Masaki Nakagome, Kazuo Maki, and Hideto Ide (2014), “The Voice of Conscience Can Defeat the Effect of Bad Apples on Moral Judgment: A Brain Decoding Study on the Effectiveness of Adam Smith’s “an Impartial Spectator,” Working Paper Series, Institute of Economic Research, Aoyama-Gakuin University no.2014-1.

The Voice of Conscience Can Defeat the Effect of Bad Apples on Moral Judgment: A Brain Decoding Study on the Effectiveness of Adam Smith's "an Impartial Spectator"

By Masaki Nakagome, Kazuo Maki and Hideto Ide

Abstract

Our moral judgment is seriously affected by bad apples, persons who ride free and exhibit moral hazard behaviors. However, the voice of conscience is the major force preventing moral hazard. Adam Smith's "The Theory of Moral Sentiments" provides a famous example of the voice of conscience that is represented by "the impartial spectator" or "the man within the breast." To develop Adam Smith's paradigm, we should specify the social and economic conditions under which the voice of conscience can overcome the incentives of agents who are easily tempted by moral hazard. We present an experiment utilizing brain decoding, a powerful brain reading method for interpreting subjects' neural activities and understanding the motivations affecting decision-making in the conscious and unconscious mind.

Acknowledgments

This study was supported in part by a grant from the Japanese Ministry of Education, Culture, Sport, Science, and Technology (Grant in Aid for Scientific Research, No.2020002) awarded to Masaki Nakagome and by a grant from the Research Institute at Aoyama Gakuin University (Grant in Aid for Research in Social Science Areas) awarded to Masaki Nakagome and Hideto Ide. Many thanks are owed to Tetsuji Oyama for his technical support to develop our brain reading method.

1. Introduction

Adam Smith's "The Theory of Moral Sentiments" has been intensively studied in the history of economic thought. Recently, Smith's theoretical framework has been reevaluated from behavioral economic and neuroscience perspectives. The concepts of sympathy and imaginary change of situations were pioneered in Smith's work, which influenced the development of the studies on mirror neuron system and mentalization in the theory of mind.[1] Our study contributes to this body of literature by evaluating the significance of Smith's work utilizing newly developed methodologies in neuroscience.

"The Theory of Moral Sentiments" is a normative analysis of the relationship between empathy and moral judgment. Smith considers the desire to be sympathized with to be part of human nature. Ideally, agents should examine whether they are sympathized with by "an impartial spectator" or "the man in the breast" when they judge the virtue and ethics of their own behavior. The voice of conscience of "the impartial spectator" should dominate the various incentives to select moral hazard.

To develop Smith' paradigm, the normative study must be supported with positive studies. These positive studies examine whether the voice of conscience dominates the incentives of moral hazard in particular situations. We compare the effect of Smith's moral judgment with that of other types of moral judgment. For example, we consider the effect of bad apples on moral judgment. We

[1] Kiesling (2012) examined Adam Smith's analysis of the relationship between empathy and moral judgment which is compatible with studies of the functioning of the mirror neuron system studied by Rizzolatti et al. (1996), Iacoboni et al. (1999), Rizzolatti et al. (2004), Iacoboni (2008) and others. Studies of mentalization by Amodio-Frith (2006), Singer (2006), Keysers-Gazzola (2007), etc. support these discussions of empathy and social cognition.

sometimes judge the virtue and ethics of our behavior by utilizing others' behavior as a benchmark or yardstick by which to measure our actions. Unfortunately, when we are surrounded by bad apples, persons who exhibit free rider and moral hazard behaviors, we easily forgive our own moral hazard without feeling guilty. A well-known Japanese proverb, "if everyone crossed a red light, there is no reason to feel guilty," criticizes this moral decadence. This type of moral judgment, relative to a bad apple, is different from Smith's reference point for moral judgment. Smith's reference point is produced by the voice of conscience. However, a bad apple reference point results from the empathy produced by observing others' behavior and provides narrow and shortsighted incentives to ignore the voice of conscience.

The image of "the impartial spectator" is idealized and is only one of many methods of moral judgment. We actually utilize various types of moral judgment in social contexts of our daily life. Positive analysis asks when and where we utilize Smith's moral judgment. We designed an experiment to compare the effect of a bad apple with the effect of the voice of conscience on moral judgment. By changing the experimental conditions, we examined which of these exerted stronger effects on the subjects' moral judgment.

Figure 1 illustrates the aim of our experiment. Moral judgment is potentially affected not only by the bad apple but also by the voice of conscience of the impartial spectator. The figure includes Agent X, Bad Apple Y and Impartial Spectator Z. Relation (1) indicates that Agent X's judgment is affected by the reflected passion produced by empathizing with the moral hazard behavior of Bad Apple Y. Relation (2) indicates that Agent X's judgment is affected by empathizing with the voice of conscience of Impartial Spectator Z. Adam Smith claimed that empathizing with a fictive impartial spectator allows for cool and far-sighted judgments that can dominate reflected passion with real persons. However, such cool and far-sighted judgment is not always possible because sometimes the effect of Relation (1) is stronger than the effect of Relation (2). We demonstrated the experimental conditions under which the effect of Impartial Spectator Z is stronger than the effect of Bad Apple Y.

Figure 1 Moral Judgment is affected by the Bad Apple and by the voice of conscience of the Impartial Spectator

We utilized brain decoding in the first trial of its kind, a neuroeconomic study of Adam Smith's paradigm. Brain decoding is a powerful brain reading method that overcomes the limits of questionnaires. While questionnaires can investigate the characteristics of self-conscious decision-making, brain decoding can directly analyze subjects' neural activity to interpret both conscious and unconscious activities. We classified neural activity into two groups that represented different mental states to examine whether the effect of the impartial spectator dominated the effect of the bad apple on moral judgment. The introduction of brain reading extends traditional studies of the history of economic thought (especially of Adam Smith's work) and opens the door to new research in the neuroscience concerning moral judgment.

2. Methods

2.1 Subjects and Tools for Brain Decoding

Eighteen healthy right-handed subjects (nine males; nine females) aged 20 -23 years played the experimental games. All subjects were students at Aoyama Gakuin University. While each subject played the games, we obtained the necessary brain decoding data at 20 different times to provide enough data to execute brain decoding 360 times. The subjects were not allowed to eat for two hours before the experiment to ensure clear neural reactions to the experimental tasks. Before beginning the experiment, the experimental procedures, safety and security information and procedure for obtaining payment for participation were explained to the subjects. Informed consent

was obtained from the subjects. Our experimental plans and procedures were approved by the Research Ethics Committee of Aoyama Gakuin University, Tokyo, Japan.

As illustrated in Figure 2, we utilized a functional near-infrared spectroscopy (fNIRS), a simpler and more convenient tool for examining brain activity than the more widely utilized functional magnetic resonance imaging (fMRI). The use of fNIRS results in minimal stress to the subjects. We utilized the Spectratech OEG-SpO2 model (updated from the OEG-16 model, sampling rate 6.10Hz, manufactured by Spectratech Inc., Tokyo) for fNIRS, which is based on a modified Beer-Lambert law, to scan the frontal cortex of the brain. This fNIRS equipment utilizes small, lightweight, 16-channel digital sensors on a headband to obtain event data through a dynamic high-sensitivity optical signal that reflects how in vivo hemoglobin combines with oxygen in blood vessels with high or low cortical activation. fNIRS provides three types of event-related neural data: changes in oxyhemoglobin (ΔCoxyHb), changes in de-oxyhemoglobin (ΔCdeoxyHb) and aggregate changes in the two types of hemoglobin (ΔCoxyHb + ΔCdeoxyHb). We utilize the changes in oxyhemoglobin for brain decoding. Strangman et al. (2002) found a strong correlation between fMRI variables and fNIRS measures, and the oxyhemoglobin data provided the strongest correlation. Therefore, utilizing the oxyhemoglobin data will produce results for fNIRS brain decoding that correspond to those of fMRI studies. We claim that this method enables us to perform efficient and low-stress experiments in brain decoding.

