Interpersonal Comparison of Utility by Measuring Neural Activity*

Kaosu Matsumori[§1, 2, 3]
Kazuki Iijima[1, 4]
Yukihito Yomogida[1, 5, 6]
Kenji Matsumoto[§1]

Aggregating welfare across individuals is a fundamental problem in our society. There is no rational aggregation procedure without interpersonally comparable utility (Arrow's impossibility theorem). However, scientific methods for interpersonal comparison of utility have thus far not been available. Here, we have developed such a method based on brain signals that correlated with changes in expected utility weighted by subjective probability. The signal was larger for participants with lower household incomes than for those with higher household incomes and their ratio coincided with the estimates by "impartial spectators". These findings suggest that our interpersonal comparison method enables scientifically reasonable welfare aggregation.

JEL classification: D60, D71, D87

* We thank Tatsuya Kameda, Reiko Gotoh, Kou Murayama, and Tsuguo Mogami for discussion. This study was partly supported by a Grant-in-Aid for JSPS Research Fellow Number JP20J01726, JSPS KAKENHI Grant Number JP18H05085 and JP19H04885, and AMED under Grant Number JP18dm0207008 and JP21dm0307001. The study was approved by the ethical committee of Tamagawa University in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.
Email: kaosu.matsumori@gmail.com and matsumot@lab.tamagawa.ac.jp
§ Corresponding authors.
[1] Brain Science Institute, Tamagawa University; Machida, Tokyo, 194-8610, Japan
[2] Department of Social Psychology, University of Tokyo; Bunkyo-ku, Tokyo, 113-8656, Japan
[3] Japan Society for the Promotion of Science; Chiyoda-ku, Tokyo, 102-0083, Japan
[4] National Institute of Mental Health, National Center of Neurology and Psychiatry; Kodaira, Tokyo, 187-8553, Japan
[5] National Institute of Neuroscience, National Center of Neurology and Psychiatry; Kodaira, Tokyo, 187-8502, Japan
[6] Araya, Inc.; Chiyoda-ku, Tokyo, 100-8901 Japan

Main Text:

I. Introduction

I.A. Background in economics

Consider the situation where you decide to give a piece of candy to one of two people. If one person wants it and the other does not, your decision is easy to make. What about when both people say, "I want it"? If you want to give it to the person who will derive more pleasure from it, you have to somehow compare the potential pleasures of the two people due to the candy. You may be able to make a decision if you know the two people well. However, it is extremely difficult to make a decision with enough objectivity to convince a person with a different opinion about these two people's potential enjoyment of the candy. This situation illustrates a serious problem in modern societies that must consider policies of taxation and wealth distribution, because the utilities of large numbers of unknown people, ranging from the very poor to the very rich, have to be compared. This “interpersonal comparison of utility” is pivotal and inevitable for policy evaluation, but it is almost completely ignored in current economics, because interpersonal comparison of utility has been regarded as having no empirical meaning (Robbins 1932, 1938; Arrow 1963; Sen 2018). This stems from the history of concepts of utility.

Bentham's original sense of utility (psychological utility) is psychological pleasure (Bentham 1789). For example, if one person's pleasure when eating an apple is greater than when eating an orange, then her psychological utility is greater for the apple than for the orange. Until the beginning of the 20th century, economists regarded utility in this sense as interpersonally comparable. It is often pointed out that, in everyday life, ordinary people make interpersonal comparisons of utility with relative ease and apparent success (Davidson 2004; Rossi 2014). Most people would think, “If a poor person and a rich person receive the same amount of money, the poor person will be more pleased than the rich person.”

However, psychological utility cannot be directly measured from behavior. Therefore, along with the rise of logical positivism and behaviorism, the new concept of

utility (decision utility), which can be operationally defined by the agent's behavior, emerged as a surrogate for psychological utility. Because agents are usually thought to prefer the option with greater psychological utility, a function u such that u(preferred option) is greater than u(non-preferred option) came to be called the utility function (Kreps 2012; Moscati 2018).

