Judgments About Others' Trustworthiness: An fMRI Study

Motoki Watabe[1,*], Hiroshi Ban[2], Hiroki Yamamoto[3]

[1] Waseda Institute for Advanced Study, Waseda University, 1-6-1 Nishi-waseda, Shinjuku-ku, 169-8050, Japan

[2] School of Psychology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[3] Graduate School of Human and Environmental Studies, Kyoto University, Yoshida Nihonmatu-sho, Sakyo-ku, Kyoto, 606-8501, Japan

*Author for correspondence (mwatabe@aoni.waseda.jp)

We investigated how information about others' trustworthiness affects brain region activation in a functional MRI (fMRI) study. Participants were given statements about a person's behaviors and were asked to judge whether or not the person was trustworthy while undergoing fMRI imaging. Participants read 32 statements, half of which were relevant to making judgments about trustworthiness, and half of which were irrelevant to making judgments about trustworthiness. We found that making trustworthiness judgments when reading relevant statements was associated with differential activation in five regions: the angular gyrus (AG), anterior cingulate (AC), left frontal lobe (LF), right frontal lobe (RF), and putamen/caudate nucleus (PU/CA). Previous study using a highly abstract economic game situation has also shown activation in these regions. These regions are also related to the learning process and to theory-of-mind processing. In addition, we found that people with high or low scores on a general trust scale showed less activation than did people with middle-range scores. These results suggest that we use trial-and-error learning to decide whether to trust others, and that this learning history (represented here as general trust level) influences automatic processing of new trust judgments.

Keywords
trust, fMRI, theory of mind, learning

Introduction

How do we trust others?

In deciding whether or not one should become socially involved with another person, most people would probably make judgments about whether or not that person is trustworthy. Social scientists commonly argue that trust, an important component for social capital, helps to create and strengthen social bonds among people. Trust also enhances social and economic exchanges, because people who trust one another can exchange more valuable resources with each other (Nakayachi & Watabe, 2005a, 2005b; Yamagishi, Cook, & Yamagishi, 1998; Zucker, 1986). However, trusting others is risky because it makes people vulnerable to exploitation. Thus, people need to judge how trustworthy a person is. Many experimental game studies have demonstrated that people tend to consider others' past behaviors to avoid future exploitation (e.g., Milinski, Semmann, & Krambeck, 2002). However, little is known about the neurobiological mechanisms by which such judgments are made.

This study aims to explore how individuals make judgments about others' trustworthiness based on information about their past behaviors, using functional Magnetic Resonance Imaging (fMRI) to discover brain regions that activate during such judgments. Previous imaging studies have tried to identify active regions involved in decision-making in game situations such as the Prisoner's Dilemma or the Trust Game (e.g., Delgado, Frank, & Phelps, 2005; Rilling, Sanfey, Aronson, Nystrom, & Cohen. 2004). Participants in these studies played these games repeatedly with feedback information about their partner's decisions, the benefits accrued in the game, and, sometimes, with a picture of their partner's face. However, these studies suffer from two limitations. One is that stimuli that are extraneous to judgments about trustworthiness may activate brain regions that are not related to those judgments. For example, the amygdala is always activated when a person sees a human face, thinks about interpersonal impressions, or thinks about others' emotional states (e.g., Vuilleumier, Richardson, Armony, Driver, & Dolan, 2004). Amygdala activation may therefore occur in response to the stimulus faces, not in response to the judgment about trustworthiness. In order to avoid confounds, the most parsimonious possible design must be employed. In this study, participants used only written information about a person's past behaviors in order to make judgments about trustworthiness.

The other problem with previous studies is that they have typically involved experimental game situations including highly abstract exchanges. In these experiments, participants are given a payoff matrix with monetary rewards and are asked to maximize their own payoff. Our more parsimonious design allows us to approximate more real-life judgments about trustworthiness: Participants were simply asked how much they could trust the persons whose histories they were given. If our participants showed differential activation in the same brain

doi: 10.5178/lebs.2011.16
Received 17 May 2011.
Accepted 23 May 2011.
Published online 28 July 2011.

regions as in previous experimental game studies, this would help to establish the external validity of the experimental abstract paradigms.

Research Outline

The study had two phases. First, we identified information that was relevant to decision-making about trustworthiness, and distinguished it from information that was not relevant to such judgments. In the second phase, we used fMRI to identify regions that are differentially activated when participants are given relevant vs. irrelevant information and asked to make a judgment about trustworthiness.

Study 1: Selection of Trust Information

(a)Participants

Eighty undergraduate students in a Japanese university who were recruited in classes. All of them were Japanese men (mean age = 21.03, SD = 1.01).

