processing costs (12).

However, there is also a body of evidence pointing to a different conclusion. Schneider, Nott, and Dux (13) found that a character's false-belief can influence adults' visual attention irrespective of the relevance of the belief to the tasks adults were assigned. Both adults who were told to track a character's belief and adults who were told to track a ball's location fixated longer at an empty box before the character returned to the scene and falsely believed the box to contain the ball than when the character believed it was empty. Mirroring findings from young children, Van der Wel and colleagues (14) found that the effects of indirect belief calculation were different from the effects of direct belief judgments. Adults saw a ball and a cube disappear behind two screens. A bystander had a false-belief whilst participants had a true-belief about the objects' locations. Participants who were instructed to move a computer mouse to reach the ball's location showed involuntary tracking of belief:

their mouse movements to the ball were skewed towards where the bystander falsely-believed the ball to be. Deliberate inferences showed different effects: participants who were told to track beliefs took longer to move the mouse when their beliefs differed from the bystander's (and their mouse movements were not skewed by the bystander's beliefs). The second puzzle is thus: How can belief tracking be sometimes but not always automatic?

TWO-SYSTEMS ACCOUNT

We can solve the puzzles by supposing that mindreading architecture involves at least two systems for tracking mental states, with complementary trade-offs between efficiency and flexibility in much the sense that, on some theories, there are at least two systems for tracking number (15). The efficient mindreading system is evolutionarily and ontogenetically ancient, fast-operating, largely automatic and independent of central cognitive resources. In contrast, the flexible mindreading system is late-developing, slow-operating, making deep and lasting demands on executive control processes. Advances in executive function and language help cultivate flexible attributions about others' psychological perspectives (12). While the efficient system typically subserves responses that occur independently of a subject's task and motives (e.g., looking-behavior on some tasks), the flexible system is recruited by tasks that require declarative expressions of and/or deliberation about beliefs.

The processes that drive the efficient system may be trigged by direct cues like an agent's line of sight so that that rapid online mindreading may be supported in subjects with limited information-processing resources. Deployment of the flexible

system is not dependent upon the immediate availability of cues about what a target witnesses. Components of efficient mindreading may have non-zero cognitive costs and may place some demands on working memory, as indicated by findings suggesting that dual tasking may disrupt looking-time responses to false-belief tasks (16). The efficient system should remain relatively distinct from the more flexible system, although there might be some exchange of information between systems over the course of development (7, 12, 17).

Efficient mindreading is distinct from flexible mindreading in terms of signature limits arising from the type of model of the mental that the respective systems rely on. The flexible system uses a canonical model of the mental where belief is characterized as a propositional attitude. A propositional attitude is a state whose content can be picked out with a that-clause (e.g., Lucy believes that the Morning Star is above the horizon). A canonical model takes into account the aspectuality of beliefs, so that although the Morning Star is the Evening Star, Lucy's belief that the Morning Star is above the horizon is distinct from her belief that the Evening Star is at that location. Such flexible reasoning would support understanding of mistakes in others' representations of identity in the numerical sense, as when Lucy falsely believes the Morning Star is not the Evening Star. The efficient system, by contrast, uses a minimal model of the mental where psychological states including belief-like states are characterized as relational attitudes – states whose contents can be distinguished using relations between objects and locations or other properties.

Belief-like states can serve as proxies for beliefs: in a limited but useful range of situations, ascriptions of beliefs and belief-like states lead to identical expectations about an agent's behaviour. However, their contents are not as fine-grained as the truth conditions of beliefs proper; crucially, they are not aspectual, i.e. do not

distinguish under which guises objects and situations are represented (2). If Lucy has a belief-like relational attitude to the Morning Star and its position above the horizon, and if the Morning Star is the Evening Star, then she has the same attitude concerning the Evening Star. An efficient mindreading system will therefore display a signature limit concerning the aspectuality of belief.

Much like ascribing belief, there is also more to reasoning about perception than tracking someone's visual connection to an object; different visual experiences may represent the very same thing in different ways. An efficient mindreading system that is set to track relational attitudes will also be ill-equipped to process the aspectual nature of mental states generally. The 2-systems account therefore predicts that the efficient system can cover Level-I visual-perspective-taking tasks (tracking what is or is not perceptible from different perspectives) and simple false-belief tasks about the location of objects (the subject has to keep track of what the agent has or has not witnessed). However, this system cannot cover Level-II visual-perspective-taking (representing the particular way someone sees an object) or ascribing false-beliefs about numerical identity, giving rise to signature limits.