In this concept, for example, if an agent prefers an apple to an orange, she has greater decision utility for an apple than for an orange, and this is measurable using the agent’s choice behavior. Thus, decision utility was considered an ordinal scale because only the relationship "prefer to" was available (ordinal utility) (Fig. 1A).

von Neumann and Morgenstern attempted to turn the concept of utility into an interval scale by considering the history of the development of the temperature scale (Arrow 1963; von Neumann and Morgenstern 2007). When only the relationship "warmer than" was known, temperature was also considered an ordinal scale. Temperature became an interval scale (e.g., Celsius and Fahrenheit) through examinations of the expansion of fluids such as mercury, beginning in the 17th century. The initial validation of these attempts was based on coherence between such scales and human sensation (Chang 2004; Tal 2015). They got the idea that liquids expand with heat on the basis of their most obvious observations in winter cold and summer heat, which justified those temperature scales (Chang 2004). Temperature scales which are based on the expansion of fluids have two free parameters for unit width (α) and zero point (β), and are mathematically unique up to positive linear transformation ($x' = \alpha x + \beta$, $\alpha > 0$). The Celsius and Fahrenheit scales have different unit intervals and zero points, and they are commensurable with a linear transformation. (Subsequent developments in thermodynamics even resulted in a ratio scale for temperature (the Kelvin scale)).

von Neumann and Morgenstern developed a concept of cardinal utility on lotteries as an interval scale (Krantz et al. 1971; von Neumann and Morgenstern 2007) (expected utility) (Fig. 1B, C). Here, assume an agent whose preference order is apple, banana, and orange. In this concept of utility, if an agent is indifferent between a banana and a 50/50

gamble of an apple and an orange, the difference in utility between an apple and a banana is equal to the difference in utility between a banana and an orange (Fig. 1B, C). Thus, they succeeded in expressing utility as interval scale. This idea has been supported by later studies dealing with choice probability, decision time, and neural signals (Luce 1959; Waldner 1972; Stott 2006; Stauffer, Lak, and Schultz 2014; Ratcliff et al. 2016). Moreover, a study has shown that expected utility, weighted by subjective probability, has a linear relationship with psychological utility (Abdellaoui, Barrios, and Wakker 2007), suggesting replacement of psychological utility with the expected utility weighted by subjective probability.

However, expected utility is only based on each individual's choice behavior and only the relative position of the utility for an option between the best and worst options on an arbitrary linear scale is empirically measurable (Fig. 1B, C, E, F). In other words, the information of utility as interval scale is fully preserved even after a linear transformation with the parameters for different unit width (α_i) and for different zero point (β_i) among individuals ($u_i' = \alpha_i u_i + \beta_i$, where i is for individuals). A popular transformation is the expression on a scale from 0 for the worst option to 1 for the best option (0-1 rescaling). This indicates that the decision utilities are only *intra*personally, but not *inter*personally, comparable by definition. Note that this limitation applies to ordinal utility as well.

Thus, after the 1930's, interpersonal comparisons of utility based on lay people's intuition came to be regarded as neither objective nor scientific, and thus have been taboo for nearly a century since the concept of decision utility was popularized (Robbins 1932, 1938). This is a serious problem caused by replacing psychological utility with decision utility. With decision utility alone, the utility for a slice of bread cannot be compared even between a starving person and a billionaire (Sen 2018).

Analyses based on decision utility have shown that a competitive equilibrium based on an ideal market can realize a Pareto efficient state. However, Pareto efficiency is not considered to be a sufficient condition for proper resource allocation because Pareto efficient resource allocation does not exclude highly unequal states. Then, the next question

is “how can we reach an appropriate resource allocation method that satisfies conditions other than Pareto efficiency?” However, the existence of such a method was denied by Kenneth Arrow. He mathematically proved that there is no rational procedure that simultaneously satisfies even some very mild conditions for validity without interpersonally comparable utility (Arrow 1963; Sen 2009, 2018) (Arrow's impossibility theorem). Nevertheless, governments are often compelled to evaluate policies to conduct collective decisions in which interpersonal comparisons of utility are required. Because there is no appropriate method to compare utilities interpersonally, policy evaluations are often conducted by assuming that the utility of a penny is exactly the same at all income levels from the poor to the very well off (Boardman et al. 2018; HM Treasury 2018) (social cost-benefit analysis). This assumption is unrealistic and contrary to intuitions about interpersonal comparisons (Boardman et al. 2018; Sen 2018). If only we could find an appropriate method for interpersonal comparisons of utility, we could deal with the distribution problem properly.