Figure 2 fNIRS Multi-channel Digital Sensors on a Headband

The 16-channel digital sensors were fixed on the frontal cortex by the headband during the experiment. After each subject completed the experiment, the location of each sensor was measured utilizing 3D positioning with a digital camera (Nikon D5100) and NIRS-SPM software to allow statistical analysis of the fNIRS signals and to confirm that the channels were properly located on the frontal cortex of the brain. Figure 3 illustrates the locations of the sensors for the first subject mapped onto a canonical brain optimized for NIRS analysis. We obtained event-related, high-sensitivity optical signals from these channels.

Figure 3 Locations of the 16 fNIRS Channels in the First Subject, Mapped onto a Canonical Brain

2.2 Experimental Tasks

We presented the subjects with the tasks to be executed via a computer monitor. We obtained neural data while the subjects were considering moral judgment problems. As Figure 4 illustrates, our experiment was composed of four parts, Games A, B, C and C+.

Figure 4 The Experiment was Composed of Four Games A, B, C and C+

Games A and B were preliminary games that produced neural data to be utilized in learning and training the neural network architecture to accurately recognize the typical pattern of neural activity in Games A and B. Figure 5 illustrates the concepts of learning and training the neural network architecture. Utilizing “nntraintool” in the Neural Network Tool Box, we defined a hidden layer 10 and an output layer 2. The input was the neural data obtained in Games A and B. The output was the vector (1,0) when the input was the data from Game A and the vector (0,1) when the input was the data from Game B. The neural network architecture was obtained by training with the input data to determine the typical patterns of neural activity for either the vector (1,0) or (0,1). This is a so-called “supervised learning method.”

Figure 5 Learning the Neural Network Architecture Using the Neural Activity Data Obtained in Games A and B

In Game A, the subjects were asked to consider a moral problem after viewing an image of the impartial spectator. The bad apple was not displayed on the computer monitor in Game A. We expected the subjects to make their moral judgment in light of the voice of conscious of the impartial spectator. In Game B, the subjects were asked to consider a moral problem after observing an image of the bad apple. The impartial spectator was not displayed on the computer monitor. We expected that subjects in Game B would make judgments affected by the image of the bad apple. These data and the Neural Network Tool Box allowed us to map the neural network architecture and accurately recognize the typical patterns of neural activity in Games A and B.

Games C and C+ were the core games in which we obtained the neural data that were compared with the typical patterns utilizing a pattern recognition method. In these games, the subjects could freely consider moral problems after viewing images of both impartial spectator and the bad apple. We expected the subjects to select either the impartial spectator or the bad apple as a more effective cognitive frame for their moral judgments. Games C and C+ presented the subjects with the same experimental task, but in Game C+, the subjects executed the task following a request for additional effort. The subjects were asked to memorize seven random numbers before Game C+, for example 3825149. After the game, the subjects were asked to write down these numbers. We expected that this request would provide a cognitive burden sufficient to produce different results from those observed in game C.

We utilized the mapped neural network architecture to judge which factor, either the image of the impartial spectator or the image of the bad apple, was actually utilized when the subjects were asked to make moral judgments. The brain decoding compared the data obtained in Games C and C+ with the two typical neural patterns previously identified in Games A and B utilizing pattern recognition to obtain a rate of matching. If the rate of matching for the neural pattern of Game A was larger than that of Game B, the subject utilized the cognitive frame of the impartial spectator to make free moral judgments.

Games A, B, C and C+ were composed of short tasks. These short tasks were repeated 3 times in each game. Figure 6 illustrates the short tasks. We explain the short tasks for Games A and B. The first screen displayed a white cross on a black ground to indicate that the subjects should begin the game in the state of relaxation. After 10 seconds, the second stage of the games began. First, the subjects were presented with the moral judgment problem on the screen, which requested that they would decide whether to charge their cell phones in an empty classroom. The next screen of the second stage in Game A displayed an image of the impartial spectator who asked us to refrain from moral hazard, while the screen in Game B displayed a bad apple who was a free rider. The bad apple said, “Let’s sneak some cakes despite prohibitions and warnings,” a proposition that could lead to moral hazard. The screens in the second stage alternated every second. After 8 seconds, the third stage began. In the third stage, the subjects were again presented with the moral judgment problem of whether to charge their cell phones in the classroom. The subjects reflected on the moral judgment problem for 7 seconds, and neural activity data were obtained by fNIRS during this consideration period. Then, the final stage began. The screen displayed the message, “Push the First Button to Confirm Your Consideration of the Moral Judgment Problem.” Pushing the button prevented subjects from getting tired of the games and maintained attention throughout the games.

Figure 6 The Short Tasks Executed in Games A, B, C and C+ (repeated 3 times)

Figure 6 also illustrates the short task assigned in Games C and C+, which presented the subjects with the same experimental task. In the task, the subjects made their moral judgment freely. The first screen displayed a white cross on a black ground to begin the game with relaxation. After 10 seconds, the second stage commenced. The first screen of the second stage displayed the moral judgment problem of whether the subjects would charge their cell phones in the classroom. The second screen of the second stage displayed the image of the impartial spectator, and the third screen displayed the bad apple. The three screens for the second stage alternated every second. The

tasks in Games C and C+ differed from the tasks in Games A and B only in the three types of screens displayed in the second stage. We expected that the subjects would be simultaneously affected by the images of impartial spectator and the bad apple when they judged the moral problem. After 12 seconds, the third stage began. The subjects were given 7 seconds to decide whether they would charge their cell phones. We obtained the subjects' neural activity data utilizing fNIRS during their consideration period. The final screen requested the subjects to push the button to confirm.

2.3 Random Selection of Neural Data for Brain Decoding

For brain decoding, as Figure 7 illustrates, we randomly selected the data obtained by the fNIRS during the experiment in two steps. First, we randomly selected a sample of 40 neural data points per subject from Games A and B to determine the neural network architecture and establish the typical neural patterns. Next, we randomly selected a sample of 20 data points per subject from Games C and C+. Each of the 20 data points was sorted into one of the two types of typical neural patterns from the preliminary games by pattern recognition to obtain the rates of matching. If the rate of matching with Game A was larger (or lower) than the rate of matching with Game B, the subject's moral judgment was mainly influenced by the cognitive frame of the impartial spectator (or the bad apple). The random selection was conducted for all 18 subjects. Data existed for 360 cases of brain decoding with pattern recognition.

Figure 7 Random Selection of Neural Data for Brain Decoding

3. Results

We obtain the following four results in this experiment. The implications and interpretations of these results will be considered in detail in the next section.

