Theory of Mind

Chapter · October 2018

DOI: 10.1007/978-3-319-16999-6_2376-1

2 authors, including:

Evan Westra
York University
17 PUBLICATIONS **128** CITATIONS

SEE PROFILE

T

Theory of Mind

Evan Westra[1] and Peter Carruthers[2]
[1]Univeristy of Rochester, Rochester, NY, USA
[2]University of Maryland, College park, MD, USA

Synonyms

Folk psychology; Mentalizing; Mental-state attribution; Mind reading

Definition

The capacity to predict and interpret behavior by using representations of hidden, causally efficacious mental states.

Introduction

"Theory of mind" consists in the ability to use concepts of intentional mental states, such as beliefs, emotions, intentions, goals, and perceptual states, in order to predict and interpret behavior. Functional magnetic resonance imaging studies have revealed a distinctive network of neural regions that is active during theory-of-mind tasks, including the temporal-parietal junction, the posterior superior temporal sulcus, the medial prefrontal cortex, the precuneus, and the temporal poles (Van Overwalle 2009). Deficits in theory-of-mind abilities, which are common in autism spectrum disorder (Tager-Flusberg 2007), typically correlate with broader difficulties in social understanding.

Many scholars have suggested that theory of mind is an innate adaptation for social cognition, emerging very early in development and playing a crucial role in social learning and the acquisition of language (Baron-Cohen 1997). However, others have argued that theory of mind is the product of largely domain-general learning processes and is acquired gradually over the course development through social experience (Wellman 2014). A third view argues that humans possess two systems for theory of mind: an innate, domain-specific "implicit" system and a learned, domain-general "explicit" system (Apperly and Butterfill 2009).

The False-Belief Task Controversy

The classic experimental paradigm for testing theory-of-mind abilities in children is the false-belief task (Wimmer and Perner 1983). In the most common version of this task, participants observe an agent place an object in a box and then leave the room. While the first agent is gone, a second agent moves the object to another hidden location. When the first agent returns, the participant is asked to predict where the agent will look for the object (or in some versions of the task, to say where she thinks the object is). The correct

T.K. Shackelford, V.A. Weekes-Shackelford (eds.), *Encyclopedia of Evolutionary Psychological Science*,
https://doi.org/10.1007/978-3-319-16999-6_2376-1

answer is "in the box." In order to pass the task, the participant must represent that the agent has a *belief* about the location of the object that differs from what is really the case and use that representation to accurately predict what she will do next. Similar tasks test the capacity to reason about other mental states, including desires, states of knowledge, perceptual access, and emotions (Wellman and Liu 2004).

Typically, children fail the false-belief task until after their fourth birthday, as do many adults with autism spectrum disorder (Wellman et al. 2001). Since these results were first discovered, there has been controversy over their interpretation. "Constructivists" have argued that children's shifting performance on this task reflects the emergence of a new concept of *belief*, amounting to a fundamental change in children's intuitive theory of the social world, analogous to theoretical changes brought on by scientific discovery (Gopnik and Wellman 1992). Children, in other words, are learning what beliefs are and the circumstances in which people have them. "Nativists," in contrast, have pointed to the selective deficits on the false-belief task displayed by people with autism spectrum disorder as evidence of an underlying theory-of-mind module, which is selectively impaired in autism (Baron-Cohen 1997). Nativists also argue that younger children's difficulties on the false-belief task don't reflect the absence of a concept of belief but rather a performance error due the immaturity of their domain-general executive capacities. According to this view, the concept of belief is innate, but children can fail to deploy it successfully in certain experimental contexts. Importantly, the nativist view doesn't have to claim that theory of mind is unaffected by experience and individual learning. Rather, the claim can be that these changes are both constrained and facilitated by a domain-specific, innately channeled, learning mechanism that comes equipped with a few basic mental-state concepts (Carruthers 2015).