SIGNATURE LIMITS ON EFFICIENT MINDREADING

There are at least three sources of relevant evidence. First, visual perspective-taking studies show that humans automatically track what is seen but not how something is seen. Samson and colleagues (18) showed adults photographs where an avatar saw all of the dots on a wall (his perspective was consistent with participants') or where the avatar saw a subset of the dots (his perspective was inconsistent with participants'). Adults were slower and more error-prone in judging how many dots they could see

when the avatar happened to have a different perspective. Furlanetto and colleagues (19) confirmed that adults experienced interference from the avatar's perspective only if they believed he could see, suggesting that interference resulted from processing of the avatar's mental states, and not merely the direction in which he was facing (cf. 20). Thus, even when calculating what others see (a Level-I perspective-taking scenario) is task-irrelevant, children and adults automatically track others' encountering and registration of objects, and this causes interference on self-judgments.

Fitting with the 2-systems account, the interference in Level-I perspective-taking scenarios does not generalize to Level-II perspective-taking scenarios, which concern how an agent represents an object. Surtees and colleagues (21, 22) found that children and adults did not automatically show such interference effects when participants had to report how they represented a rotationally asymmetrical digit (e.g., a '6') that was perceived differently from the avatar's opposite viewing angle (e.g., as a '9'). Similar patterns have been found in experiments measuring adults' eye movements during real-time discourse processing. For example, Mozuraitis, Chambers and Daneman (23) found that listeners distinguished between what they know versus what a speaker is inferred to know based on whether an object was seen, but not how it was seen.

Second, Low and Watts (24) found that 3- and 4-year-olds and adults displayed accurate looking-time responses with the usual age-related improvements in verbal predictions when construing an agent's false-belief about object-location. However, the same participants showed incorrect looking when tracking how an agent's representation of identities would lead to a false-belief that there were two objects when, in fact, there was only one object (Figure 1, Column 1). The switch from a location to an identity task did not affect declarative responding; 4-year-olds

and a majority of the adults provided accurate verbal predictions. However, participants experienced the different visual aspects of the deceptive object late in the sequence. Demands associated with revising and updating inferences about the agent's representation of identities might have impaired participants' looking responses. That said, Wang, Hadi and Low (25) found that adults still showed incorrect gaze anticipations (but correct verbal predictions) when the test object revealed its dual aspect early in the sequence (Figure 1, Column 2).

Figure-1

Third, Fizke, Butterfill and Rakoczy (26) uncovered complementary findings when measuring toddlers' helping behavior. An object that was both an [A] and a [B] (e.g., reversible rabbit-carrot toy) was put into box-1 in the agent's presence as [A]. The object was then turned into its B-aspect and returned to box-1 – in the absence of the agent in the false-belief condition (so that she was unaware of the identity A=B), but witnessed by the agent in the true-belief condition. Then the agent observed the object (as [B]) moved from box-1 to box-2. The agent struggled to open box-1 and children's spontaneous helping was recorded. Children did not behave differently between the false- and true-belief conditions: the majority of toddlers focused on goal-directed relations and opened box-1. It is not that toddlers failed to understand identity per se; 14-month-olds can disregard superficial features and sort by object identity (27). Crucially, when the false-belief task was switched to pure location tracking (cf. 28), children did differentiate true- and false-belief conditions, mostly opening box-2 in the latter and box-1 in the former.

In summary, there is converging evidence showing that the efficient mindreading system breaks down in cases involving Level-II perspective-taking and beliefs about numerical identity. Because such cases require reasoning based on a

canonical model of the mental, we can use such limits to identify whether an individual's performance on a particular task involves the efficient or flexible mindreading systems (1, 2).

AN ALTERNATIVE: EARLY-MINDREADING ACCOUNT

The two-systems account contrasts with the approach suggesting that humans have a unitary early-developing (possibly innate) psychological reasoning system that parses mental states from behavior. According to the early-mindreading account, infants and young children succeed in violation-of-expectation or anticipatory-looking tasks because those tasks only involve the belief representation process (10). Additional processes are involved in tasks that typically require making verbal responses to a question; 3-year-olds also need to select between different possible responses to the test question, and inhibit a default to answer from their own knowledge. The additional processes overwhelm 3-year-olds' limited executive functioning, masking innate belief-reasoning competence.