Result (1): There were remarkable differences between the typical patterns of neural activity in Games A and B. As illustrated in Figure 8, the neural activity level in Game A was higher than that observed in Game B at all 16 channels of the frontal cortex.
Result (2): When the subjects made their moral judgment freely in Game C, brain decoding indicated that the cognitive frame of the impartial spectator had smaller effects on moral judgment than the bad apple frame. Only utilizing the results from Game C, we cannot support the claim that the voice of conscience overpowered the effects of bad apple.
Result (3): However, Game C+ produced changes that deserved our careful consideration. Brain decoding indicated that the effect of the bad apple on moral judgment decreased sharply and that the effect of the impartial spectator increased slightly. The subjects' moral judgment was affected by the cognitive burden of memorizing seven numbers before playing the game.
Result (4): We examined whether the subjects could correctly write the seven random numbers after finishing Game C+. We found that each subject was able to recall the numbers without mistakes. The subjects were expected to execute the game after carefully memorizing the seven numbers.

Figure 8 The Average Difference in the Change in Oxyhemoglobin in the Frontal Cortex During Games A and B

Result (1)

We mapped the neural network architecture to recognize each subject's typical patterns of neural activity in Games A and B accurately utilizing the algorithms and progress stop conditions of the Neural Network Tool Box listed in Table 1(a). Table 1(b) lists the seed numbers that were

utilized to generate a random number sequence for neural network weights initialization and to partition the initial data into a training set for learning and a validation set. Seed numbers were determined to maximize the performance in the mapping of neural network architecture. The numbers enable reproduction of our analytical results when the same numbers are utilized with the same experimental data.

The typical patterns of neural activity in Games A and B were differed sharply. Figure 8 illustrates the average neural activity of the 18 subjects during Games A and B in terms of the change in oxyhemoglobin (ΔCoxyHb) at each of the 16 channels located on the frontal cortex. The black bars represent the average pattern of neural activity in Game A, and the grey bars represent the average pattern in Game B. The change in oxyhemoglobin was measured in the standard unit, mMmm (millimole x millimeter). As noted in the section 2.1, Strangman et al. (2002) identify a strong correlation between fMRI variables and fNIRS measures, and oxyhemoglobin data provide the strongest correlation between these measures. Figure 9 indicates that neural activity during Game A was higher than that observed during Game B. This pattern implies that the cognitive frame of the impartial spectator required a higher level of neural activity than the frame of the bad apple to make moral judgments.

Table 1(a) The Algorithms and Process Stop Conditions for Determining the Neural Network Architecture Using the Neural Network Tool Box (nntraintool)

Table 1(b) The Seed Numbers Used to Begin the Sequence of Random Numbers for Learning with the Neural Network Tool Box for Each Subject

Result (2)

We executed the brain decoding by matching the neural data produced during Game C with the typical patterns of neural activity observed in Games A and B. For each subject, the 20 neural data points from Game C were matched with his/her typical neural patterns utilizing pattern recognition to obtain a rate of matching. Figure 9(a) illustrates the results from our 360 (= 20 x 18) brain decoded values for Game C. The horizontal axis of the diagram measures the rate of matching that corresponds to the probability of utilizing the image of the voice of conscience to make moral judgments. The vertical axis measures the rate of matching that implies the probability of utilizing the image of the bad apple to make moral judgments. These two rates of matching were calculated utilizing the Neural Network Tool Box. The points in the lower right section of Figure 9(a) possess larger values along the horizontal axis than along the vertical axis; these points can be considered cases in which the subjects relied on the image of the voice of conscience. Conversely, the points located in the upper left section of the figure can be considered cases in which the subjects mainly relied on the image of the bad apple to make moral judgments. In cases in which the matching rates were located in the lower right section or upper left section, brain decoding was able to classify the neural activity data into the two groups clearly. Figure 9(a) illustrates, however, that some cases could not be clearly sorted into these two groups. These observations were not produced by a technological failure of the brain decoding, but rather reflected the usage of other cognitive capacities, such as physiological reasons.

Figure 9(a) The Scatter Diagram Obtained by Brain Decoding of Neural Data in Game C

We calculated the average value of the matching rates, which are illustrated in Figure 9(a). The average matching rate for the voice of conscience was 0.2472, whereas the average matching rate for the bad apple was 0.7487. The average matching rate for the bad apple was larger than that for the impartial spectator by 0.5015. This result implies that subjects primarily relied on the image of the bad apple during Game C and that the cognitive frame of the impartial spectator produced smaller effects on their moral judgments than the bad apple. We cannot support the claim that the voice of conscience can defeat the effects of a bad apple only utilizing the results from Game C.

Result (3)

In Game C+, the subjects were asked to memorize seven random numbers before playing the game. After the game, they were asked to write down these numbers. We expected that the cognitive burden of memorizing these random numbers would produce different results in Game C+ from those observed in Game C. Figure 9(b) illustrates the results in a scatter diagram produced by the 360 brain decoding values for Game C+. To indicate the difference between the results of Games C and C+, we compared the average value of the matching rates in Game C+ with that in Game C. In Game C+, the average matching rate with the usage of the voice of conscience increased from 0.2472 to 0.2940, and the average matching rate with the bad apple decreased from 0.7487 to 0.5828. The difference between the average matching rates of the bad apple and the impartial spectator decreased sharply from 0.5015 to 0.2888. The change in the difference between the average matching rates was statistically significant ($p < 0.01$). We claim that the voice of conscience can defeat the effects of a bad apple under conditions with cognitive burdens. Figure 10 illustrates the change in the average matching rates. This figure indicates that the effect of the bad apple on moral judgment sharply decreased and that the effect of the impartial spectator slightly increased.

Figure 9(b) The Scatter Diagram Obtained by Brain Decoding of Neural Data in Game C+

Figure 10 The Change in the Average Matching Rates from Game C to Game C+

Result (4)

To confirm memorization of the seven numbers, subjects were asked to recall these numbers after completing the game. We confirmed that every subject could correctly write his or her seven numbers. The subjects utilized a portion of their cognitive capacity to remember these random numbers. The subjects were expected to execute the experiment despite this restricted capacity.

Discussions

The aim of this study is to examine the possibility that Adam Smith's type of moral judgment can defeat the effect of bad apples when making moral judgments. We reconsider the implications of the results that were outlined.

Figure 11 indicates our strategy for examining Adam Smith's type of moral judgment. The figure illustrates a representative state of cognitive capacity that was utilized for moral judgment. Area IS represents an area of cognitive capacity where the image of an impartial spectator was utilized, whereas Area BA is an area of capacity where the image of a bad apple was utilized. Area O represents an area in which other cognitive frames such as the subjects' physiological reasons were utilized. The dual process approach reflects the condition that agents simultaneously hold two or more frames that sometimes compete and sometimes cooperate.[2] This condition is illustrated by the three mental states and the cognitive capacities noted in Figure 11.

Figure 11 Restricted Cognitive Capacities and Differential Usage of Frames

---

[2] The dual process approach has been used in social and cognitive psychology to explain how actions change, for example, in the conscious and unconscious mind. Economics is too conservative to introduce such an innovative psychological approach. Economic studies have obstinately assumed the homo economicus with a rational mind to maximize his utility or profits. The recent studies by Kahneman (2003,2011) depart from this conservative tradition in economics by introducing the dual process approach to formulate behavioral economic models.

Figure 11 illustrates the change in brain decoding observed during Game C+. Result (2) indicates that the subjects relied on the image of the bad apple in Game C, and Area BA is larger than Areas IS and O. However, Result (3) suggests remarkable changes during Game C+. When the subjects were asked to memorize seven numbers, a portion of their cognitive capacity was occupied with memorization. Result (4) confirmed the faithful memorization of the seven random numbers. The decrease in available capacity forced the subjects to economize the capacity used for moral judgment, and they changed their method of moral judgment to one suitable for the restrictive case. As Result (3) of Game C+ suggests, the subjects sharply decreased their use of the image of the bad apple, whereas they increased their use of the impartial spectator.