Therefore, it is essential to extend the informational basis of policy evaluation by developing an appropriate interpersonal comparison method, which enables us to escape from Arrow's impossibility. An appropriate method for interpersonal comparison of utility has to satisfy the following two conditions. First, the interpersonal comparison method must rest upon scientific demonstration rather than upon ethical principle (Robbins 1932, 1938). This condition warrants the objectivity and reproducibility of results about interpersonal comparison of utility. Second, the interpersonal comparison method must effectively capture human intuitions about utility and be validated by their coherence, as was done for the fluid expansion-based temperature scale in its initial stages (Chang 2004). Satisfying this condition must serve as a starting point for subsequent scientific inquiry (Chang 2004). One popular idea among both economists and philosophers is that, in order to solve the problem of interpersonal comparisons of utility, we have to look at how ordinary people make such comparisons in everyday life (Harsanyi 1955; Rossi 2014), because it is often pointed out that ordinary people make interpersonal comparisons of utility with relative ease and apparent success (Davidson 2004; Rossi 2014).

In the present paper, we developed a scientific method for interpersonal comparison of utility based on brain-derived signals obtained by functional magnetic resonance imaging (fMRI). In addition, we validated the interpersonal comparison method by comparing utility based on MRI signals with impartial spectators' subjective estimations. Finally, we applied the interpersonal comparison method to an actual distribution problem.

I.B. Background in neuroscience

To construct a method for interpersonal comparison of utility that satisfies the above two conditions, we focused on neural representation of utility, for the following reasons. First, every economic choice behavior that reveals decision utility consists of nothing but voluntary muscle movements generated by signals originating in the brain. Moreover, mental states, including psychological utility, are realized by brain states. Interpersonally comparable utility may be found in the activities of the responsible brain areas.

In the past quarter century, neuroscience studies have uncovered neural circuits that represent subjective values, known as the "value system" (Glimcher and Fehr 2013; Bartra, McGuire, and Kable 2013; Caplin et al. 2010). The value system consists of a set of brain regions, including the medial prefrontal cortices and anterior striatum, that learn the values of choice options by receiving inputs from the midbrain dopaminergic nuclei (Lewis et al. 1988; Gaspar et al. 1989). Neuroscientific studies using rodents, monkeys, and humans have repeatedly shown that damage to the value system or the midbrain dopaminergic nuclei causes impairments in choice behaviors (Butter, McDonald, and Snyder 1969; Cardinal et al. 2001; Frank, Seeberger, and O'Reilly 2004; Camille et al. 2011). In addition, a recent study reported that electrical stimulation to a region (ventral prefrontal cortex) in the value system caused changes in choice behaviors of monkeys (Ballesta et al. 2020). Furthermore, functional neuroimaging studies on human participants have shown that the activities of the value system represent the value of various rewards (e.g., food, money, and social praise) on a single common scale (Izuma, Saito, and Sadato 2008; Levy and Glimcher 2012; Glimcher and Fehr 2013; Bartra, McGuire, and Kable 2013). Moreover, the

activities of the value system are less responsive to food when not hungry (Gottfried, O'Doherty, and Dolan 2003; Cassidy and Tong 2017), and less responsive to monetary rewards when the reward is delayed (McClure et al. 2004; Kable and Glimcher 2007), suggesting that the activities of these regions represent the subjective aspect of the value of rewards. Therefore, utility information is expected to be found in activity in the value system.

One study using monkey subjects showed that neuronal activity in the midbrain dopaminergic nuclei, which project to the value system, encoded a utility difference (Δu), the prediction error of decision utility (i.e., the difference between the utility of an obtained outcome and the expected utility of a lottery) (Stauffer, Lak, and Schultz 2014; Caplin et al. 2010; Schultz, Dayan, and Montague 1997; Sutton and Barto 2018). If we find a utility representation in human brains, it may allow us to objectively measure utility differences between options depending on individual mental states, in contrast to measurement based on choice behavior alone (Fig. 1D). The information of such a utility representation should be preserved after a linear transformation with the parameters for common unit width (α) and for different zero points (β_i) among individuals ($u_i' = \alpha u_i + \beta_i$, where i is for individuals).

II. RESULTS

II.A. Neural representation of utility

To examine neural representation of utility, we identified a representation which correlates with the prediction error of decision utility (Δu), not with that of reward amount, and commonly works irrespective of reward type (i.e., food and money).

