Brain Regions Involved in Theory of Mind

Daniel Borowski

Abstract

Humans have the ability to interpret and infer mental statues such as intentions, feeling, desires, and beliefs of other people through a mechanism known as "Theory of Mind." The original task to measure one's ability of attributing mental states, in this case false belief, was the Sally-Anne Task created by Baron-Cohen, Leslie, and Frith (1985). In the task, Sally has a marble with two boxes in front of her. She places the marble into a box and then leaves the room. While she is gone Anne comes in, moves the marble to the other box, and then leaves. The question now asked to children is where will Sally look for the marble when she enters the room again. To correctly predict Sally's behavior, it is necessary to take into account both Sally's desire for the marble and Sally's belief concerning the location of the marble (Leslie, 2000). There have been numerous tasks similar to this one over the years testing children's abilities to attribute various mental states to different situations and to different people or objects. With the technology available to us now, Theory of Mind and its relation to our brain areas has been gathering quite a lot of attention along with new studies being performed to determine where in the brain this mechanism for Theory of Mind is controlled. This paper will provide a description of the brain areas involved in Theory of Mind processing network and how damage to these specific areas can affect the attribution of mental states to others and can also contribute to other mental issues that may arise in individuals such as autism, alexithymia, and schizophrenia.

There is an interest in the brain basis of Theory of Mind because a better understanding of where it is processed in the brain and whether or not it is modular can ultimately lead to a clearer picture of our neural wiring. Before the year 2000, there were very few reports published in regards to using neuroimaging techniques to study Theory of Mind. In 1994 (Baron-Cohen S, Ring H, Moriarty J, Schmitz B, Costa D, Ell P.) had used single photon emission computerized tomography (SPECT) to determine an increase of cerebral blood flow in the right orbito-frontal cortex. A study run by (Goel V, Grafman J, Sadato N, and Hallett M.) in 1995 determined there was an activation in the left medial frontal lobe and left temporal lobe using a PET scan. Another study run by (Fletcher PC, Happea F, Frith U, Baker SC, Dolan RJ, Frackowiak RSJ, and Frith CD.) in 1995 had people read stories that involved mental state attribution and stories that were similar but involved physical objects instead of people and found an activation in the left medial prefrontal gyrus as well as in the posterior cingulate gyrus. One of the first experiments using fMRI technology was run by (H.L. Gallagher, F. Happe b, N. Brunswick, P.C. Fletcher, U. Frith, and C.D. Frith.) using a similar method to that of Fletcher et al. where people's neural activity would be imaged while reading stories. Their goal was to identify regions involved in Theory of Mind in the verbal and visual domains by showing people different types of pictures on a screen along with stories either invoking the process of attributing mental states to people or not. The results of showing Theory of Mind *stories* vs. non-Theory of Mind were an activation in the medial prefrontal cortex, the temporal poles bilaterally and the temporo-parietal junctions bilaterally (Gallagher et al., 2000). As for Theory of Mind *cartoons* vs. non-Theory of Mind, there was also activation in the medial prefrontal cortex along with the temporo-parietal junctions bilaterally, and also the right middle frontal gyrus, the precuneus and the fusiform. In

both cases, the medial prefrontal cortex was the area solely activated in Theory of Mind cartoons and stories. A conclusion from these experiments is that there was significant activation in the medial prefrontal cortex and the temporo-parietal junctions bilaterally in Theory of Mind stories and cartoons as opposed to non-Theory of Mind. The other parts of the brain previously listed had slight activation but not as much as the aforementioned areas. It becomes clear that the ability to mentalise is mediated by the medial prefrontal cortex since it was an area activated uniquely in Theory of Mind stories and cartoons. The temporo-parietal junctions bilaterally also had a large activation but was activated in the non-Theory of Mind controls.