Why did the subjects sharply decrease their use of the image of the bad apple? Why did the subjects increase their use of the image of the impartial spectator? We argue that the subjects decreased their usage of either the impartial spectator or the bad apple to economize their cognitive capacity during the restrictive game. We expected that the subjects more seriously considered the choice between methods to make moral judgment in the restrictive case, and the subjects were expected to decrease mainly their use of the less valuable method to avoid mistakes. Result (3) indicated that the usage of the bad apple sharply decreased while the usage of the impartial spectator did not decrease. Therefore, we conclude that the subjects considered the image of the bad apple to be less valuable than the image of the impartial spectator.

The above discussion does not allow complete understanding of why the usage of the impartial spectator increased slightly in the restricted case. To answer this question, we must clearly distinguish between a direct change in the choice of moral judgment and an indirect change. The discussion in the previous paragraph addresses only a direct change in moral judgment by the decreasing the available cognitive capacity. We should also consider an indirect effect utilizing the concept of marginal rate of substitution, a standard analytical tool in economics.

As explained above we expected that the image of the impartial spectator would be especially valued by the subjects under conditions of restricted capacity. The marginal rate of substitution between the two images for moral judgment would change in an indifference map that described the subject's preferences. Figure 11 illustrates an x increase in the usage of the impartial spectator with the change in the marginal rate of substitution. This represents the indirect substitution effect that is a second-round change in the choice of the method for moral judgment. However, the indirect change may be overwhelmed by the direct change caused by the decrease in available capacity. Result (3) indicates that the overwhelmedness was not observed in our moral judgment problem. The indirect change was larger than the direct change. The net effect between the direct and indirect changes reflected the increased usage of the impartial spectator even when cognitive capacity was restricted. We claim that Adam Smith's type of moral judgment could overwhelm the effects of a bad apple under conditions of restricted capacity and increased cognitive burden.

We explicitly illustrate our experimental results on an indifference map. An indifference map is a conventional two-dimensional map, and we assume that the area of cognitive capacity O is constant. This simple indifferent map illustrates the essence of the discussion in this study. Figure 12 indicates the individual choice between two methods of moral judgment, the impartial spectator and the bad apple. The horizontal axis represents the level of capacity IS to be utilized with the image of the impartial spectator, and the vertical axis represents the level of capacity BA to be utilized with the image of the bad apple. The budget lines T1 and T2 in Figure 13 indicate the total level of cognitive capacity. Result (1) indicates that the impartial spectator requires a higher level of neural activity than the bad apple. Therefore, T1 and T2 are steeper than the – 45-degree line. When the subjects were asked to memorize seven random numbers, the available cognitive capacity decreased and shifted T1 downward to T2. Figure 12 shows the choice of the method for moral judgment shifted from point E to point F by the downward shift of the line T1 to T2. The use of the bad apple decreased sharply and the use of the impartial spectator increased slightly.

If the indifferent curves I and II were homothetic, the use of these two methods for moral judgment would simultaneously decrease from point E to point F*. The line 0Z indicates the locus of the choice between the two methods in the homothetic case. On the 0Z line, the marginal rate of

substitution between the two frames is constant. However, the observed marginal rate of substitution between the two types of moral judgment was not constant when the available capacity decreased. The marginal rate of substitution between the two images changed. The image of the impartial spectator was more preferred in the restrictive case. The difference between our results and the theoretical result in the homothetic case is illustrated by the indirect effect in a shift from F* to F on the indifferent map. The indirect change in the methods of moral judgment results from the change in the marginal rate of substitution.

The total change from point E to point F can be divided into two types of changes: the direct change from E to F* caused by the decrease from T1 to T2 and the indirect change from F* to F caused by the effect of the marginal rate of substitution. The indifference map is useful to illustrate an overall change in methods caused by the decrease in available capacity.

Figure 12 An Illustration of the Experimental Results using an Indifference Map, a Standard Tool in Economics

5. Concluding Remarks

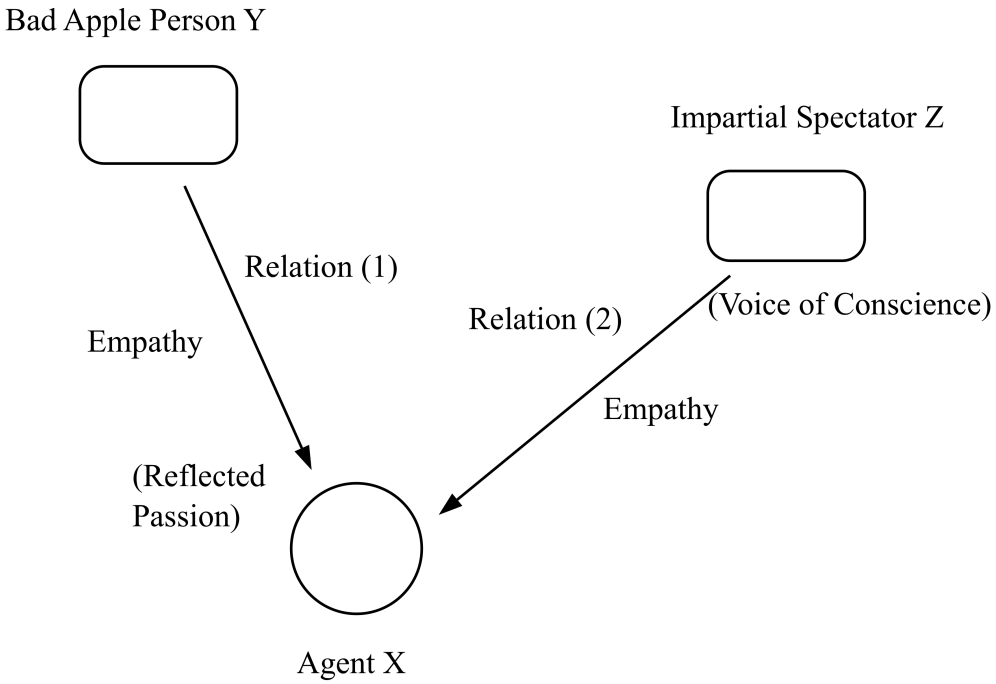
The results of our brain decoding are as follows. The cognitive frame of the impartial spectator required a higher level of neural activity to make moral judgments than the frame of the bad apple. Therefore, when subjects freely made their moral judgment, the impartial spectator was not frequently used for moral judgments. However, when the cognitive capacity was occupied for other tasks, and when the moral judgment required a smaller portion of cognitive capacity, the results changed sufficiently to deserve our careful consideration. The subjects sharply decreased their use of the bad apple, whereas they slightly increased their use of the impartial spectator. This change might seem to be an anomaly that is difficult to explain. This change was not consistent with our intuitive expectations that the method requiring a higher level of neural capacity would be less frequently used in the restricted capacity case. However, our discussion demonstrates that the change in the methods of moral judgments was not anomalous but indicates the greater effectiveness of the impartial spectator. Each method of moral judgment was valued by the subjects by the intrinsic effectiveness at avoid mistakes of judgment. The less effective method should be used only in easy cases where the cognitive capacity was not restricted and where the subjects were off guard without seriously considering the choice of the method of moral judgment. We expect that the image of the impartial spectator was rated as more effective and valuable than the image of the bad apple in the restrictive capacity case. Therefore, to economize on cognitive capacity in the restricted case, the subjects sharply decreased their use of the less effective image, whereas they did not decrease the use of the effective image. To explain this change more completely, we used an indifference map, a standard tool in economics. The change in the method of moral judgment was divided into a direct change and an indirect change, and our experimental results suggest that the indirect change was stronger than the direct effect. The indifference map helps us to explain the experimental result of sharply decreased use of the image of the bad apple and slightly increased use of the impartial spectator in the restricted capacity case.[3]

