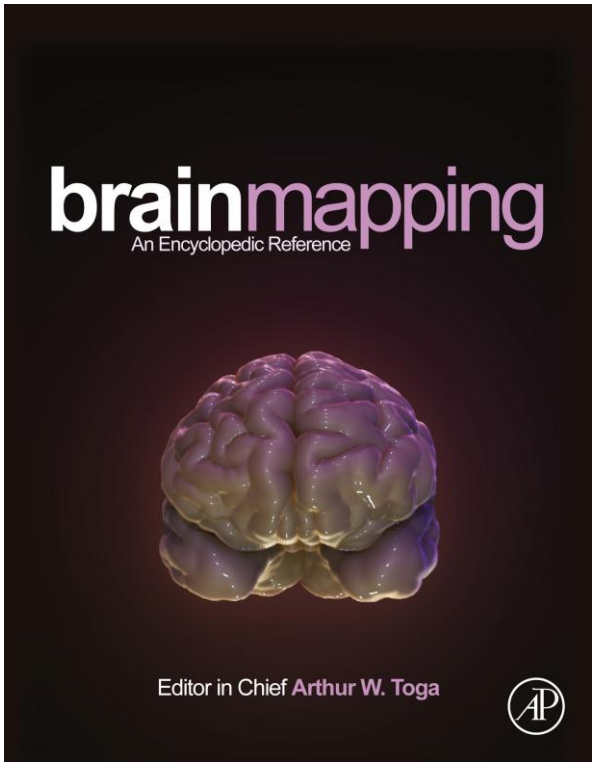

Wagner D.D. (2015) Mentalizing. In: Arthur W. Toga, editor. *Brain Mapping: An Encyclopedic Reference*, vol. 3, pp. 143-146. Academic Press: Elsevier.

Mentalizing

DD Wagner, The Ohio State University, Columbus, OH, USA

The ability of humans to understand, predict, and influence the thoughts, emotions, and beliefs of others is, if not uniquely human, then at least among the most sophisticated examples of social cognition in the animal kingdom. Although terminology may vary, the ability to *mentalize* has been the subject of great interest in disciplines as diverse as philosophy, animal cognition, developmental psychology, social psychology, and, more recently, social and cognitive neuroscience. With regard to the neural basis of mentalizing, the impetus for research on the topic arose primarily from two sources: research on theory of mind in developmental and neuropsychology and, separately, social psychological research on how we perceive, categorize, and form impressions of other persons (i.e., person perception).

Theory of Mind

In a seminal paper on chimpanzee cognition, Premack and Woodruff (1978) posed the question: 'does the chimpanzee have a theory of mind?' That is to say, do chimpanzees understand the behavior of their conspecifics as reflecting desires and personal motives distinct from their own? Although the issue of whether or not chimpanzees are capable of 'mentalizing' remains controversial (e.g., Call & Tomasello, 2008), this early attempt to study mentalizing sparked an interest in the phenomenon by hinting at a method for testing for the presence or absence of the ability to mentalize. This method, which was realized several years later (Wimmer & Perner, 1983), became known as the false-belief task. The false-belief task is perhaps the first and certainly the most ubiquitous measure for the presence of a theory of mind. The test is designed to assess whether or not an individual possesses the capacity to comprehend that their belief about the external world may conflict with another person's otherwise false belief. Developmentally normal children and adults typically have little trouble passing this test, whereas children with autism encounter difficulty (Baron-Cohen, Leslie, & Frith, 1985).