R. Saxe and N. Kanwisher ran an experiment that also involved people reading stories, but their interest lied more specifically in the activation of the temporo-parietal junction. They attempted to prove that the TPJ did in fact activate only in Theory of Mind stories and the reason Gallagher et al. had the issue of it being activated in Theory of Mind and non-Theory of Mind stories was because their non-Theory of Mind stories still invited inferences about a characters true beliefs (Saxe et al., 2003). They had devised their own version of the false-belief story task adopted from Fletcher et al. for use in their own experiment. They created stories that dealt with a characters true or false belief, but they also contrived two other types of stories: the first involved mechanical inferences (i.e. forces of gravity, inertia, melting etc.) and the second was simply a description of nonhuman objects. Now the researchers would be able to isolate brain areas so that when a story was being read, they could clearly identify the brain areas involved in the processing of mental states of others. Their experiment revealed that the TPJ bilaterally, or TPJ-M, had an increase in activation during Theory of Mind stories as opposed to the mechanical inference stories. This activation is robust and reliable across individual subjects... Our results

confirm that the TPJ-M response to verbal descriptions generalizes to human actions based on true beliefs (Saxe et al., 2003). An interesting finding though, was that the left TPJ-M was primarily activated in verbal descriptions, while the right TPJ-M was activated in non-verbal descriptions, such as pictures or videos. After the Saxe et al. and H.L. Gallagher et al. experiments, a conclusion that can be drawn is that the medial prefrontal cortex and the temporo-parietal junction bilaterally, or TPJ-M, are both heavily activated in the brain when subjects must attribute some sort of mental states to other humans, thereby verifying that Theory of Mind processing is modular within specific areas of the brain.

After several more imaging studies were run by various researchers to determine the function and network of Theory of Mind, more areas have been discovered in its involvement. Apart from TPJ-M and the medial prefrontal cortex, the anterior paracingulate cortex has been discovered to be heavily involved in allowing individuals to attribute mental states to others and themselves which is also known as mentalizing (some refer to damage to this area as Mind-Blindness). Two recent studies suggest that the anterior paracingulate cortex is the key region for mentalizing (McCabe, K. et al., 2001; Gallagher, H.L. et al., 2002). Instead of the typical experiment where subjects read stories and their neural activity is measured, this time Gallagher et al. had people play a game of "rock, papers, scissors" against a computer. The people were told that they were either going to play against a strategically capable machine, a machine with a random set of moves, or another human being. The philosopher Dan Dennett describes playing against the hidden human being and attributing some sort of mental states to it such as goals and desires as an "intentional stance." After imaging the individuals brains as they were playing, they discovered one area that had a surge of activity when told they were playing against a human

being (in fact all of them were playing against computers), and that was the anterior paracingulate cortex. No further regions appeared even when the statistical threshold was lowered to $p = 0.1$ (Gallagher, H.L. et al., 2002). In the experiment run by McCabe, K. et al., they had also discovered increased activity in the same area. It requires the ability to infer each other's mental states to form shared expectations over mutual gains and make cooperative choices that realize these gains (McCabe, K. et al., 2001). As it turns out in both studies, people were playing "games" against either other human beings or computers, and there was activity in the anterior paracingulate cortex whenever people thought that they were either playing against or cooperating with another human being with their own beliefs, desires, and goals. Gallagher and Frith state that the reason they believe in previous experiments there was activity in the medial prefrontal cortex and other areas in the brain was because other cues were available to the person such as eye gaze and expressive body language which may have activated other brain areas. The only difference between the mentalizing condition and the control conditions in both of these studies lay in the "stance" adopted by the volunteers (H.L. Gallagher, C. D. Frith, 2003). Based off these two studies, the ability to attribute mental states now seems to lie in only one area, the anterior paracingulate cortex, and the previously mentioned activated brain areas could have resulted from other cues or uncontrolled stimuli during the "story" experiments.