Following the early-mindreading account, some experiments suggest that 17- to 18-month-olds can already attribute false-beliefs about identity (29, 30). However, these experiments could just as well suggest that infants are tracking beliefs about the types of objects present rather than about numerical identity (2). That said, Scott and colleagues (31) provided other evidence suggesting that infants' mindreading may be relatively sophisticated. Specifically, 17-month-olds watched a thief attempt to steal a preferred object (a rattling toy) when its owner was momentarily absent by substituting it with a less-preferred object (a non-rattling toy). Infants looked longer when the thief substituted the preferred object with a non-visually-matching silent toy

compared to when the thief substituted it with a visually-matching silent toy. The authors postulated that infants can ascribe to the thief an intention to implant in the owner a false-belief about the identity of the substituted toy. The authors further suggested that infants make such ascriptions only when the substitution involves a visually-matching toy and the owner will not test whether the toy rattles on her return.

However, Scott et al.'s (31) explanations also require postulating that infants take the thief to be strikingly inept; despite having the opportunity simply to pilfer from a closed box known to contain at least three rattling toys, the thief engages in elaborate deception which will be uncovered whenever the substituted toy is next shaken and the thief, as sole suspect, easily identified. A further difficulty is that factors unrelated to the thief's mental states vary between conditions, such as the frequencies with which toys visually matching one present during the final phase of the test trial have rattled. These considerations jointly indicate that further evidence would be needed to support the claim that humans' early mindreading capacity enables them to ascribe intentions concerning false-beliefs involving numerical identity.

In support of the early-mindreading account, Carruthers (32) suggests that performance issues can also explain findings showing non-automatic belief attribution. With respect to Apperly et al.'s (11) study, Carruthers worries about the interval between the belief cues and belief questions being longer than the interval between the reality cues and reality questions. Adults might be slower at attributing beliefs because they had to retrieve information about the agent's beliefs (which had been automatically inferred) from long-term memory when responding to unpredictable probe questions. Carruthers spotlights Cohen and German's (33) study arguing that adults automatically inferred beliefs when there was a shorter interval between belief cues and questions. However, in Cohen and German's study, the

context of the agent putting a marker on the wrong container just before the belief probe could just as well prompt adults to spontaneously (rather than automatically) infer the agent's false-belief as a relevant explanation for her mistaken endorsement. Indeed, Back and Apperly (34) found that task context motivated adults to make spontaneous inferences about an agent's beliefs, and could maintain them over time, even though they did not need to. In the absence of such motivation, however, participants did *not* automatically make belief inferences even when the stimulus afforded such inferences.

The broader developmental evidence is also not entirely consistent with the explanation that contradictions in responses to false-belief scenarios reflect completely incidental demands on executive processing. Cultural differences in inhibitory control are not linked to corresponding differences in performance on standard false-belief tasks (35). Three-year-olds do not even find selection-less false-belief tasks easier than standard false-belief tasks (36). As Wellman (36) notes, it is also unclear why certain indirect tasks (e.g., violation-of-expectation paradigm) are assumed to be free of inhibition demands when infants apparently face the same problem of controlling a default reading of the situation in terms of where the object really is located to track beliefs instead. The notion of underlying belief-reasoning competence being masked by incidental task demands to inhibit a tendency to answer from one's own knowledge would also need stretching to account for interference effects on reality judgments. Studies show that adults and children find it difficult to even hold others' false-beliefs in mind, resulting in slower and incorrect judgments about reality (37, 38). These considerations suggest that constraints on information-processing play a deeper and more nuanced developmental role in the construction, maintenance and use of belief concepts, in addition to lasting roles in the mature

mindreading system (12). The two-systems account fits better with the diverse literatures where mindreading is studied.

CONCLUSIONS

The two-systems approach to mindreading is theoretically motivated and we are starting to see its predictions tested and confirmed. The view is committed to an efficient system present in infants having representational powers limited by the (minimal) model of the mental it relies on. There is accumulating evidence—involving different ages, populations and paradigms—showing that an efficient system tracking belief-like states can handle some visual-perspective and false-belief problems, but not others. Research is needed to map the terrain of the efficient (versus flexible) mindreading system, whether it is limited to handling certain kinds of agents, desire-like states, trait impressions, and perspective-based utterances of low complexity (1, 39, 40). Studying the temporal course of behavioral and neural activity associated with tracking belief-like states versus ascribing belief in real-time settings will also illuminate circumstances where information might pass between systems, and delineate precise moments in time when mindreading inferences are constructed, stored and used. New thinking about the cognitive architecture of human mindreading as involving multiple systems, models and signature limits may be necessary for making sense of dissociations both between different response classes and also between non-automatic and automatic processing.

REFERENCES

1. Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116,* 953–970. doi:10.1037/a0016923.