It is on the basis of these observations of impaired theory of mind in autism that the first neuroimaging studies of mentalizing were constructed. For instance, in the first studies to examine the neural substrates of mentalizing, Fletcher and colleagues, using positron emission tomography (PET), contrasted brain activity during the reading and comprehension of stories requiring the inference of mental states (i.e., stories in which the behavior of a character can only be understood with respect to their false belief about a situation) with brain activity when participants read stories describing physical events, such as an alarm going off following an intrusion. This study demonstrated that reading the mentalizing stories resulted in greater brain activity in the dorsal medial prefrontal cortex (MPFC), temporal poles, posterior cingulate, and superior temporal sulcus (STS) (Fletcher et al., 1995). A second study appearing around the same time also used PET to show that inferring whether historical characters would have knowledge of certain concepts (e.g., would Columbus know about telephones?) led to increased brain activity in a nearly identical set of regions, namely, the dorsal MPFC, temporal poles, and STS (Goel, Grafman, Sadato, & Hallet, 1995). Strikingly, the results of these first two studies of mentalizing are still very much in accord with present-day neuroimaging results using improved methods and technology. Since that time, other researchers have gone on to examine specific facets of mentalizing (e.g., false-belief understanding; Saxe & Kanwisher, 2003), as well as to examine mentalizing in other modalities, such as cartoons (Gallagher et al., 2000), animations (Castelli, Happe, Frith, & Frith, 2000), and movies (Mar, Kelley, Heatherton, & Macrae, 2007), all with similar results. Another method of studying mentalizing under a more naturalistic setting can be seen in the work of Gallagher, Jack, Roepstorff, and Frith (2002) in which people were engaged in a competitive game against an ostensibly real person or a computer (in truth, they were always playing against the computer). Interestingly, when people were led to believe they were playing against a real person, they demonstrated greater activity in the dorsal MPFC as compared with when they thought they were playing against a computer opponent (Gallagher et al., 2002).

Person Perception and Social Cognition

In the previous section, we discussed the role of developmental and neuropsychological findings in driving interest in the brain systems underlying mentalizing. However, social psychologists have long been concerned with how people understand the social behavior of others, and thus, it is no surprise that a significant portion of research on mentalizing originated from social psychologists interested in the neural processes underlying person perception.

Person perception is generally concerned with how people perceive, construe, categorize, and form impressions of others. For example, people readily form impressions of others, and these impressions help organize person knowledge into schemas, thereby increasing the ability to recall information about other people (e.g., Hamilton, Katz, & Leirer, 1980). This tendency to organize information about other people into distinct impressions is thought to be largely automatic (Uleman, 1999). For instance, people readily form impressions of others in the absence of any requirement to do so (Todorov & Uleman, 2002), even going so far as to falsely remember seeing words describing their impressions in a subsequent memory test (Todorov & Uleman, 2004).

Findings, such as those outlined earlier, served as a springboard for social psychologists who wished to investigate the brain basis of person perception. Although the tasks used by these researchers differed from the false belief-type tasks described in the previous section, they nevertheless can be

construed as another approach to studying mentalizing. Early work in this vein demonstrated that forming impressions of other people recruited essentially the same regions as false-belief and other theory of mind tasks (e.g., the MPFC, temporal poles, and STS). For instance, in one of the earliest functional neuroimaging studies of impression formation, it was found that the dorsal MPFC represents semantic knowledge about people (i.e., psychological traits) but not about objects, suggesting that the neural representation of person knowledge is dissociable from semantic knowledge for nonsocial categories (Mitchell, Heatherton, & Macrae, 2002).

Subsequent work demonstrated that forming impressions of others and making trait inferences also recruit the dorsal MPFC as well as other regions of the mentalizing system such as the temporal poles, precuneus, and STS (Cloutier, Kelley, & Heatherton, 2011; Mitchell, Macrae, & Banaji, 2004; Mitchell, Neil Macrae, & Banaji, 2005; Todorov, Gobbini, Evans, & Haxby, 2007). This work also demonstrated that the mentalizing system was specifically recruited when forming impressions of people and nonhuman animals (i.e., animate agents that either have mental states like humans or invite mental state attributions, like animals) but not when forming impressions of inanimate objects (Mitchell, Cloutier, Banaji, & Macrae, 2006; Mitchell et al., 2005).

Another line of research on person perception has examined the neural correlates of viewing familiar others. In general, this work has shown that viewing faces of personally familiar others, such as friends and family (Gobbini, Leibenluft, Santiago, & Haxby, 2004; Leibenluft, Gobbini, Harrison, & Haxby, 2004), or of unfamiliar others who have become familiar through the acquisition of person knowledge (e.g., Cloutier, Kelley, & Heatherton, 2011; Todorov et al., 2007) leads to increased activity in the mentalizing system, particularly the dorsal MPFC. Taken together, this work on impression formation and familiar others suggests that the dorsal MPFC is involved in representing person knowledge not only during impression formation but also when cued by viewing the faces of familiar others for which people possess intimate person knowledge (for a more detailed review, see Wagner, Haxby, and Heatherton, 2012).