(b)Procedure

Participants were given 58 statements about a person's behavior. Eighteen of the statements were quoted from Kosugi & Yamagishi (1998), and 40 were composed by the researchers. Half of the 58 statements were intended to be relevant to judging trustworthiness (e.g., "Person A cheated on an examination"), and the rest were intended to be irrelevant (e.g., "Person A wears glasses.") All of the relevant statements were negative in valence (i.e., intended to influence participants to judge the target person as untrustworthy). Participants were asked to read each statement and to evaluate each target person's trustworthiness using a four-point Likert scale ranging from "1: Surely Trustworthy" to "4: Surely Untrustworthy." If the participants were not able to evaluate trustworthiness based on the statement, they were asked to check "5: Don't know."

Distinction between trust information and positive impression to a person

Trustworthiness is an important component for positive impression to a person. So, it is impossible to separate trustworthiness from general positive impression. However we did the following two things to focus only on trustworthiness and tried to get rid of the other factors for positive impression. First, we were careful to choose the episodes describing only target person's trustworthiness and tried to avoid the episodes describing just prosocial behaviors in general. Second, we asked participants 1) how trustworthy this person is, and 2) how much they like this person. We picked out the statements according to the evaluation score for trustworthiness only. Not surprisingly, there is a positive correlation between the scores of trustworthiness and likability (r = .59, p < .001).

Analysis

We selected the 16 statements that had the highest average untrustworthiness rating (M = 3.714, SD = 0.496) as "relevant information" for use in the experimental condition of the imaging study. We selected the 16 statements that were checked as "5: don't know" by the largest number of participants (70.3% on average) as "irrelevant information" for use in the control condition of the imaging study.

Study 2: fMRI Experiment

(a)Participants

Twenty-three undergraduates in a Japanese university. All of them were Japanese male (mean age = 21.65, SD = 0.98). None of the study 1 participants participated in study 2.

(b)Procedure

Participants came to the lab individually. They were asked to complete a general trust scale (Yamagishi & Yamagishi, 1994), comprising seven items answered on a seven-point Likert scale ranging from "Completely Disagree" to "Completely Agree." The general trust scale measures respondents' general level of trust in other people. Its validity and reliability has been confirmed in more than 10 studies (e.g., Gheorghiu, Vignoles, & Smith, 2009). Participants then lay flat on their backs in the fMRI machine. Figure 1 illustrates the experimental situation. During the first two to three minutes, participants were asked to simply relax as their brain was initially scanned to measure its size and shape.

After the initial scan, participants were asked to read each of the 32 statements selected from Study 1. After reading each statement, they were asked to judge whether the person described in the statement was trustworthy or not by pressing one of three optical-fiber switches labeled "Trustworthy," "Not trustworthy," or "Cannot judge." They were instructed to make judgments as quickly as possible after being presented with each statement. The statements were presented for 5 seconds each, with a 25-second pause between statements. Statements were presented in random order. To reduce measurement errors, each participant completed the experimental task twice.

Results

Behavioral Data

According to the post-experimental interview, 15 of the all participants reported that they made mistakes on pressing the buttons for the trust judgment because the buttons were small and lined with small interval (about 4mm). Thus, the data of judged trustworthiness was not reliable and we gave up using the behavioral data for analysis.

Regions of Interest (ROI)

fMRI data for two participants who fell asleep during the procedure were discarded. Data for the remaining 21 participants were analyzed.

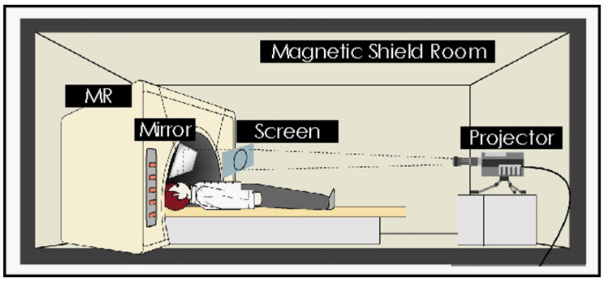

Figure 1 Experimental situation.

The average reaction time was 732 ms, but 12 participants reported that they were confused about which buttons they should press, and made errors during the experiment. We therefore analyzed brain activation at a constant 600 ms after each statement was shown, using a standard GLM analysis .According to a paired t-test, the following regions were significantly activated (p < .005) in response to relevant information: the **angular gyrus (AG), anterior cingulate cortex (AC), left frontal lobe (LF), right frontal lobe (RF), and putamen/caudate nucleus (PU/CA).** Figure 2 shows selected activations and Figure 3 gives a summary illustration of all of the activated regions. In previous research, these regions have been found to be activated during social judgments and complicated tasks. Especially noteworthy is that our observations overlap with Delgado et al. (2005). Whereas Delgado and colleagues' subjects played a complex and abstract trust game, our participants were simply considering information about a person's past actions. This result is evidence that abstract game experiments are valid for making conclusions about trust behaviors in natural settings.