---

[3] In this study, we present an experimental task with a small cognitive burden that requests the memorization of seven random numbers. In the case of the small and appropriate cognitive burden, we demonstrate the dominance of Adam Smith's type of moral reasoning over the bad apple. However, if the cognitive burden were too heavy for the subjects to cope with, the subjects are expected to produce different behaviors from those observed in our study. An alternative analysis would be required to consider the possibility of irrational herd behavior produced by extraordinary stress. Lux (1995), Lynch (2000) and Shiller (2000) have explained irrational heard behavior in financial markets. A brain decoding study on herd behavior caused by extraordinary stress will be considered in the future. We argued for the necessity of the neuroeconomic study of herd behavior in a preliminary study (Nakagome et al. (2012)).

We claim that the cognitive frame of the impartial spectator can defeat the frame of a bad apple, and the dominance of the impartial spectator can be realized under conditions of restricted cognitive capacity. Restricted capacity occurs in our complex and fast-paced society, which requires simultaneously performing many tasks. Therefore, the implications of our study appear to be valuable. In addition, our claim supports the old proverb, “an idle brain is the devil’s workshop,” and strengthens the effectiveness of Adam Smith’s approach to moral reasoning.

Figure 1 Moral Judgment is Affected by the Bad Apple and by the Voice of Conscience of the Impartial Spectator*



* Agent X is a person who is making a moral judgment. Bad Apple Y is a free rider with moral hazard behavior. Impartial Spectator Z is the man in the breast who provides the voice of conscience.

Figure 2 fNIRS Multi-channel Digital Sensors on a Headband

Figure 3 Locations of the 16 fNIRS Channels on the First Subject, Mapped onto a Canonical Brain

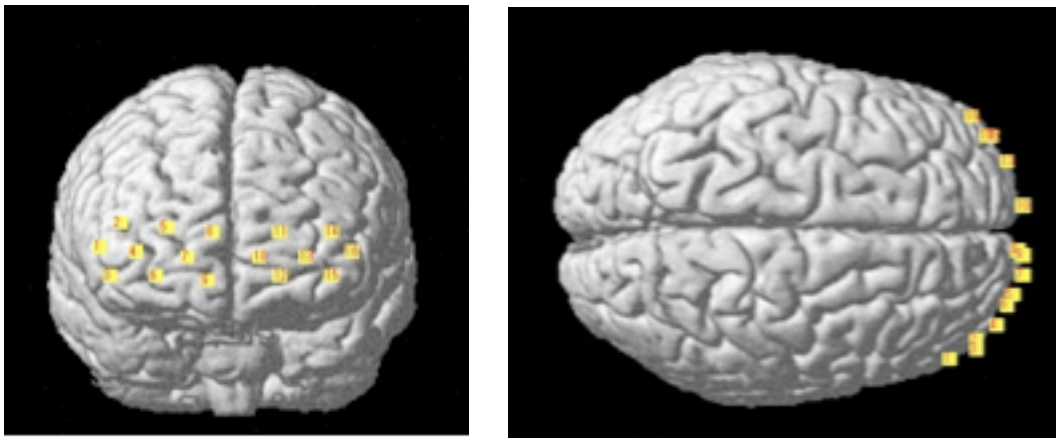
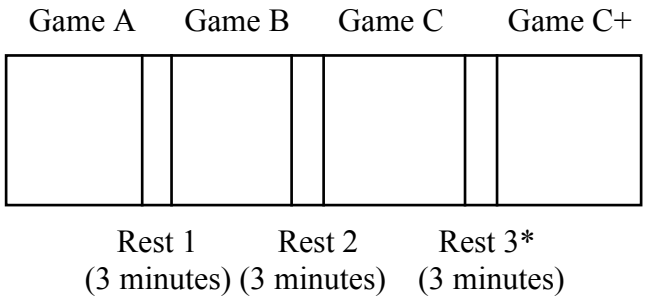


Figure 4 The Experiment was Composed of Four Games A, B, C and C+



* During Rest 3, the subjects were asked to memorize seven random numbers.

Figure 5 Learning the Neural Network Architecture Using the Neural Activity Data Obtained in Games A and B

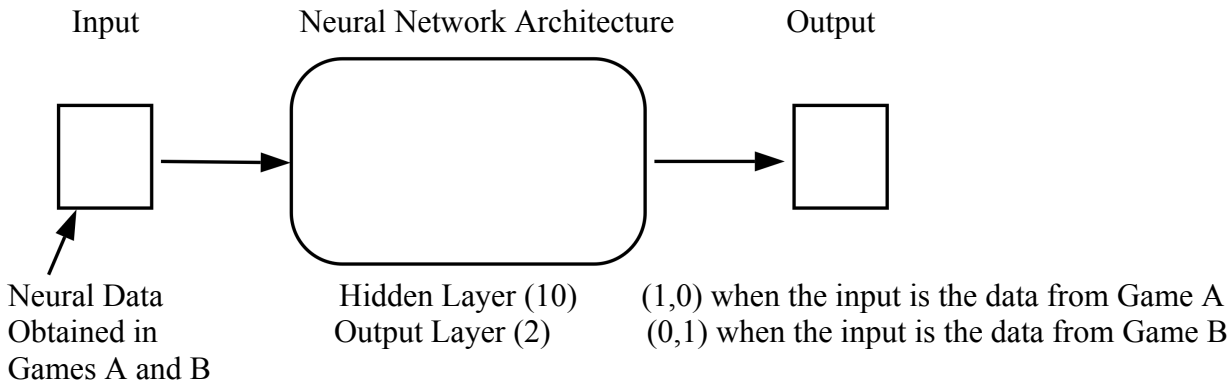
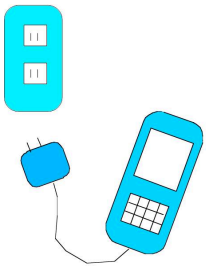


Figure 6 The Short Tasks Executed in Games A, B, C and C+ (repeated 3 times)

(I)The first screen with a white cross on a black ground (10 seconds)
This screen implied that the subjects should begin the game by relaxing.

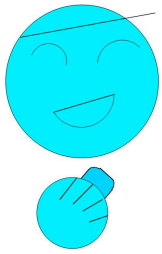
(II)The second stage displayed the following screens (Screens (a) and (b) displayed in Game A, screens (a) and (c) displayed in Game B, screens (a), (b) and (c) displayed in Games C and C+)
The screens alternated every second. (repeated 4 rounds)

(a) Do you charge your cell phone in the empty classroom?



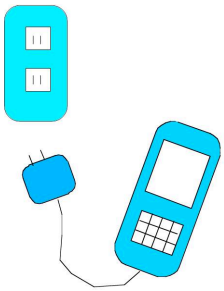
(b) Can you hear the voice of conscience?