Are There Brain Regions Necessary for Mentalizing?

As discussed earlier, studies of mentalizing and person perception both recruit a relatively circumscribed set of brain regions that we refer to as the mentalizing system. The degree to which each of these brain areas are considered essential for normal mentalizing remains a topic of some debate, although there is some evidence suggesting that damage to these areas leads to impairments on mentalizing and person perception tasks. For instance, damage to the prefrontal cortex has been found to impair performance on theory of mind tasks (Stuss, Gallup, & Alexander, 2001, although see Bird, Castelli, Malik, Frith, & Husain, 2004), and with regard to atypical populations, patients with autism show reduced activity in the MPFC when viewing social animations meant to elicit mentalizing. Additional evidence pointing to a potentially critical role for the MPFC in mentalizing and interpersonal behavior comes from a case study of two patients with extensive MPFC damage. In this study, it was found that, following the trauma, these patients experienced the sudden development of what has been described as autistic-like personality traits. Specifically, these patients underwent a change in personality that was marked by a reduction in the ability to empathize with others (Umeda, Mimura, & Kato, 2010).

Finally, a study using transcranial magnetic stimulation (TMS) provides strong evidence that at least one region of the mentalizing system, the posterior STS/temporoparietal junction (TPJ), is critical for mentalizing about other's moral behavior (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). In this study, the authors found that inactivation of the right TPJ by repetitive TMS impaired people's ability to make normal judgments of intentional moral transgressions, leading them to judge these less harshly compared with when TMS was applied to a control site.

The Relationship Between Action Understanding and Mentalizing

One common route for understanding the behavior of others is by attempting to understand the goal of their immediate actions. Research on the neural basis of action understanding has focused primarily on a frontoparietal network of brain areas that are involved in both action observation and planning. For example, the intraparietal sulcus is often found to be activated not only when people observe goal-directed actions such as reaching and grasping (Hamilton & Grafton, 2006) but also when they observe more complex familiar movements such as dancing and smoking (Cross, Hamilton, & Grafton, 2006; Wagner, Cin, Sargent, Kelley & Heatherton, 2011).

A number of studies have examined the interplay between action understanding and mentalizing. For instance, when attempting to infer the emotions that others are feeling, people can either attempt to infer the underlying mental states that contribute to outward displays of emotion or attempt to understand emotions by analyzing the body language and facial expressions of the person experiencing them. Interestingly, as people switch from one strategy to the other, separate neural systems are recruited such that when people rely more on outward facial expressions, they recruit brain regions involved in action observation, whereas when they rely more on a mentalizing strategy, they recruit regions involved in mentalizing (Zaki et al., 2009). Outside the domain of emotion, similar findings have been shown for understanding common everyday behaviors in terms of how people are performing them (i.e., the motor actions contributing to a behavior) or why people are performing them (i.e., the underlying motives for a behavior). As in the previously mentioned work, when people adopt a 'why' stance, they recruit brain regions involved in mentalizing to a greater degree than when asking 'how' the behavior is being performed. In contrast to the mentalizing system, regions implicated in action understanding where not modulated by the participants particular strategy, showing equal recruitment for both 'why' and 'how' (Spunt, Satpute, & Lieberman, 2011).

Spontaneous Mentalizing

In much of daily life, forming impressions of others and attributing mental states to them are a largely automatic process. Neuroimaging methods have afforded a unique opportunity to

examine how brain regions involved in mentalizing are modulated by task demands and stimulus content. A number of studies have now shown that many regions implicated in mentalizing are recruited spontaneously in the absence of any explicit mentalizing or impression formation goal, so long as the stimulus invites a mental state inference. For instance, studies have shown that observing deception (German, Niehaus, Roarty, Giesbrecht, & Miller, 2004), watching social interactions (Iacoboni et al., 2004; Wagner, Kelley, & Heatherton, 2011), and even viewing abstract animations that give the appearance of a social interaction (Castelli, Frith, Happe, & Frith, 2002; Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007) recruit areas of the mentalizing system despite the fact that subjects are not explicitly tasked with mentalizing.