2. Butterfill, S. A. & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28,* 606–637. doi:10.1111/mila.12036

3. Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103–128. doi:10.1016/0010-0277(83)90004-5

4. Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72,* 655–684. doi:10.1111/1467-8624.00304

5. Rakoczy, H., Bergfeld, D., Schwarz, I., & Fizke, E. (2015). Explicit theory of mind is even more unified than previously assumed: Belief ascription and understanding aspectuality emerge together in development. *Child Development, 86,* 486–502. doi:10.1111/cdev.12311

6. Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9,* 377–395. doi:10.1016/0885-2014(94)90012-4

7. Low, J. (2010). Preschoolers' implicit and explicit false-belief understanding: Relations with complex syntactical mastery. *Child Development, 81,* 579–615. doi:10.1111/j.1467-8624.2009.01418.x

8. Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology, 80,* 201–224. doi:10.1006/jecp.2001.2633

9. Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308,* 255–258. doi:10.1126/science.1107621

10. Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Science, 14,* 110–117. doi:10.1016/j.tics.2009.12.006

11. Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C. & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science, 17,* 841–844. doi:10.1111/j.1467-9280.2006.01791.x

12. Apperly, I. A. (2010). *Mindreaders: The cognitive basis of "Theory of Mind*. Hove, East Sussex, UK: Psychology Press.

13. Schneider, D, Nott, Z. E., & Dux, P. E. (2014). Task instructions and implicit theory of mind. *Cognition, 133,* 43–47. doi:10.1016/j.cognition.2014.05.016

14. Van der Wel, R. P. R. D., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition, 130,* 128–133. doi:10.1016/j.cognition.2013.10.004

15. Carey, S. (2009). *The origins of concepts*. New York, NY: Oxford University Press.

16. Schneider, D., Lam, R. Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts theory of mind processing. *Psychological Science, 23,* 842–847. doi:10.1177/0956797612439070

17. Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology, 30,* 172–187. doi:10.1111/j.2044-835X.2011.02067.x

18. Samson D., Apperly I. A., Braithwaite J. J., Andrews B. J., & Bodley Scott S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception & Performance, 36,* 1255–1266. http://dx.doi.org/10.1037/a0018729

19. Furlanetto, T., Becchio, C., Samson, D., & Apperly, I. A. (2016). Altercentric interference in Level 1 visual perspective-taking reflects the ascription of mental states, not submentalizing. *Journal of Experimental Psychology: Human Perception and Performance*, 42, 158-163. http://dx.doi.org/10.1037/xhp0000138

20. Santiesteban, I., Catmur, C., Coughlan Hopkins, S., Bird, G., & Heyes, C. (2014). Avatars and arrows: Implicit mentalizing or domain-general processing? *Journal of Experimental Psychology: Human Perception and Performance, 40,* 929–937. http://dx.doi.org/10.1037/a0035175

21. Surtees, A. D. R., Butterfill, S. A., & Apperly, I. A. (2012). Direct and indirect measures of Level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology, 30,* 75–86. doi:10.1111/j.2044-835X.2011.02063.x

22. Surtees, A. D. R., Samson, D. & Apperly, I. A. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition, 148,* 97–105. doi:10.1016/j.cognition.2015.12.010

23. Mozuraitis, M., Chambers, C. G., Daneman, M. (2015). Privileged versus shared knowledge about object identity in real-time referential processing. *Cognition, 142,* 148–165. doi:10.1016/j.cognition.2015.05.001

24. Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science, 24,* 305–311. doi: 10.1177/0956797612451469

25. Wang, B., Hadi, N. H., & Low, J. (2015). Limits on efficient human mindreading: Convergence across Chinese adults and Semai children. *British Journal of Psychology, 106,* 724–740. doi: 10.1111/bjop.12121

26. Fizke, E., Butterfill, S. A., & Rakoczy, H. (2013). *Toddlers' understanding of false belief about an object's identity*. Poster presented at the Meeting of the Society for Research in Child Development. Seattle, USA.

27. Cacchione, T., Schaub, S., & Rakoczy, H. (2013). Fourteen-month-old infants infer the continuous identity of objects on the basis of non-visible causal properties. *Developmental Psychology, 49,* 1325–1329. http://dx.doi.org/10.1037/a0029746

28. Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition, 112,* 337–342. doi:10.1016/j.cognition.2009.05.006

29. Buttelmann, F., Suhrke, J., & Buttelmann, D. (2015). What you get is what you believe: Eighteen-month-olds demonstrate belief understanding in an

unexpected-identity task. *Journal of Experimental Child Psychology, 131,* 94–103. doi:10.1016/j.jecp.2014.11.009

30. Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development, 80,* 1172–119. doi:10.1111/j.1467-8624.2009.01324.x

31. Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive Psychology, 82,* 32–56. doi:10.1016/j.cogpsych.2015.08.003.