There are two other regions that are activated across a number of Theory of Mind experiments. They are the superior temporal sulcus (STS) and the temporal poles bilaterally. These two new areas are not completely understood in the roles they play in Theory of Mind, but Gallagher et al. (2000) has found that the right STS seems to be associated with understanding the meaning of stories involving people with or without the act of attributing mental states. There

are also some functional imaging studies that show STS activity when perceiving hand actions (Grezes, J. et al., 1998), body movements (Bonda, E. et al., 1996), mouth movements and lip-reading (Puce, A. et al., 1998), and eye-movements and gaze direction (Wicker, B. et al., 1998). The temporal poles bilaterally are generally associated with object and face recognition in primates (Nakamura, K. and Kubota, K., 1996). There are also studies dealing with semantic and episodic memory that also activate the temporal poles bilaterally (Dolan, R.J. et al., 2000; Fink, G.R. et al., 1996). There may be several reasons why people's temporal poles bilaterally are activated in Theory of Mind experiments. An individual may need to consider if they had been deceived in the past by a specific person so they may need to rely on their memory of that person. A memory of someone may also help us realize that the person is trying to be deceitful and that they should not be believed. The STS and the temporal poles bilaterally seem to clearly play a role in the Theory of Mind network by allowing us to use other faculties of our mind in the attribution of mental states to others. These two areas, along with the other areas forms a network of five possible brain regions contributing to the Theory of Mind network: the medial prefrontal cortex, TPJ bilaterally, STS, temporal poles bilaterally, and the anterior paracingulate cortex.

Armed with knowledge of this neural network for mentalizing, Dominic Marjoram, et al., (2006) performed a study where they displayed visual jokes to patients with schizophrenia and examined their brains under an fMRI to investigate their Theory of Mind capabilities. Their hypothesis was that significant differences in activity would be found in the STS, the temporal poles, and the medial prefrontal cortex when compared to normal, healthy adults. They showed the different groups three types of cartoons: one which required Theory of Mind reasoning such

as attributing a false belief to a character, another cartoon which was a slapstick, physical type of humor that didn't require any Theory of Mind reasoning, and a third type which wasn't meant to contain any jokes but instead was composed of jumbled images. After a detailed analysis between different risk levels of schizophrenics and of healthy people, one conclusion reached was that in contrast estimate plots for these maximal voxels show that in all of these the control group was activating less than most of the high-risk subject groups (Dominic Marjoram, et al., 2006). They claim that high-risk individuals have an impaired Theory of Mind network due to their high risk for schizophrenia which could possibly require some form of compensation from activity in other brain areas. The groups hypothesis was that they would see key activation differences in the STS, prefrontal cortex areas, and the temporal poles bilaterally, but in a few of their analyses there was actually only a major difference in the middle and medial prefrontal cortices. They also hypothesize that it could be that that appreciation of visual jokes does not require these episodic memories of social scripts to the extent of other Theory of Mind paradigms, particularly verbal stories (Dominic Marjoram, et al., 2006).

Another group of researchers wanted to study how the impairment of some areas in the Theory of Mind network affect alexithymic individuals. These individuals have difficulty in recognizing and describing emotions in themselves. Autistic spectrum disorders, including Asperger's syndrome, are characterized by an impairment of Theory of Mind (Baron-Cohen et al., 1985). Asperger's syndrome is also associated with high alexithymia scores (Berthoz and Hill, 2005). The researchers Yoshiya Moriguchi, et al. (2006), set out to investigate neuronal activity in individuals with high alexithymia scores in a Theory of Mind task using fMRI technology. The hypothesis was that alexithymia would be associated

with decreased neuronal activity in the medial prefrontal cortex, the TPJ bilaterally, and the temporal pole. Subjects were asked to assess silent visual animations. After testing women and men with alexithymia and without, they demonstrated differences between the individuals in behavioral and neural responses in mentalizing such as a decreased activity in the medial prefrontal cortex in alexithymics. Their results showed relative inactivity in the right medial prefrontal cortex in alexithymic individuals. In our study, individals with high alexithymia showed less activation in the MPFC in responese to the mentalizing task. Thus, there could be a common neural component for both normal individuals with high alexithymia and autistic people with impaired mentalizing ability, which also indicates a common component of understanding the self and others (Yoshiya Moriguchi, et al., 2006). The conclusion from their experiment demonstrates that hypoactivity in the medial prefrontal cortex can be associated with a deficiency in the ability to identify and describe feelings of the self and feelings of others which is an impairment in the ability to mentalize.