Similar findings have been found in the person perception domain showing that when participants read statements that communicated specific information about a personality trait (i.e., trait diagnostic) versus statements that did not (e.g., he photocopied the article), they show greater activity in the mentalizing system even when not explicitly engaged in impression formation (Mitchell, Cloutier, Banaji, & Macrae, 2006). Another example of spontaneous mentalizing comes from a study showing that when participants are playing a virtual driving game in which their task is to taxi around a passenger, they show greater activity in the MPFC and temporal poles during those parts of the task that they later report they were thinking about the mental states of the passenger (Spiers & Maguire, 2006).

Finally, it is interesting to note that there are also individual and developmental differences that are related to spontaneous mentalizing. For instance, compared to healthy individuals, patients with autism show reduced recruitment of the dorsal MPFC when viewing social animations (Castelli, Frith, Happe, & Frith, 2002). Similarly, among the healthy population, individual differences in empathizing (a measure of people's propensity to engage in mentalizing and experience empathy) are correlated with the degree to which people spontaneously recruit the dorsal MPFC when viewing scenes of social interactions (Wagner, Kelley, & Heatherton, 2011).

Mentalizing and Its Relationship with Everyday Social Behavior

Much of the work reviewed thus far focuses on the neural systems involved in mentalizing and how these are modulated by tasks, goals, or stimulus content. More recently, researchers have started to examine whether the recruitment of these regions can be used to predict everyday social behaviors. Prior work has demonstrated that, for instance, activity in the MPFC during a person perception task is related to individual differences in empathizing (Wagner, Kelley, & Heatherton, 2011); however, it is only recently that researchers have investigated whether activity in this region is related to prosocial behavior. One of the first studies to examine this possibility had participants engage in a mentalizing task in which they judged the preferences and opinions of a confederate. Following the mentalizing task, they completed a monetary allocation task in which they could allocate money to themselves or the other person. What they found was that activity in the dorsal MPFC during the mentalizing task predicted how much money participants would later donate to the confederate. Moreover, activity in this region also predicted how time participants later volunteered to help the confederate in a separate problem solving task (Waytz, Zaki, & Mitchell, 2012).

Outside the realm of prosocial behavior, brain activity during mentalizing has also been shown to predict the persuasiveness of an idea. In this study, a set of participants were asked to watch a series of videos explaining a set of fictive television pilot ideas; these participants were then tasked with communicating these ideas to a new group of people who would then go on to judge which television show idea should be recommended for production. Here, activity in the dorsal MPFC during encoding of the initial television show ideas in the first set of participants predicted which shows the other group of participants would go on to approve. That is to say, the degree to which participants engaged the mentalizing system when first learning about the show predicted which shows the separate group of judges would find most worth pursuing (Falk et al., 2013).

See also: **INTRODUCTION TO SOCIAL COGNITIVE NEUROSCIENCE**: Mentalizing and Psychopathology in Schizophrenia, Depression, and Social Anxiety; Neural Correlates of Social Cognition Deficits in Autism Spectrum Disorders; Person Knowledge and Attribution; Self-Knowledge; Strategic Mentalizing: The Neural Correlates of Strategic Choice; The Default Network and Social Cognition; The Neural Correlates of Social Cognition and Social Interaction; Trust Perception.

References

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

Bird, C. M., Castelli, F., Malik, O., Frith, U., & Husain, M. (2004). The impact of extensive medial frontal lobe damage on "Theory of Mind" and cognition. *Brain*, *127*(4), 914–928.

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

Castelli, F., Frith, C., Happe, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, *125*, 1839–1849.

Castelli, F., Happe, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, *12*(3), 314–325.

Cloutier, J., Kelley, W. M., & Heatherton, T. F. (2011). The influence of perceptual and knowledge-based familiarity on the neural substrates of face perception. *Social Neuroscience*, *6*, 63–75.

Cross, E. S., Hamilton, A. F., & Grafton, S. T. (2006). Building a motor simulation de novo: Observation of dance by dancers. *NeuroImage*, *31*, 1257–1267.

Falk, E. B., Morelli, S. A., Welborn, B. L., Dambacher, K., & Lieberman, M. D. (2013). Creating buzz: The neural correlates of effective message propagation. *Psychological Science*, *24*(7), 1234–1242.

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., et al. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*, 109–128.

Gallagher, H. L., Happe, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, *38*, 11–21.

Gallagher, H. L., Jack, A., Roepstorff, A., & Frith, C. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, *7*, 77–83.