In the present study, we engaged 63 participants in two kinds of gambling tasks, a food gambling task and a monetary gambling task. We used brain imaging data from 56 and 60 participants during the food gambling task and the monetary gambling task, respectively, after excluding the data from 7 and 3 participants because of head movements

during the scan, insufficient motivation to obtain the food reward, and a technical problem in data storage (see APPENDIX in detail).

Because utility for moderate amounts of money is approximately linear intrapersonally (Wakker and Deneffe 1996; Abdellaoui, Bleichrodt, and L'Haridon 2008; Booij and Kuilen 2009), we used the food gambling task to extract brain regions that correlate with the prediction error of decision utility (Δu) rather than that of reward amount.

In the food gambling task, participants chose between a sure payoff of food (snack) tickets and a lottery entailing a 50/50 chance of gaining one of two quantities of food tickets (Fig. 2A). We determined each participant's 0-1 scaled utility function for food tickets using the fractile method, with considering probability weighting function (Stauffer, Lak, and Schultz 2014) (see APPENDIX, Supplementary Fig. 1). The slope of the utility function for most of the participants decreased as the amount of food increased (Fig. 2B), indicating that their marginal utilities diminished in the range from 1 to 300 food tickets at least.

Because we included trials that offered a random number of food tickets, but these trials were not used for the fractile method in the food gambling task (see APPENDIX), we could use them to validate the shape of each participant's utility function. The percentage of trials in which participants chose the offer with larger estimated utility was 77.13% on average (SD 10.28%) and the choice behaviors were, after consideration of probability weighing, more accurately predicted by expected utility than expected reward amount ($t(55) = 4.58$, $P < 0.001$ two-sided), suggesting that the utility functions account well for the participants' behaviors.

Next, we analyzed the fMRI data to identify brain regions whose activation correlated with utility. Because we had an a priori hypothesis focusing on the activations of the value system (Glimcher and Fehr 2013; Bartra, McGuire, and Kable 2013), the analysis was performed for the voxels within the explicit anatomical mask that covers the medial prefrontal cortices and anterior striatum. We found that the activity of a brain region consisting of the anterior cingulate cortex and its adjacent ventromedial prefrontal cortex

(ACC/vmPFC) (peaked at [0, 38, -4] MNI coordination), correlated with the prediction error of utility (Δu) (GLM1) (red and yellow areas in Fig. 3A, t(55) = 4.79, FWE corrected *P* = 0.008 one-sided). Importantly, the activity of this region, as well as others, did not correlate with the prediction error of the number of food tickets (GLM2). This indicates that the activity of this region correlates with the prediction error of utility rather than that of reward amount per se.

Next, we examined whether the activity of the ACC/vmPFC region that correlated with utility prediction errors in the food gambling task also correlated with utility prediction errors in the monetary gambling task. In the monetary gambling task, participants chose between a sure amount of money and a lottery for money with a 50% payoff probability in 2 different conditions of blocks: the narrow block, which had a narrow range of reward amounts (¥-200~200) and the wide block, which had a wide range of reward amounts (¥-400~400) (1 USD ≈ 100 JPY (¥)) (Fig. 2C). This allowed us to examine whether brain activity was influenced by different ranges of reward amounts. We compared the percent signal changes (PSC) of the ACC/vmPFC region evoked by the utility prediction errors per ¥1 between the narrow block and the wide block, using not only classical hypothesis testing but also Bayesian hypothesis testing based on Bayes factor (BF). We found moderate evidence indicating an absence of influence by different ranges of reward amounts (t(59) = 0.36, *P* = 0.360 one-sided, BF_{+0} = 0.193, median posterior δ = 0.044 with 95% CI [-0.201, 0.290]) (GLM3) (Supplementary Fig. 2). Therefore, we ran another GLM (GLM4) and identified the voxels representing utility prediction errors in our anatomical mask without discriminating between the narrow and wide blocks (peaked at [-9, 5, -10] and [9, 8, -7] (t(59) = 9.05, t(59) = 7.68) in the striatum and at [-6, 38, -4] (t(59) = 6.57) in the vmPFC) (FWE corrected *P*s < 0.001 one-sided) (blue and yellow areas in Fig. 3A).

The utility representation for use of interpersonal comparison must represent utility irrespective of varying reward type. We sought the brain region that represents the utility prediction errors for both food and money. Most of the ACC/vmPFC region which represent the utility prediction error for food also correlated with the utility prediction error

for money, hereafter called the "utility region" in this paper (yellow area in Fig. 3A). Thus, the "utility region" was correlated with the prediction error of utility rather than that of reward amount irrespective of reward type, and we refer to the PSCs of the utility region as Δu_{neural} in this paper.