Two Views on the Evolution and Development of Theory of Mind

In the background of the nativist approach is a theoretical commitment concerning the importance of theory of mind in the evolution of human social intelligence. On this view, innate adaptations for theory of mind are thought to have emerged early in the hominid line. In the highly social environments of ancestral hominids, accurate prediction and interpretation of the behavior of conspecifics would have been crucial for survival. Accounts of the earliest emergence of theory-of-mind abilities in primates emphasize the adaptive importance of theory of mind in social competition, deception, and manipulation, as individuals sought to gain mating opportunities and to improve their status within their group's social hierarchy (Byrne and Whiten 1988). This "Machiavellian intelligence" hypothesis is supported by evidence that modern great apes seem to demonstrate at least simple forms of theory-of-mind abilities in competitive but not cooperative experimental contexts (Call and Tomasello 2008).

Accounts of the evolution of more highly developed, and perhaps specifically human, theory-of-mind abilities, in contrast (including capacities to reason about the false beliefs of others), tend to place greater emphasis on the cooperative functions of theory of mind, particularly when coordinating multiple agents in the pursuit of mutually shared goals, such as group hunting and foraging (Tomasello 2014). Inferences about beliefs and intentions are also thought to have played a crucial role in the emergence of early systems of gestural communication, such as pointing and pantomime, which require inferences about mental states on the part of both the communicator and the audience (Scott-Phillips 2014). These cooperative environments are thought to have created an adaptive feedback loop, where selection pressures for more complex theory-of-mind abilities led to more complex forms of cooperation, which created further selection pressures on our theory of mind. Thus, theory of mind is thought to have fueled the evolution of highly complex forms of mutualistic cooperation,

in addition to supporting the development of these same social abilities in ontogeny.

Constructivists typically accept that human beings possess some special adaptations for social intelligence but generally deny that these amount to a genuine theory of mind. For example, some constructivists acknowledge that neonates are innately biased to attend to faces and eyes and are innately disposed to engage in imitative behavior (e.g., Meltzoff 2007). According to constructivist accounts, these low-level, noncognitive mechanisms serve as a scaffold for children's early theory-of-mind development by directing their attention toward socially relevant phenomena. The latter then serve as inputs for domain-general learning procedures, such as statistical learning (Ruffman et al. 2012). Many constructivists also believe that children's acquisition of a mature theory of mind depends on exposure to specific linguistic inputs, such as clausal complementation syntax or mental-state vocabulary (e.g., de Villiers and Pyers 2002). Thus, while nativists posit that the ancestral emergence of domain-specific cognitive adaptations for theory of mind made complex forms of cooperation and linguistic communication possible, constructivists hold that the existence of complex cooperative environments lay the developmental foundation for children to acquire a theory of mind via individual learning, which implies that it is an evolutionarily recent, culturally dependent phenomenon (Heyes and Frith 2014).

Does Theory of Mind Emerge Early or Late in Development?

Support for the constructivist view comes from evidence that theory-of-mind development is influenced by both social experience and language. For instance, having older siblings tends to lead to earlier success on the false-belief task, as do greater amounts of parental mental-state discourse (Ruffman et al. 2012). Strikingly, deaf children who do not learn sign language until later in life also have persistent difficulties on the false-belief task, even in adulthood (Pyers and Senghas 2009). These findings seem incompatible with the nativist's performance-error account of children's performance on the false-belief task, suggesting instead that children's social and linguistic environment plays an important role in determining their theory-of-mind abilities.

In response to these findings, some nativists have argued that they reflect the fact that children must learn how to *apply* their innate theory-of-mind abilities in different social and linguistic contexts and that variations in social environment impact this learning process (Westra 2017). In effect, the suggestion is that younger children fail at the tasks because their grip on discourse pragmatics is weak, leading them to misunderstand the point of the questions they are asked, and it is this that is impacted by social experience.

In support of the nativist view, a large body of evidence has emerged more recently, demonstrating a range of theory-of-mind abilities in children in the first 2 years of life, well before they pass verbal forms of the false-belief task. For example, a number of studies have adapted the classic false-belief task to make it suitable for infants. Importantly, these tasks never ask children to make explicit, verbal predictions. Instead, they rely on children's spontaneous behaviors to measure their theory-of-mind abilities. For example, one such task presented 15-month-olds with a scene in which an experimenter hid an object in one of two boxes and then left. While the experimenter was absent, the object was moved to the other box. When the experimenter returned, infants either saw her reach toward the first, empty box, or the second box, where the object then was. Results show that infants look far longer when the experimenter reaches for the second box, suggesting that they find this behavior surprising (Onishi and Baillargeon 2005). The infants were seemingly expecting the agent to reach into the box where she *believed* the goal object to be.