Differential Activation by General Trust Level

We also analyzed the degree of activation in each region across different levels of general trust. We divided participants' general trust scale scores into three categories labeled "high trusters," "middle trusters," and "low trusters." The distribution of the trust scale score (i.e., the mean score across all seven items) approximated a Gaussian distribution, with an overall mean of 4.26 (SD = 1.04). High trusters (n = 8) were defined as those with mean scores greater than 4.33, middle trusters (n = 6) were defined as those with a mean score of exactly 4.33, and low trusters (n = 7) were defined as those with mean scores less than 4.33. Figure 4 shows the mean activation levels of the five regions across high, middle and low trusters. Though the middle trusters show the greatest mean level of activation in every region, this pattern was significant only in the angular gyrus (AG), F(2, 18) = 4.44, p < .05.

Discussion

There are two major findings in this study. First, we identified five brain regions that are apparently related to making judgments about trustworthiness. Because our observations were congruent with studies using experimental game paradigms,

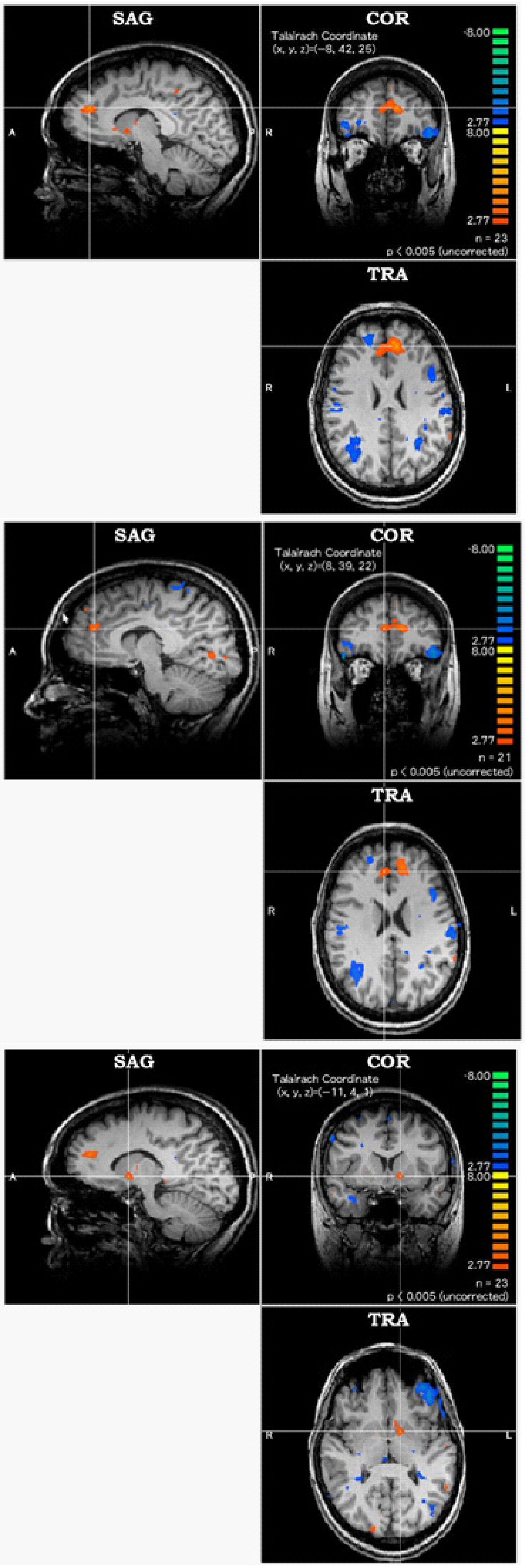

Figure 2 Differential activation in selected brain regions during a trustworthiness judgment task. Top: Anterior cingulate cortex. Middle: Left and right frontal lobes. Bottom: Putamen/caudate nucleus.