(c) Let's sneak some cakes despite prohibitions and warnings.



(III)The third stage presented the moral judgment problem (7 seconds)
Neural activity data were obtained by fNIRS during this stage.

Do you charge your cell phone in the empty classroom?



(IV)The final screen (4 seconds)

"Push the First Button to Confirm Your Consideration of the Moral Judgment Problem."

(V)Rest (4 seconds) and return to the first screen in the second stage.

Figure 7 Random Selection of Neural Data for Brain Decoding

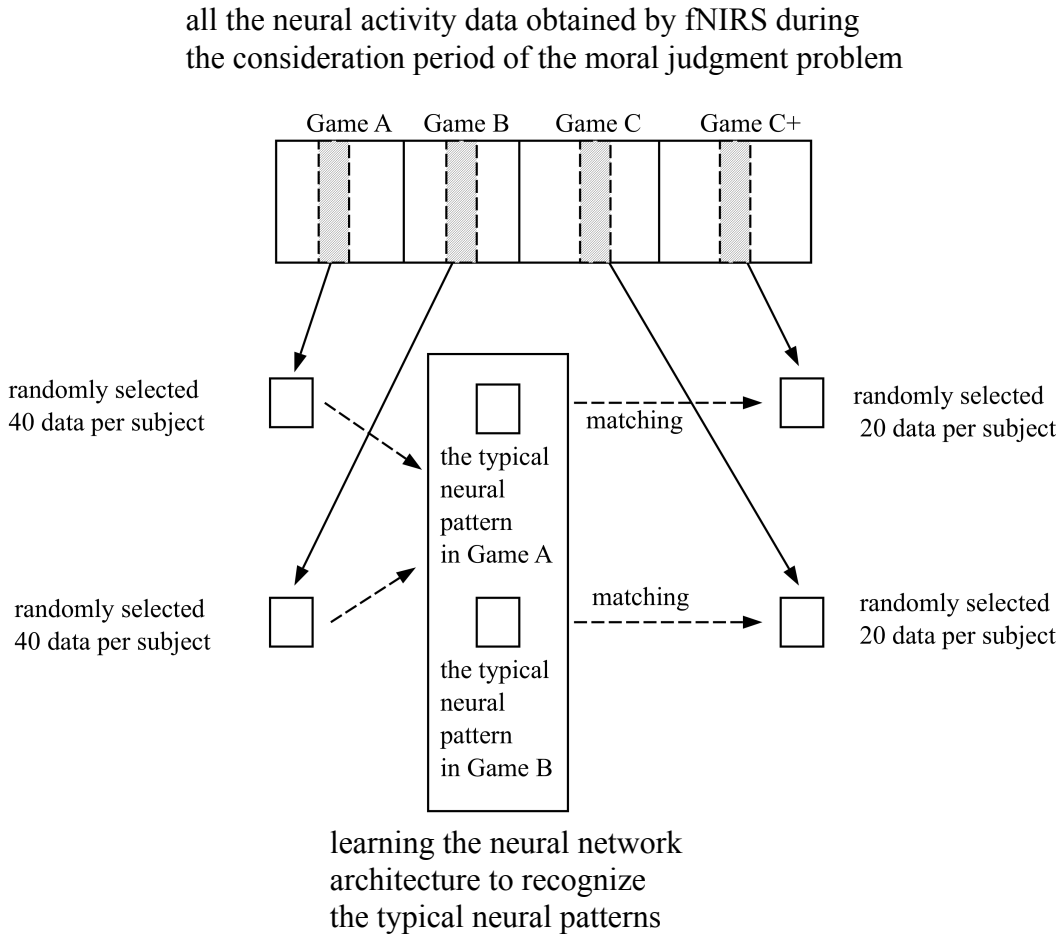
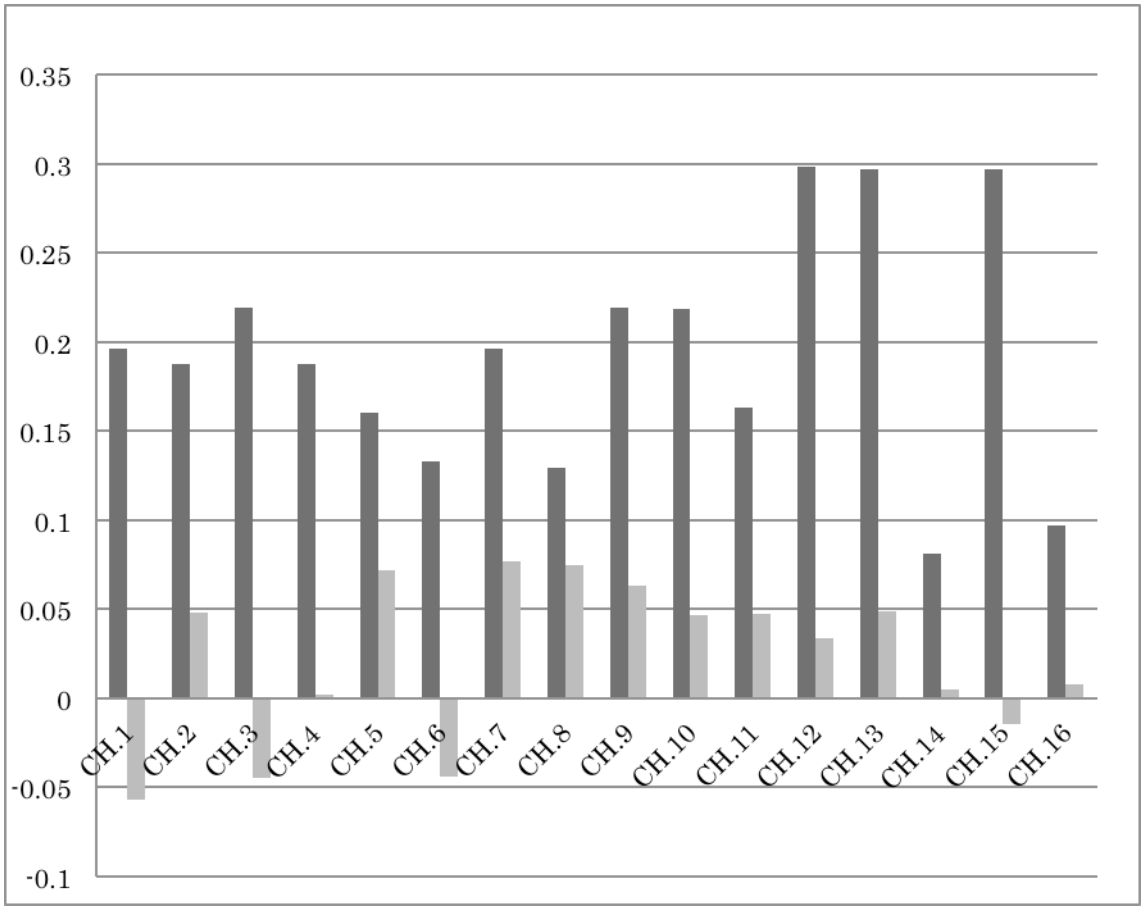


Figure 8 The Average Difference in the Change in Oxyhemoglobin in the Frontal Cortex During Games A and B*



* The typical patterns of neural activities in Games A and B were sharply different. We can illustrate the difference by comparing the average patterns of the 18 subjects in Game A and Game B. Figure 9 illustrates the average difference in the change in oxyhemoglobin (ΔCoxyHb) at each of the 16 channels located on the frontal cortex. In the figure, the black bars indicate the average pattern in Game A, and the grey bars indicate the average pattern in Game B. The change in oxyhemoglobin was measured by using the standard unit, mMmm (millimole x millimeter).