32. Carruthers, P. (2015). Mindreading in adults: Evaluating two-systems views. *Synthese*. Advance online publication. doi:10.1007/s11229-015-0792-3

33. Cohen, A., & German, T. (2009). Encoding of others' beliefs without overt instruction. *Cognition, 111,* 356–363. doi: 10.1016/j.cognition.2009.03.004.

34. Back, E., & Apperly, I. A. (2010). Two sources of evidence on the non-automaticity of true- and false-belief ascription. *Cognition, 115,* 54–70. doi: 10.1016/j.cognition.2009.11.008.

35. Sabbagh, M., Xu, F., Carlson, S., & Moses, L., & Lee, K. (2006). The development of executive functioning and theory-of-mind: A comparison of Chinese and U.S. preschoolers. *Psychological Science, 17,* 74–81. doi: 10.1111/j.1467-9280.2005.01667.x

36. Wellman, H. M. (2014). *Making minds*. New York, NY: Oxford University Press.

37. Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adult performance on a non-inferential theory of mind task. *Cognition, 106,* 1093–1108. doi:10.1016/j.cognition.2007.05.005

38. Low, J., & Simpson, S. (2012). Effects of labeling on preschoolers' explicit false-belief performance: Outcomes of cognitive flexibility or inhibitory control? *Child Development, 83,* 1072–1084. doi: 10.1111/j.1467-8624.2012.01738.x.

39. Wang, J. J., Ali, M., Frisson, S., & Apperly, I. A. (2015). Language complexity modulates 8- and 10-year-olds' success at using their theory of mind abilities in a communication task. *Journal of Experimental Child Psychology*. Advance online publication. doi:10.1016/j.jecp.2015.09.006

40. Schneider, D., & Low, J. (2016). Efficient versus flexible mentalizing in complex social settings: Exploring signature limits. *British Journal of Psychology, 107,* 26–29. doi:10.1111/bjop.12165

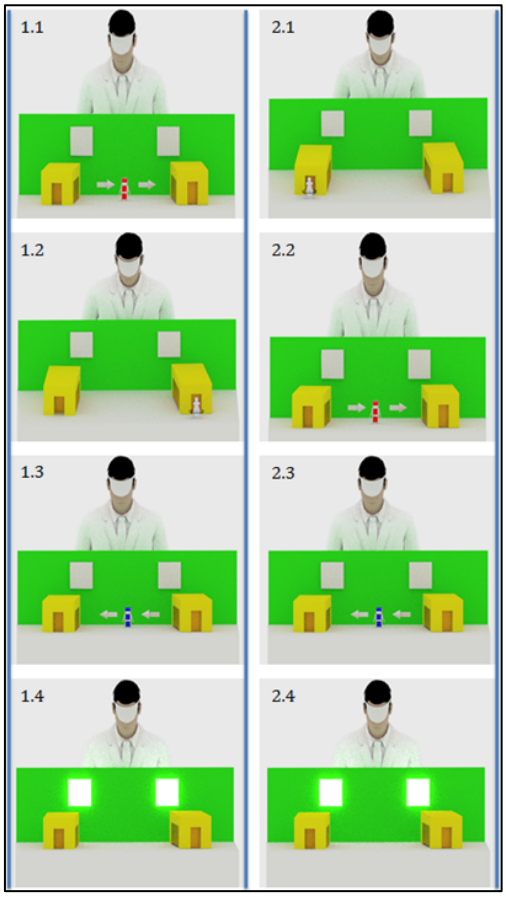

Figure 1. In Low and Watts' (2013) identity task (Column 1), the robot's red and blue aspects are revealed after it moved from the right-side box-A to the left-side box-B (Frame 1.2). Inside box-B, visible only to participants, the robot spun around to reveal its red and blue sides. Then the robot, with its blue aspect facing participants, moved back to box-A. If viewers represent object identities, they should anticipate that the agent falsely believes that there is another (blue) robot inside box-B. The agent (having a blue-color preference, for example) would have reason to reach into box-B. If participants tracked object registrations, then the robot is inside box-A and the agent should search there. In Wang, Hadi and Low's (2015) modified version (Column 2), dual identity was revealed inside box-A before the robot's first movement (Frame 2.1). In both versions, participants showed incorrect looking responses (to box-A) with age-related increases in accuracy of verbal predictions.