II.B. Interpersonal comparison of utility based on neural signals

To realize an interpersonal comparison of utility based on each person's household income, the relationship between the income and Δu_{neural} must be determined. We applied a multiple regression analysis with regressors for income, age, and sex. The results showed that Δu_{neural} was significantly affected by income ($\beta_{income^{-1}} = 0.247 \pm 0.132$ (mean ± SE), $t(56) = 1.874$, $P = 0.033$ one-sided; Fig. 3C), but not the other regressors ($\beta_{age} = -0.100 \pm 0.132$, $t(56) = -0.757$, $P = 0.452$ two-sided; $\beta_{sex} = -0.095 \pm 0.130$, $t(56) = -0.731$, $P = 0.468$ two-sided; $\beta_0 = -0.235 \pm 0.179$, $t(56) = -1.318$, $P = 0.193$ two-sided). This indicated that Δu_{neural} was inversely proportional to income, corresponding to the hypothesis that the larger the income, the smaller the Δu_{neural}.

In the above analysis, we assumed that utility of income can be approximated by a logarithmic function. Since this assumption was arbitrary, we compared Δu_{neural} between the participants whose household incomes were in the lower half (up to 6 million yen) among all participants (the "poor" group) and the participants whose household incomes were in the upper half (the "rich" group) (Fig. 3B), to subsequently consider a distribution method for the MRI experiment participants that does not require any specific utility function (see **Application to a distribution problem**). We found a significant difference in Δu_{neural} between the income groups ($t(58) = 2.39$, $P = 0.010$ one-sided, $BF_{+0} = 5.723$, median posterior $\delta = 0.535$ with 95% CI [0.046, 1.046]) (Fig. 3D), but no difference in subjective probabilities (Supplementary Fig. 3). The posterior means of Δu_{neural} across participants, $E[\mu_{PSCpoor}]$ and $E[\mu_{PSCrich}]$, were $4.624 * 10^{-4}$ and $2.007 * 10^{-4}$, respectively ($E[\mu_{PSCpoor}]/E[\mu_{PSCrich}]$ was 2.304). Thus, we were able to objectively determine an *inter*personally comparable scale of utility difference by using Δu_{neural} (Fig. 1D). The scale

of utility difference would not be interpersonally comparable if it was based on choice behavior alone.

II.C. Coincidence of the ratio of Δu_{neural} and impartial spectators' estimation

In order to validate our scientific interpersonal comparison method using Δu_{neural}, we developed a behavioral task to measure impartial spectators' estimates of the utility difference ratio of the poor group and the rich group. An additional 15 participants completed the impartial spectator task. Two participants were excluded from the following analysis, because they did not pass the instructional manipulation check (see APPENDIX). We considered the remaining 13 participants "impartial spectators" for this task, because they were impartial and empathetic to the MRI participants (Smith 1759; Harsanyi 1977) (see APPENDIX). After seeing a histogram showing the household income distribution of the MRI participants for the gambling experiment, the impartial spectator participants gave their intuitive estimates of the amounts of money necessary to please participants in the rich group equivalently to those in the poor group receiving amounts of ¥400, ¥500, and ¥600 (see APPENDIX). Their answers are shown in Fig. 4A. From the results, impartial spectator participants' intuitive estimates of the utility difference ratio of poor group to rich group were calculated (median, 2.000) (Fig. 4B). The utility difference ratio of poor group to rich group based on the impartial spectator participants' intuitive estimates coincided with that based on Δu_{neural} with moderate evidence ($BF_{10} = 0.165$, median posterior $\delta = 0.014$ with 95% CI [-0.456, 0.211]) (Fig. 4B, C, Supplementary Fig. 4). Thus, the interpersonal comparison method based on MRI signals from the brain was coherent to impartial spectators' estimation.