Table 1

Table 1(a) The Algorithms and the Process Stop Conditions for Determining the Neural Network Architecture Using the Neural Network Tool Box (nntraintool)

Algorithms

Data division function: random data division function
Training function: scaled conjugate gradient training function
Performance function: mean squared error performance function
Derivative function: default derivative function

The process stop conditions for learning

Epoch: 1000
Performance: 0.00
Gradient: 1.00 e-10
Validation Checks: 6

Table 1(b) The Seed Numbers Used to Begin the Sequence of Random Numbers for Learning with the Neural Network Tool Box for Each Subject

subject	1	2	3	4	5	6	7	8	9
seed	-15	12	-16	-1	4	10	11	2	-3

subject	10	11	12	13	14	15	16	17	18
seed	9	5	6	-6	-16	4	-47	1	1

Figure 9(a) The Scatter Diagram Obtained by Brain Decoding of Neural Data in Game C

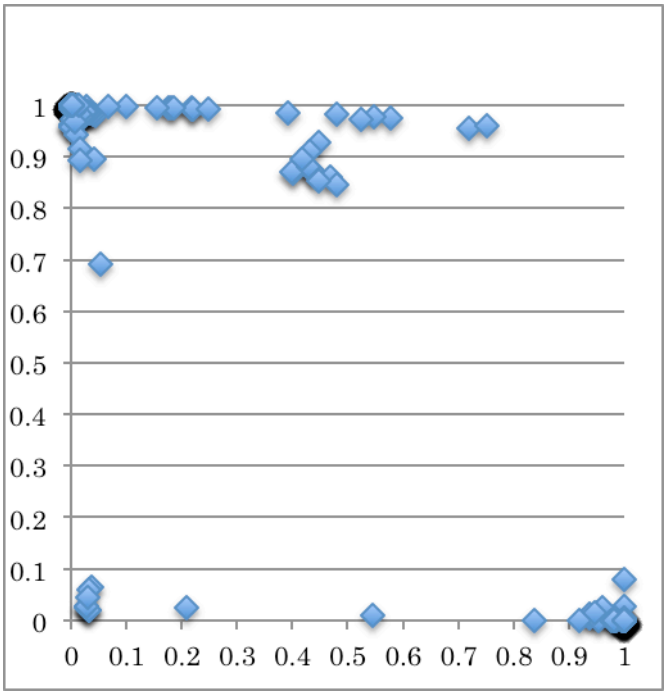


Figure 9(b) The Scatter Diagram Obtained by Brain Decoding of Neural Data in Game C+

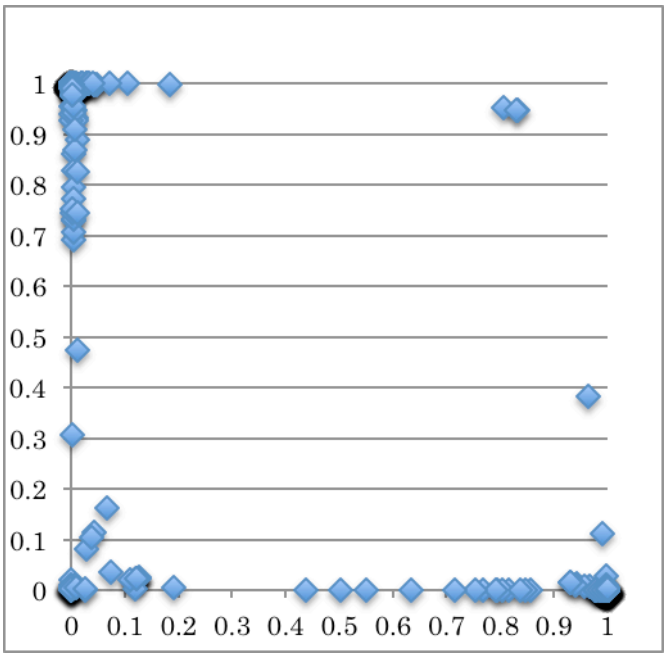
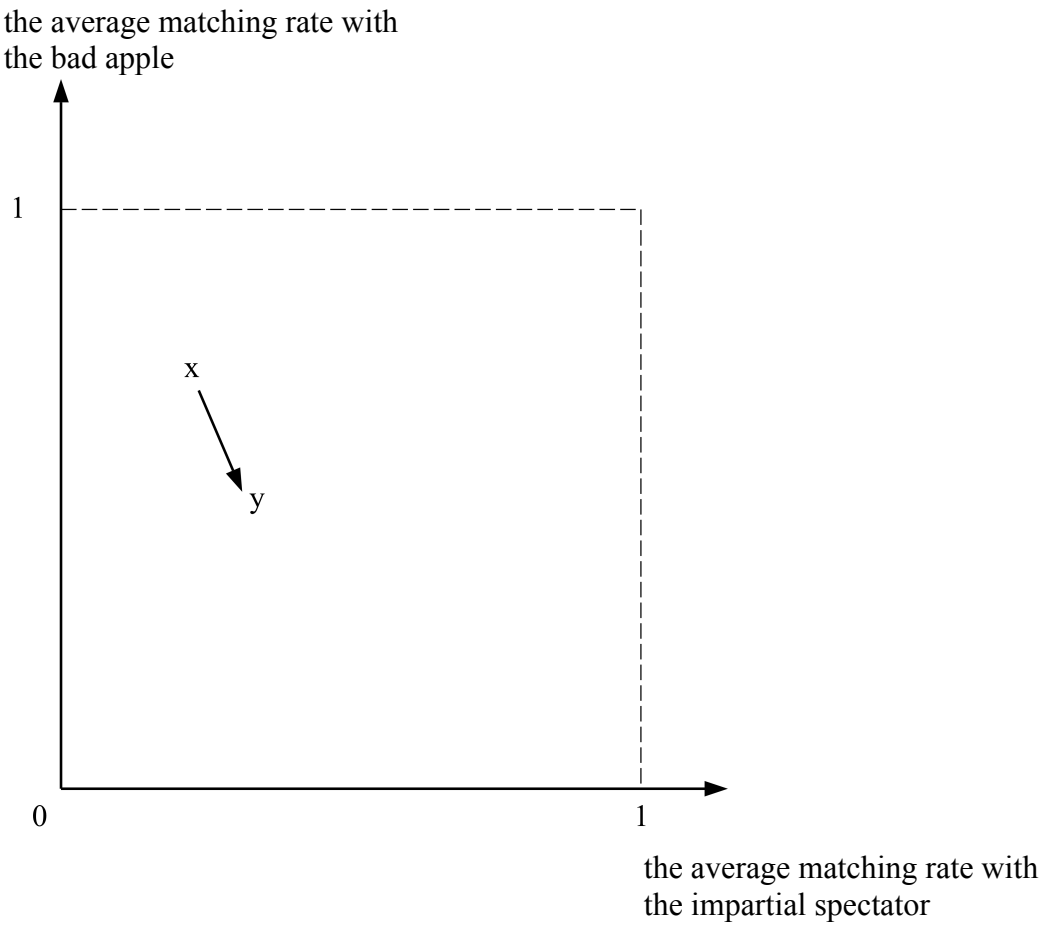
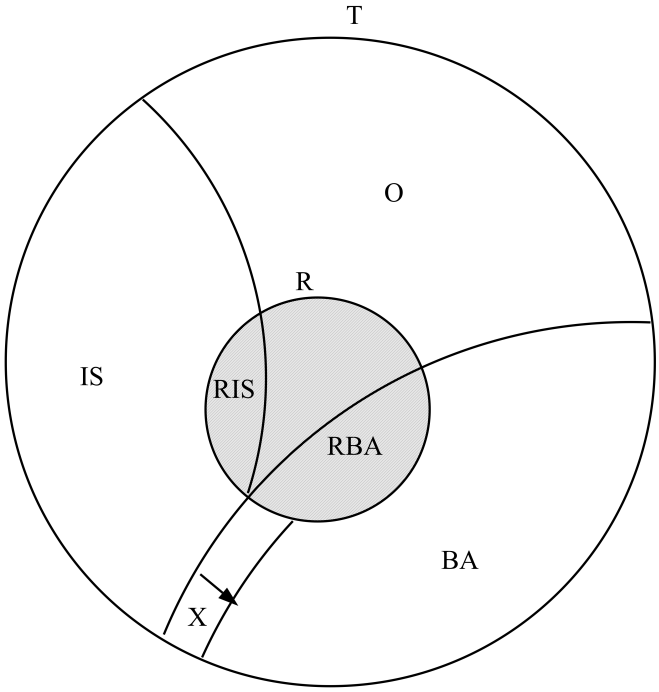