There are also other disorders that have hypoactivity in the medial prefrontal cortex. Individuals with autistic spectrum disorder have showed less activation in the MPFC when tested on mentalizing tasks (Goel et al., 1995).

The five brain regions, namely the medial prefrontal cortex, the temporal parietal junction bilaterally, superior temporal sulcus, temporal poles bilaterally, and the anterior paracingulate cortex, along with a few slightly less significant areas make up what is known as the Theory of Mind network or circuit. Damage to the anterior paracingulate cortex can lead to what is known as Mind Blindness (McCabe, K. et al., 2001; Gallagher, H.L. et al., 2002). Various studies in the field of studying Theory or Mind using neuroimaging techniques point to the APC as the critical

area where Theory of Mind processing takes place. The other areas all aid in some way to this process either by relying on episodic memories, by measuring eye gaze or body movements, or by some other ways. Many researchers have set out to study various mental disorders involving Theory of Mind processing to see how they relate to damage to specific areas in the neural network. Schizophrenic, alexithymic, and autistic spectrum disorder patients all seemed to have an impairment to their medial prefontal cortices. Although the anterior paracingulate cortex area seemed to be what allowed people to have an "intentional stance" when confronted with another person, it still seems as though the MPFC plays a major role in the Theory of Mind network.

Baron-Cohen S, Leslie A, Frith U. (1985). Does the autistic child have a “theory of mind”?

Baron-Cohen S, Ring H, Moriarty J, Schmitz B, Costa D, Ell P. (1994). The brain basis of theory of mind: the role of the orbito-frontal region. British Journal of Psychiatry.

Berthoz, S., Hill, E.L., (2005). The validity of using self-reports to assess emotion regulation abilities in adults with autism spectrum disorder. Eur. Psychiatry.

Bonda, E. et al. (1996) Specific involvement of human parietal systems and the amygdala in the perception of biological motion. J. Neuroscience.

Dolan, R.J. et al. (2000) Dissociable temporal lobe activations during emotional episodic memory retrieval. NeuroImage.

Dominic Marjoram, et al. (2006). A visual joke fMRI investigation into Theory of Mind and enhanced risk of schizophrenia. NeuroImage.

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RSJ, Frith CD. (1995). Other minds in the brain: a functional imaging study of `theory of mind' in story comprehension. Cognition.

Gallagher H.L, F. Happe b, N. Brunswick, P.C. Fletcher, U. Frith, and C.D. Frith. (2000). Reading the mind in cartoons and stories: an fMRI study of `theory of mind' in verbal and nonverbal tasks. Neuropsychologia.

Gallagher, H.L. et al. (2002) Imaging the intentional stance. NeuroImage.

Gallagher H.L, Frith C. D. (2003). Functional imaging of ‘theory of mind’. TRENDS in Cognitive Sciences.

Goel V, Grafman J, Sadato N, Hallett M. (1995). Modelling other minds. Neuroreport.

Grezes, J. et al. (1998) Top-down effect of strategy on the perception of human biological motion: a PET investigation. Cognitive Neuropsychology.

Leslie A. (2000). “Theory of Mind” as a Mechanism of Selective Attention. THE NEW COGNITIVE NEUROSCIENCES.

McCabe, K. et al. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. Proc. Natl. Acad. Sci. U.S.A. 98.

Nakamura, K. and Kubota, K. (1996) The primate temporal pole: its putative role in object recognition and memory. Behav. Brain Res.

Puce, A. et al. (1998) Temporal cortex activation in humans viewing eye and mouth movements. J. Neurosci.

Saxe R, Kanwisher N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind”. NeuroImage.

Wicker, B. et al. (1998) Brain regions involved in the perception of gaze: a PET study. Neuro-Image.

Yoshiya Moriguchi, et al. (2006). Impaired self-awareness and theory of mind: An fMRI study of mentalizing in alexithymia. NeuroImage.