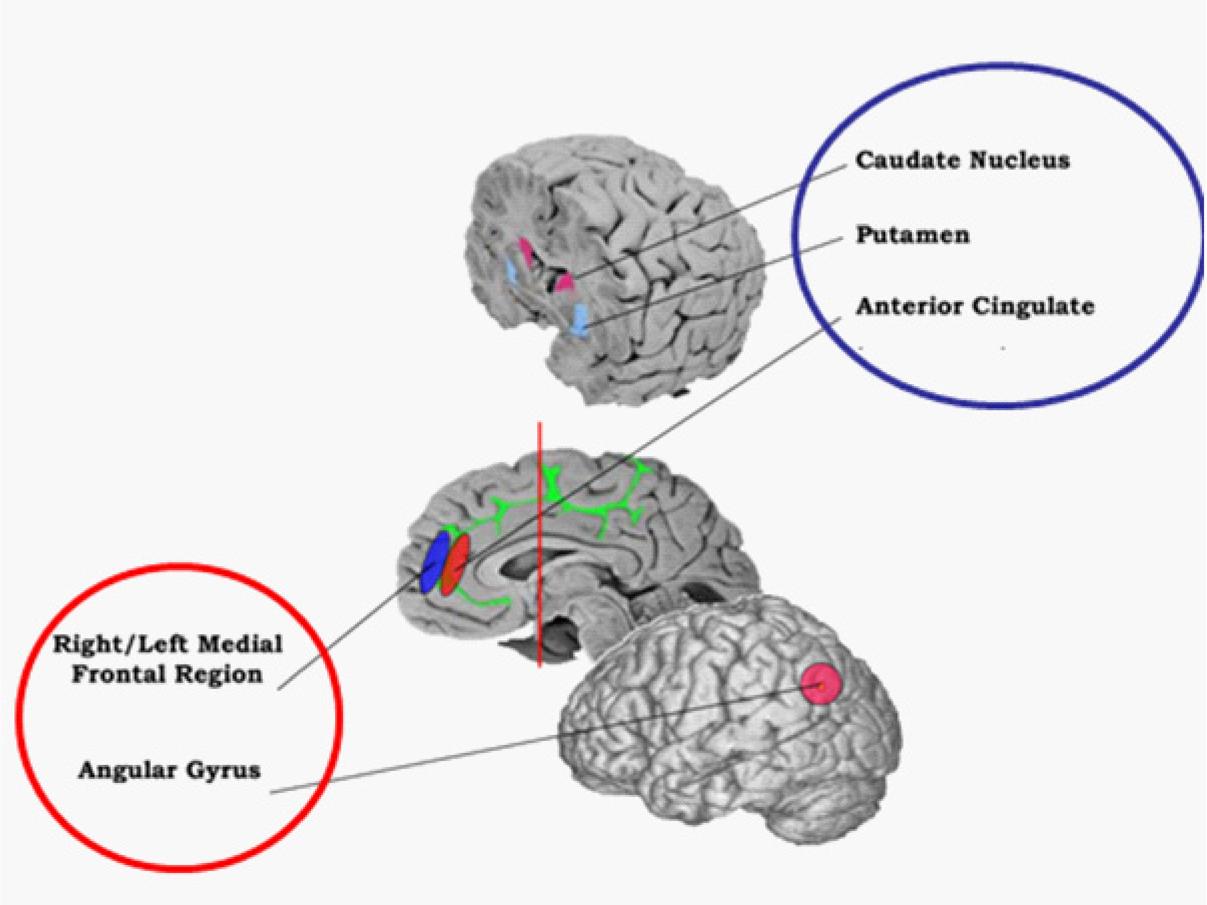

Figure 3 Summary illustration of activated brain regions.

we concluded that these paradigms are valid for making conclusions about trust behaviors in real-life social interactions.

A more interesting question is how these regions are involved in making judgments about trust. This is difficult to answer. Each ROI could influence various behaviors, and it is hard to identify a particular process of decision-making from images of differential brain activation. However, we can guess at some possibilities. First, it is known that the PU/CN is active in temporal difference learning (TDL; e.g., O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003). TDL is often used to account for reinforcement learning in animals and humans. In TDL, people attempt trial-and-error learning repeatedly, using feedback information to adjust future trials and, gradually, leading to optimal decisions (Doya, 2007). The AC, RF, and LF are also related to TDL. A TDL approach to trust suggests that humans learn how to trust others based on observations of others' behaviors, attempting to optimize their judgments over multiple trials. If this is the case, social environments are crucial for the creation of trust. Second, theory-of-mind experiments have shown that the AG is active when one person attempts to model another's cognitions. The AC, RF, and LF are also related to theory of mind. We conclude that judgment of others' trustworthiness is a skill that is developed iteratively and that is based in theory of mind. This argument is, however, still speculative: It remains to be clarified how each ROI works, and for what specific kinds of information processing and behavior each is active.