Figure 10 The Change in the Average Matching Rates from Game C to Game C+



x: the average matching rate in Game C
y: the average matching rate in Game C+

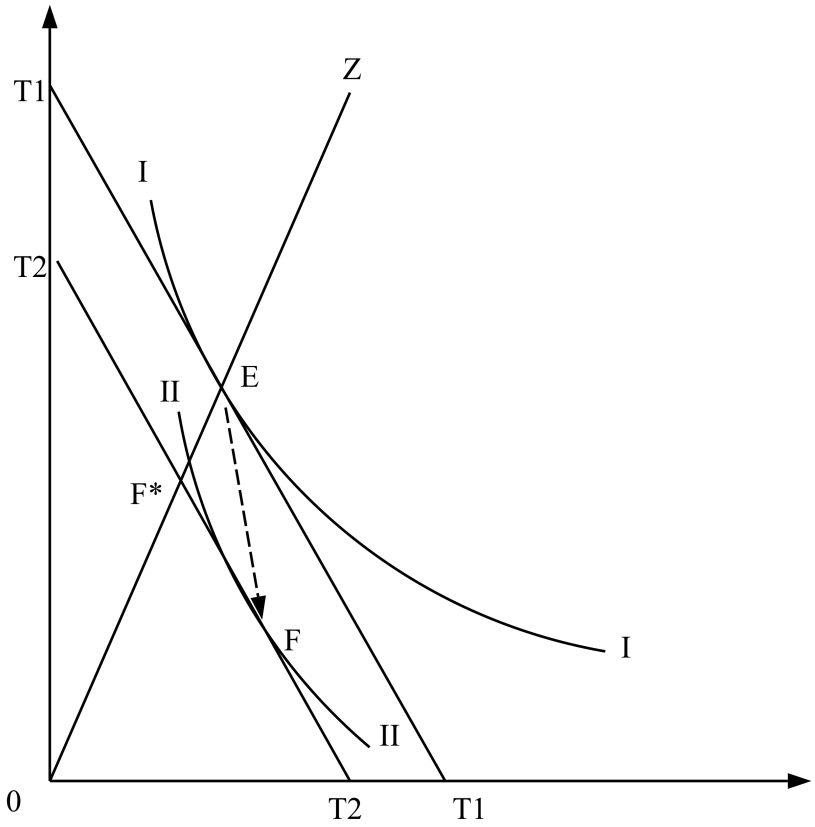
Figure 11 Restricted Cognitive Capacities and Differential Usage of Frames



T … total cognitive capacity
IS … cognitive capacity to be used for the impartial spectator
BA … cognitive capacity to be used for image the bad apple
O …cognitive capacity to be used for reasons such as the subjects' physiological reasons
R … a decrease in cognitive capacity produced by the request for memorizing the seven numbers
RIS … a direct decrease in the capacity IS by the decrease of capacity R
RBA …a direct decrease in the capacity BA by the decrease of capacity R
X … an indirect decrease in the capacity BA by the change in the marginal rate of substitution between the impartial spectator and the bad apple
(= an indirect increase in the capacity IS by the change in the marginal rate of substitution between the impartial spectator and the bad apple)

Figure 12 An Illustration of the Experimental Results using an Indifference Map, a Standard Tool in Economics

the capacity to be used
to imagine the bad apple BA



the capacity to be used to imagine
the impartial spectator IS

References

Amodio, D.M. and Frith, C.D. (2006),"Meeting of Minds: The Medial Frontal Cortex and Social Cognition," *Nature Review Neuroscience*, vol.7, no.4, pp.268-277.
Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C. and Rizzolatti, G. (1999), "Cortical Mechanisms of Human Imitation," *Science,* vol.286, pp.2526-2528.
Iacoboni, M. (2008), *Mirroring People: The New Science of Empathy and How We Connect with Others,* New York: Farrar, Straus and Giroux.
Kahneman, D. (2003), "A Perspective on Judgment and Choice," *American Psychologist,* vol. 58, pp. 697-720.
Kahneman, D. (2011), *Thinking, Fast and Slow,* New York: Farrar, Straus and Giroux.
Keysers, C. and Gazzola, V. (2007), "Integrating Simulation and Theory of Mind: From Self to Social Cognition," *Trends in Cognitive Science,* vol.11, pp.194-196.
Kiesling, L.L. (2012), "Mirror Neuron Research and Adam Smith's Concept of Sympathy: Three Points of Correspondence," *Review of Austrarian Economics,* vol.25, no.4, pp.299-313.
Lux, T. (1995), "Herd Behavior, Bubbles and Crashes," *Economic Journal*, vol.105, pp.881-896.
Lynch, A. (2000), "Thought Contagions in the Stock Market," *Journal of Psychology and Financial Markets*, vol.1, pp.10-23.
Nakagome, M., Maki, K., Fujimori, H. and Ide, H. (2012), "A Brain Decoding Analysis of Framing Effects on the Change in Characteristics of Herd Behavior in Laboratory Financial Markets," Working Paper Series, Institute of Economic Research at Aoyama Gakuin University, no.2, October, 2012.
Rizzolatti, G., Fadiga, L., Fogassi, L. and Gallese, V. (1996), "Premotor Cortex and the Recognition of Motor Actions," *Cognitive Brain Research,* vol.3, pp.131-141.
Rizzolatti, G. and Craighero, l. (2004), "The Mirror-neuron System," *Annual Review of Neuroscience*, vol.27, pp.169-192.
Shiller, R. J. (2000), *Irrational Exuberance,* Princeton: Princeton University Press.
Singer, T. (2006), "The Neural Basis and Ontogeny of Empathy and Mind Reading: Review of Literature and Implications for Future Research," *Neuroscience and Biobehavioral Review,* vol.30, pp.855-863.
Strangman, G., Culver, J.P., Thompson, J.H. and Boas, D.A. (2002), "A Quantitative Comparison of Simultaneous BOLD fMRI and NIRS Recordings During Functional Brain Activation," *Neuro Image,* vol.17, pp.719-731.