II.D. Application to a distribution problem

We applied the interpersonal comparison method based on Δu_{neural} to a distribution problem from a social planner's perspective (right balloon in FIG. 5). Here, we used a

decision rule that maximizes utilitarian social welfare, because a utilitarian social welfare function is the most common when utility differences are interpersonally comparable (d'Aspremont and Gevers 1977; Roberts 1980; Narens and Skyrms 2020). Suppose the case in which the planner will distribute ¥1,000 to all participants in addition to their usual income (Policy A). Since $E[\mu_{PSCpoor}]$ and $E[\mu_{PSCrich}]$ were $4.624 * 10^{-4}$ and $2.007 * 10^{-4}$, respectively, the expected social welfare increase would be 19.893 *util* (= $(4.624 + 2.007) * 10^{-4} * 1{,}000 * 30$), assuming that 1 util corresponds to a 1% change in the BOLD signal. Next, we considered the case in which the planner distributes a certain amount of money selectively to the poor participants. For the selection process, we assumed a regulation cost should be required. Suppose the case in which the planner will distribute ¥1,500 to every poor participant with the assumption of regulation cost of ¥500 (Policy B). In this case, the expected social welfare increase will be 20.808 util (= $4.624 * 10^{-4} * 1{,}500 * 30$) for Policy B. Because the social welfare increase is higher for Policy B than Policy A, we decided to use Policy B to distribute ¥1,500 to every poor participant instead of ¥1,000 to all participants, and actually carried this out for the participants who agreed to receive additional compensation. Without interpersonal comparison of utility, it cannot be argued that such a greater social welfare increase will result from a policy that is more generous to the poor than the rich.

According to the ratio of posterior means of Δu_{neural} ($E[\mu_{PSCpoor}]/E[\mu_{PSCrich}] = 2.304$) described above, for the problem of choosing between the policy of distributing ¥k to all participants and the policy of distributing ¥(k + ρ) to every poor group participant, the optimal decision rule, for an expected utilitarian social welfare maximizer, was generalized as follows:

$$\begin{cases} \text{giving ¥}k \text{ to all participants} & (\text{if } k/\rho < 2.304) \\ \text{either is fine} & (\text{if } k/\rho = 2.304) \\ \text{giving ¥}(k+\rho) \text{to the poor group} & (\text{otherwise}). \end{cases}$$

This decision rule tells us that the social welfare increase resulting from distributing ¥1,200 to every poor participant with the assumption of regulation cost, ¥800, would be less than giving ¥1,000 to all participants (left balloon in Fig. 5).

III. Discussion

In order to develop a scientific method for interpersonal comparison of utility, we measured functional MR signals from participants with different household incomes while they were performing food and monetary gambling tasks. We obtained the following findings. First, we found that the activity in an ACC/vmPFC region was correlated with utility difference irrespective of reward type. Second, we showed that the signal strength from this brain region (Δu_{neural}) was inversely proportional to household income across participants. Third, we found that the ratio of Δu_{neural} from lower-income participants to those of higher-income participants coincided with estimates of their psychological pleasure by impartial spectators. Finally, on the basis of the ratio of Δu_{neural}, we actually solved a distribution problem.

We succeeded in developing a scientific method for interpersonal comparison of utility based on neural signals. Scientific measurement methods are necessary for the interpersonal comparison of utility, because interpersonal comparison methods based solely on impartial spectators' estimates are unlikely to work well when judgments are polluted by systematic biases (Müller-Lyer 1889; Chang 2004). Without scientific measurement methods, we cannot determine whether systematic biases are operating. Furthermore, non-scientific interpersonal comparison methods based on empathy may work in a small group but not in a large society, because empathy is too information-intensive to reach people living in different cultural contexts (Elster and Roemer 1991; Bloom 2016). Our new interpersonal comparison method based on a scientific measurement (human brain imaging) may serve to extend empathy and be applicable to a larger society.

Our finding that the activity of ACC/vmPFC correlated with the prediction error of utility rather than that of objective amount of reward is consistent with many previous studies that indicate ACC/vmPFC encodes subjective value (Matsumoto et al. 2007; Glimcher and Fehr 2013; Bartra, McGuire, and Kable 2013). However, in contrast to these previous studies, the striatum was not extracted in our study. This may be due to

differences between our methods and theirs. We extracted brain activity that correlated, during the food gambling task in particular, with utility by using a rigorous method to determine the utility curve (i.e., the fractile method). Therefore, it is possible that the ACC/vmPFC represents the prediction error of utility but the striatum does not. This interpretation is consistent with the finding that striatum activation was apparently correlated with the prediction error of utility in the monetary gambling task; however, it was not distinguishable from the prediction error of reward amount because of their linearity (Seymour et al. 2007). Moreover, some previous studies also suggested that the ACC/vmPFC represents more subjective information than the striatum (Aoki et al. 2014; Murayama et al. 2015).