Researchers have also designed ways of testing early false-belief competence by exploiting the fact that young children are highly motivated to help other people achieve their goals (Buttelmann et al. 2009). In these tasks, 18-month-olds observe an experimenter place a toy in one of two boxes and close the lid. The experimenter then leaves the room, and the children see a second experimenter

move the toy from the first box to the second box. Then the first experimenter returns and begins to struggle with the lid of the first box. The authors predicted that if children understood that the experimenter desired the toy but had a false belief about its location, they should respond by retrieving the toy from the second box. In a true-belief control task in which the first experimenter observes the location change, in contrast, children should instead help the experimenter open the first box, presuming that she must want something inside it. Indeed, this is what they found.

The interpretation of these results, like those of the original false-belief task, has been a subject of great controversy. While nativists have used infant false-belief paradigms to support the claim that the capacity for representing beliefs is innate (Baillargeon et al. 2010), critics of these paradigms have offered various alternative, low-level explanations of the same findings. According to some of these interpretations, infants' successful performance on these tasks may not reflect an abstract concept of belief but rather the tracking of statistical regularities in observable behavior (Ruffman et al. 2012). Nativists have responded to these criticisms by pointing out that these alternative explanations have tended to be post hoc and have failed to issue in new data (Scott 2014). The nativist framework, in contrast, has generated a continuous stream of new results, employing an ever-widening set of experimental paradigms. However, these debates are currently ongoing.

Can Apes Pass False-Belief Tasks?

We noted in the section "Two Views on the Evolution and Development of Theory of Mind" that there is evidence that other great apes have at least a simple form of theory of mind, one that allows them to track and reason about the goals and states of knowledge or ignorance of other agents. But until recently, all tests for false-belief understanding in other primates had proven negative (Call and Tomasello 2008). Recently, however, researchers have successfully adapted for use among primates some of the methods for testing false-belief competence in young children, with striking results. In one study, the experimenters showed chimpanzees, bonobos, and orangutans videos depicting an interaction between an actor dressed as a zookeeper and an actor dressed as a gorilla. The videos began by showing the gorilla hitting the zookeeper and then running into one of two haystacks arrayed at either side of the screen. While the zookeeper turned away to fetch a stick (which, in familiarizations, was used to hit the haystack containing the gorilla), the gorilla moved from one haystack to the other. Then the zookeeper turned around with the stick raised, poised to attack. The researchers used an eye tracker to measure whether the apes would look in anticipation toward the actual location of the gorilla or to the location where the zookeeper will think the gorilla is hiding (i.e., anticipating and reasoning from the zookeeper's false belief). The results showed that all three species of ape looked reliably more toward the false-belief location, suggesting that they were indeed tracking the zookeeper's beliefs (Krupenye et al. 2016). In another study, researchers adapted the active helping design of Buttelmann et al. (2009) (see above) for use with chimpanzees, bonobos, and orangutans. They, too, found that great apes seemed to use information about false beliefs in order to help an experimenter retrieve an object from a locked box (Buttelmann et al. 2017).

Some critics have argued that these results, like the infant false-belief-task results, should not be interpreted in rich, mentalistic terms; instead, we should prefer explanations that only attribute to apes the ability to make predictions about behavior based on low-level, observable regularities, such as the appearance and disappearance of the colored shirt of the zookeeper (Heyes 2017). To test this hypothesis, Krupeneye and colleagues designed a control task that matched their original anticipatory looking paradigm but replaced the actors with inanimate colored shapes (Krupenye et al. 2017). They found that, in contrast to the task involving actors, the participants were no more likely to look toward either target location when observing the same interaction between inanimate shapes. This provides compelling evidence that the apes' predictive gaze behavior was specifically sensitive to the social nature of the

stimuli, as opposed to its low-level properties. All of these results are, however, quite recent, and further research is necessary before conclusions can confidently be drawn.