German, T. P., Niehaus, J. L., Roarty, M. P., Giesbrecht, B., & Miller, M. B. (2004). Neural correlates of detecting pretense: Automatic engagement of the intentional stance under covert conditions. *Journal of Cognitive Neuroscience*, *16*(10), 1805–1817.

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal Cognitive Neuroscience*, *19*, 1803–1814.

Gobbini, M. I., Leibenluft, E., Santiago, N., & Haxby, J. V. (2004). Social and emotional attachment in the neural representation of faces. *NeuroImage*, *22*, 1628–1635.

Goel, V., Grafman, J., Sadato, N., & Hallet, H. (1995). Modeling other minds. *NeuroReport*, *6*, 1741–1746.

Hamilton, A. F., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. *Journal of Neuroscience*, *26*, 1133–1137.

Hamilton, D. L., Katz, L. B., & Leirer, V. O. (1980). Cognitive representation of personality impressions: Organizational processes in first impression formation. *Journal of Personality and Social Psychology*, *39*(6), 1050–1063.

Iacoboni, M., Lieberman, M. D., Knowlton, B. J., Molnar-Szakacs, I., Moritz, M., Throop, C. J., et al. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage*, *21*, 1167–1173.

Leibenluft, E., Gobbini, M. I., Harrison, T., & Haxby, J. V. (2004). Mothers' neural activation in response to pictures of their children and other children. *Biological Psychiatry*, *56*, 225–232.

Mar, R. A., Kelley, W. M., Heatherton, T. F., & Macrae, C. N. (2007). Detecting agency from the biological motion of veridical versus animated agents. *Social Cognitive and Affective Neuroscience*, *2*, 199–205.

Mitchell, J. P., Cloutier, J., Banaji, M. R., & Macrae, C. N. (2006). Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Social Cognitive and Affective Neuroscience*, *1*, 49–55.

Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 15238–15243.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2004). Encoding-specific effects of social cognition on the neural correlates of subsequent memory. *Journal of Neuroscience*, *24*, 4912–4917.

Mitchell, J. P., Neil Macrae, C., & Banaji, M. R. (2005). Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. *NeuroImage*, *26*, 251–257.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 515–526.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind" *NeuroImage*, *19*, 1835–1842.

Spiers, H. J., & Maguire, E. A. (2006). Spontaneous mentalizing during an interactive real world task: An fMRI study. *Neuropsychologia*, *44*, 1674–1682.

Spunt, R. P., Satpute, A. B., & Lieberman, M. D. (2011). Identifying the what, why, and how of an observed action: An fMRI study of mentalizing and mechanizing during action observation. *Journal Cognitive Neuroscience*, *23*, 63–74.

Stuss, D. T., Gallup, G. G., Jr., & Alexander, M. P. (2001). The frontal lobes are necessary for 'theory of mind'. *Brain*, *124*(2), 279–286.

Todorov, A., Gobbini, M. I., Evans, K. K., & Haxby, J. V. (2007). Spontaneous retrieval of affective person knowledge in face perception. *Neuropsychologia*, *45*, 163–173.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personal and Social Psychology*, *83*, 1051–1065.

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology*, *87*(4), 482–493.

Uleman, J. S. (1999). Spontaneous versus intentional inferences in impression formation. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 141–160). New York: Guilford.

Umeda, S., Mimura, M., & Kato, M. (2010). Acquired personality traits of autism following damage to the medial prefrontal cortex. *Social Neuroscience*, *5*, 19–29.

Wagner, D. D., Cin, S. D., Sargent, J. D., Kelley, W. M., & Heatherton, T. F. (2011). Spontaneous action representation in smokers when watching movie characters smoke. *Journal of Neuroscience*, *31*, 894–898.

Wagner, D. D., Haxby, J. V., & Heatherton, T. F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*, 451–470.

Wagner, D. D., Kelley, W. M., & Heatherton, T. F. (2011). Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. *Cerebral Cortex*, *21*, 2788–2796.

Waytz, A., Zaki, J., & Mitchell, J. P. (2012). Response of dorsomedial prefrontal cortex predicts altruistic behavior. *Journal of Neuroscience*, *32*, 7646–7650.

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.

Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *PNAS*, *107*, 6753–6758.

Zaki, J., Weber, J., Bolger, N., & Ochsner, K. (2009). The neural bases of empathic accuracy. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 11382–11387.