A previous study showed that midbrain dopamine neurons encode the prediction error of utility, not the reward amount, using the fractile method in monkeys (Stauffer, Lak, and Schultz 2014). Because the ACC/vmPFC receives inputs from the dopamine neurons in both monkeys and humans (Lewis et al. 1988; Gaspar et al. 1989), the dopaminergic inputs may supply the information about prediction error of utility in ACC/vmPFC in humans.

We identified the “utility region” in ACC/vmPFC by extracting the brain region that represented utility prediction errors in both food and monetary gambling tasks. This is consistent with previous neuroimaging studies that reported that specific brain regions including ACC/vmPFC represent the value of various rewards (e.g., food, money, and social praise) on a single common scale (Izuma, Saito, and Sadato 2008; Levy and Glimcher 2012; Glimcher and Fehr 2013; Bartra, McGuire, and Kable 2013) like utility in economics.

A recent study argued that inter-individual (or inter-group) differences in fMRI studies are difficult to interpret if the variances of explanatory variables were different between groups (Lebreton et al. 2019). However, that is not the case in the present study. In the monetary gambling task, the variances of the explanatory variable (i.e., prediction error of monetary amount) were not different between poor and rich groups, because all offers

were common among participants, and subjective probabilities were not different between the poor and rich groups (Supplementary Fig. 3).

Our interpersonal comparison method can be applied to other kinds of goods by following certain procedures, although we used only monetary and food rewards in our experiment. First, willingness to pay for other goods can be measured to find out how much the good is worth to the person, and then willingness to pay can be converted to the interpersonally comparable utility of other goods by multiplying the coefficients of the utility measured using our proposed method.

We did not find significant modulation of ACC/vmPFC activity by the two reward-range (narrow and wide) conditions in the monetary gambling task. This result indicated that the ACC/vmPFC activity was robust to different ranges of reward amounts in our experiment. However, we do not argue that this activity is totally free from the normalization that previous studies have suggested (Kobayashi, Pinto de Carvalho, and Schultz 2010; Louie, Glimcher, and Webb 2015; Burke et al. 2016). The apparent lack of the influence by the changes between conditions might depend on the experimental design specific to the present study. The task consisted of 3 sessions of 40 trials (20 successive trials each of the narrow and wide conditions) with 1 min break between the sessions, and we explicitly informed the participants of this during instructions for the monetary gambling task. This instruction might have caused the participants to treat the narrow-wide paired structured trials as a single block of trials and suppress the influence of the changes between the conditions. The univariate method we applied to analyze the brain data might also cause robustness to the changes between conditions (Burke et al. 2016). A multivariate analysis may be more efficient for detecting signals that are sensitive to such changes (Burke et al. 2016; Woo et al. 2017).

In the present study, we found a neural representation of utility prediction error. Because utility functions are usually defined on consequential states but not on processes that lead to them, some readers may think that the neural representation of utility is also a function on the consequential states. However, previous studies have shown that the activity of the reward system including vmPFC is modulated not only by consequential

states (Bartra, McGuire, and Kable 2013) but also by processes like opportunity equality (Aoki et al. 2014). Thus, it is not likely that utility-related information in the brain is limited to the consequential states usually considered in utility functions.

We focused on utility difference, not on utility level, in comparing utility interpersonally, because it could be treated quantitatively in the experiment. By comparing utility differences interpersonally, we actually conducted a redistribution to maximize the utilitarian social welfare of the participants within our budget constraints. Here, we note that the present study itself does not derive a normative conclusion (Hume's law (Hume 2003)). Utility level is also important to realize a full version of interpersonal comparison of utility (d'Aspremont and Gevers 1977; Roberts 1980). Future studies should conduct interpersonal comparison of utility level, which enables distributions according to variable principles discussed in modern justice theories other than utilitarianism (d'Aspremont and Gevers 1977; Roberts 1980; Hirose 2014).

We would like to note that the present study has a limitation based on the participant pool: we recruited only a few dozen participants from the student body of a specific university. In future studies, it will be necessary to examine whether neuroscientific methods for interpersonal comparison of utility are applicable to people beyond this population with larger samples.