Thus, two recent studies seem to show that three other species of great ape are capable of passing false-belief tasks. This result, if it is valid, has striking implications for the debate about the language dependence of theory of mind. If nonlinguistic primates are able to represent beliefs, this casts serious doubt on the claim that such an ability requires linguistic experience, or is uniquely human. These findings also suggest that belief representation may in fact be evolutionarily quite ancient and potentially shared by the last common ancestor of humans, chimpanzees, bonobos, and orangutans.

Such inferences should be drawn cautiously, however: the presence of robust theory-of-mind ability in modern great apes does not necessarily imply that it is an innate adaptation. For it is possible that these apes acquired their abilities through individual learning (and hence that ancestral apes might have done so as well). In other words, even if great apes do reason about false beliefs, this doesn't necessarily provide support for nativism about theory of mind in human beings. What it does do, however, is undermine the claim that false-belief understanding is specifically dependent on either *linguistic* experience or experience of distinctively human forms of collaboration and joint action.

The two-Systems View

Another set of theorists has attempted to resolve the dispute between nativists and constructivists by proposing an ecumenical solution that incorporates both constructivist and nativist elements (Apperly and Butterfill 2009). According to these "two-systems" accounts, infants' early competence on "implicit" false-belief tasks (and by parity of reasoning, the recent successful performance of other great ape species on similar tasks) does reflect a domain-specific adaptation for tracking mental states. This "implicit" theory-of-mind system is said to be fast, effortless, automatic, and largely encapsulated from executive systems; it also persists unchanged into adulthood and is shared with our nearest primate relatives. Thus, in many respects this implicit mindreading system resembles the kind of domain-specific adaptation posited by nativists. However, two-systems theorists hold that due to its automatic and encapsulated architecture, the implicit mindreading system is subject to "signature limits." In particular, while it can represent beliefs about the *locations* of objects (e.g., "Bill believes that the apple is in the box"), it cannot handle beliefs about the *identity* of objects (e.g., "Bill believes that the apple is really a pear"). This is because the implicit system is thought to *track* beliefs and other mental states without representing them as such and, in particular, without representing the aspectual nature of belief states. (Famously, one can believe that Jocasta is beautiful without believing that one's mother is beautiful, even though Jocasta is in fact one's mother. Here one and the same person is thought about under two different aspects.) The implicit system cannot, therefore, fully capture the richness and flexibility of the mature concept of belief.

In humans, one of the functions of the implicit system is to scaffold the acquisition of a second, parallel, explicit theory-of-mind system that develops much more gradually. This system is said to be slow, effortful, and heavily reliant upon executive resources, such as working memory. It develops gradually via domain-general learning in response to social and linguistic input and only emerges after children's fourth birthday (thus enabling them to pass the explicit false-belief task). Unlike the implicit system, this one isn't subject to signature representational limits. However, because it relies heavily on working memory, it must be directed by top-down goals and is compromised under cognitive load. The explicit mindreading system closely resembles the conception of mindreading posited by constructivists, albeit supported by a relatively more complex set of domain-specific cognitive adaptations. Thus, by adulthood, human beings are said to possess two parallel theory-of-mind systems, each with a distinct information-processing profile.

To test the prediction that the implicit mindreading system is subject to signature limits, two-systems theorists have compared adults' and young children's implicit theory-of-mind predictions in scenarios where agents have false beliefs about either an object's location or its identity (Low and Watts 2013). To test participants' ability to track false beliefs about identity, the experimenters constructed a scenario in which participants were familiarized with an agent who demonstrated a consistent preference for a certain color, always reaching for blue items rather than red items, for example. Next, out of sight of the agent, participants were familiarized with a paper cutout figure that appeared as a red robot from one side and as a blue robot on the other. They then watched as the agent observed what she would have seen as a blue robot enters one of two boxes. Next, unbeknownst to the agent, the figure rotated 180° and moved to the second box while presenting to the agent as a red robot. Because the agent didn't know that the red and blue robots are the same individual, she should believe that there was still a blue robot in the first box. Thus, if participants were tracking beliefs about identity, they should expect the agent to reach toward the first box. However, neither adults nor children reliably looked in anticipation toward the first box, suggesting that they were insensitive to the agent's beliefs about object identity. Meanwhile, both groups showed correct anticipatory looking on the control task that only involved beliefs about the object's location.