Our second finding was that AG activation differed depending on participants' general level of trust in others. This suggests that low and high trusters employed automatic information processing, whereas middle trusters incurred cognitive costs. Though they are not statistically significant, the other ROIs' patterns are similar, suggesting that high and low trusters tended to make judgments about trustworthiness more automatically than did middle trusters. It is plausible that people who do not have a cognitive bias will incur more cognitive costs for information processing during a task related to the content area they are not biased about. In this sense, our low trusters and high trusters could be described as cognitive misers (Fiske & Taylor, 1991). One interesting question arising from this result is whether cognitive miserliness is adaptive for making judgments in specific social situations.

Finally we would like to note that this finding may be inconsistent with past finding that high trusters are more sensitive to trust information than low trusters (Kosugi & Yamagishi, 1998). It should be examined how "sensitivity to trust information" represents in brain activity to account for the inconsistency.

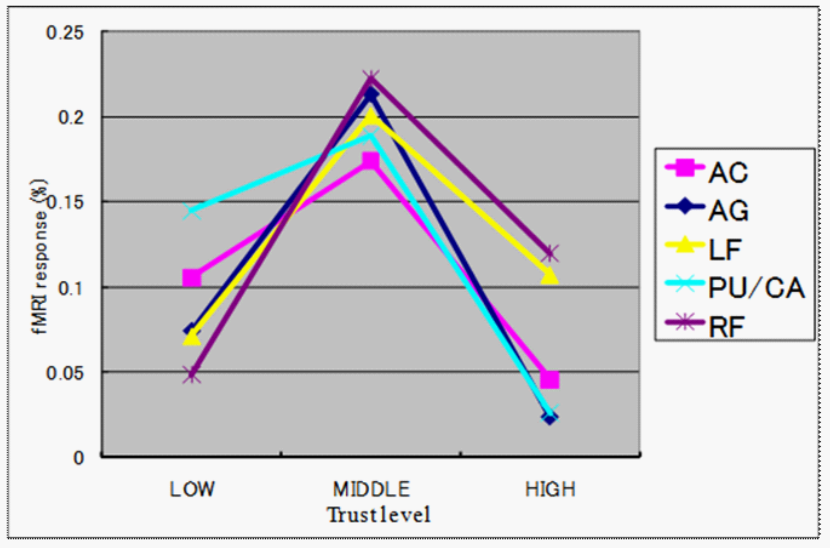

Figure 4 Degree of activation of brain regions by trust level (*p < .05).

References

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. Nature Neuroscience, 8, 1611-1618. (doi:10.1038/nn1575)

Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. HFSP Journal, 1, 30-40. (doi:10.2976/1.2732246/10.2976/1)

Fiske, S. T., & Taylor, S. E. (1991). Social cognition (2nd ed.). New York: McGraw-Hill.

Gheorghiu, M. A., Vignoles, V. L., Smith, P. B. (2009). Beyond the United States and Japan: Testing Yamagishi's emancipation theory of trust across 31 nations. Social Psychology Quarterly, 72, 365-383. (doi:10.1177/019027250907200408)

Kosugi, M, & Yamagishi, T. (1998). Generalized trust and judgments of trustworthiness. The Japanese Journal of Psychology, 69, 349-357. (In Japanese)

Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the 'tragedy of the commons.' Nature, 415, 424-426. (doi:10.1038/415424a)

Nakayachi, K., & Watabe, M. (2005a). Influences of hostage posting on estimation of trustworthiness: The effects of voluntary posting and reliable results. Japanese Journal of Psychology, 76, 235-243.

Nakayachi, K., & Watabe, M. (2005b). Restoring trustworthiness after adverse events: The signaling effects of voluntary “hostage posting” on trust. Organizational Behavior and Human Decision Processes, 97, 1-17. (doi:10.1016/j.obhdp.2005.02.001)

O’Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron, 38, 329-337 (doi:10.1016/S0896-6273(03)00169-7)

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. NeuroReport, 15, 2539-2543. (doi:10.1097/00001756-200411150-00022)

Vuilleumier, P., Richardson, M. P., Armony, J. L. Driver, J., & Dolan, R. J. (2004). Distant influences of amygdala lesion on visual cortical activation during emotional face processing. Nature Neuroscience, 7, 1271-1278. (doi:10.1038/nn1341)

Yamagishi, T., Cook, K. S. C., & Watabe, M. (1998). Uncertainty, trust, and commitment formation in the United States and Japan. American Journal of Sociology, 104, 165-164. (doi: 10.1086/210005)

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. Motivation and Emotion, 18, 129-166. (doi:10.1007/BF02249397)

Zucker, L. G. (1986). Production of trust: Institutional sources of economic structure, 1984-1920. Research in Organizational Behavior, 8, 53-111.