These results provide support the claim that while implicit theory of mind accurately tracks beliefs about the locations of objects, it doesn't track beliefs about the identities of objects. However, critics of this study and others like it have pointed out that the object-identity task, which involves effortful forms of mental rotation, is likely to place additional demands on executive function that are not present in the object-location control task. This makes it unclear whether the relevant failure is due to signature limitations on implicit theory of mind or rather limitations on working memory (Carruthers 2015). It is therefore unclear whether this evidence truly supports a two-systems view. Nevertheless, two-systems accounts of theory of mind have significantly influenced debates about the cognitive architecture of our social cognition abilities.

Conclusion

As can be observed in the debate about the false-belief task, there is considerable disagreement in the literature about the role of theory-of-mind abilities in socio-cognitive development: while nativists believe that theory-of-mind is a basic adaptation that plays an important role in early social learning, constructivists hold that theory of mind is itself the *product* of social learning. These views correspond to distinct narratives about the emergence of theory of mind in the hominid line: if the nativist view is right, then theory of mind is evolutionarily ancient; if the constructivist view is correct, then it is likely to be a far more recent phenomenon. The two-systems view attempts to carve a middle ground between these two camps, allowing that some aspects of theory of mind are indeed evolutionarily ancient and early developing, while others are evolutionarily novel and acquired via social learning. However, the evidence in favor of this view is itself a source of controversy.

Cross-References

- ▶ Autism
- ▶ Communication and social cognition
- ▶ Cultural Intelligence Hypothesis
- ▶ Does the chimpanzee have a theory of mind?
- ▶ Intentional stance, the
- ▶ Michael Tomasello
- ▶ Nativism
- ▶ Simon Baron-Cohen
- ▶ Social intelligence hypothesis
- ▶ Theory of mind and nonhuman intelligence
- ▶ Theoy of mind and evidence of brain modularity

References

Apperly, I., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970.

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

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

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.

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PLoS One, 12*(4), e0173793.

Byrne, R. W., & Whiten, A. (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford: Clarendon Press.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187–192.

Carruthers, P. (2015). Two systems for mindreading? *Review of Philosophy and Psychology, 6*, 2.

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development, 17*, 1037–1060.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7*(1–2), 145–171.

Heyes, C. (2017). Apes submentalise. *Trends in Cognitive Sciences, 21*(1), 1–2.

Heyes, C., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science, 344*(6190), 1243091–1243091.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science, 354*(6308), 110.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2017). A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. *Communicative & Integrative Biology, 10*, e1343771.

Low, J., & Watts, J. (2013). Attributing false-beliefs about object identity is a signature blindspot in humans' efficient mindreading system. *Psychological Science, 24*(3), 305–311.

Meltzoff, A. N. (2007). "Like me": A foundation for social cognition. *Developmental Science, 10*(1), 126–134.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.

Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science, 20*(7), 805–812.

Ruffman, T., Taumoepeau, M., & Perkins, C. (2012). Statistical learning as a basis for social understanding in children. *The British Journal of Developmental Psychology, 30*(Pt 1), 87–104.

Scott, R. M. (2014). Post hoc versus predictive accounts of children's theory of mind: A reply to Ruffman. *Developmental Review, 34*(3), 300–304.

Scott-Phillips, T. (2014). *Speaking our minds: Why human communication is different, and how language evolved to make it special* (Vol. 3). Basingstoke: Palgrave Macmillan.

Tager-Flusberg, H. (2007). Evaluating the theory-of-mind hypothesis of autism. *Current Directions in Psychological Science, 16*(6), 311–315.

Tomasello, M. (2014). *A natural history of human thinking*. Cambridge, MA: Harvard University Press.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford: Oxford University Press.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523–541.

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

Westra, E. (2017). Pragmatic development and the false belief task. *Review of Philosophy and Psychology, 8*(2), 235.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128.