The processes of constructing a quantitative concept and standardizing its measurement methods are to be viewed as progressive processes that increase coherence among elements of theory and instrumentation (Chang 2004; Tal 2013, 2015). This is the case for constructing an interpersonally comparable utility as well. The problem of measuring utility has frequently been compared with the problem of measuring temperature (Arrow 1963; von Neumann and Morgenstern 2007). We validated our interpersonal comparison method by showing coherence between objective measures (i.e., fMRI signals) and human intuition (Harsanyi 1977; Goldman 1995; Rossi 2014), as was done for the temperature scale in its initial stages (Chang 2004).

The proposed method is prominent in the academic sense that it enables us to escape from Arrow's impossibility (Arrow 1963; Sen 2018). Moreover, it can be applied for evidence-based policy making in nations that use cost-benefit analyses or optimal taxation theory for policy evaluation (Saez 2001; Boardman et al. 2018). Thus, we have demonstrated a major advance in social science using a neuroscientific measure (Camerer, Loewenstein, and Prelec 2005). The present study itself does not directly draw normative conclusions (Hume's law) on distributive justice. However, we believe that it raises a critical epistemic issue as a starting point for subsequent scientific inquiries to standardize a measurement method for utility as an interpersonally comparable quantity (Chang 2004) and for philosophical reflections on social structures (Rawls 1971; Daniels 1979).

References

Abdellaoui, Mohammed, Barrios, Carolina, and Wakker, Peter P., "Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory," *Journal of Econometrics*, 138 (2007), 356–378.

Abdellaoui, Mohammed, Bleichrodt, Han, and L'Haridon, Olivier, "A tractable method to measure utility and loss aversion under prospect theory," *Journal of Risk and Uncertainty*, 36 (2008), 245–266.

Abdellaoui, Mohammed, Bleichrodt, Han, and Paraschiv, Corina, "Loss Aversion Under Prospect Theory: A Parameter-Free Measurement," *Management Science*, 53 (2007), 1659–1674 (INFORMS).

Aoki, Ryuta, Matsumoto, Madoka, Yomogida, Yukihito, Izuma, Keise, Murayama, Kou, Sugiura, Ayaka, Camerer, Colin F., Adolphs, Ralph, and Matsumoto, Kenji, "Social Equality in the Number of Choice Options Is Represented in the Ventromedial Prefrontal Cortex," *The Journal of Neuroscience*, 34 (2014), 6413.

Arrow, Kenneth J., "Social choice and individual values," (New Haven, Connecticut, Yale university press, 1963).

d'Aspremont, Claude, and Gevers, Louis, "Equity and the Informational Basis of Collective Choice," *The Review of Economic Studies*, 44 (1977), 199–209.

Ballesta, Sébastien, Shi, Weikang, Conen, Katherine E., and Padoa-Schioppa, Camillo, "Values encoded in orbitofrontal cortex are causally related to economic choices," *Nature*, 588 (2020), 450–453.

Bartra, Oscar, McGuire, Joseph T., and Kable, Joseph W., "The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value," *NeuroImage*, 76 (2013), 412–427.

Bentham, Jeremy, "An Introduction to the Principles of Morals and Legislation," (1789).

Bloom, Paul, "Against empathy: The case for rational compassion," (Ecco, 2016).

Boardman, Anthony, Greenberg, David, Vining, Aidan, and Weimer, David, "Cost-Benefit Analysis," (Cambridge University Press, 2018).

Booij, Adam S., and Kuilen, Gijs, "A parameter-free analysis of the utility of money for the general population under prospect theory," *Journal of Economic Psychology*, 30 (2009), 651–666.

Burke, Christopher J., Baddeley, Michelle, Tobler, Philippe N., and Schultz, Wolfram, "Partial Adaptation of Obtained and Observed Value Signals Preserves Information about Gains and Losses," *The Journal of Neuroscience*, 36 (2016), 10016.

Butter, Charles M., McDonald, John A., and Snyder, Daniel R., "Orality, Preference Behavior, and Reinforcement Value of Nonfood Object in Monkeys with Orbital Frontal Lesions," *Science*, 164 (1969), 1306.

Camerer, Colin, Loewenstein, George, and Prelec, Drazen, "Neuroeconomics: How Neuroscience Can Inform Economics," *Journal of Economic Literature*, 43 (2005), 9–64.

Camille, Nathalie, Griffiths, Cathryn A., Vo, Khoi, Fellows, Lesley K., and Kable, Joseph W., "Ventromedial Frontal Lobe Damage Disrupts Value Maximization in Humans," *The Journal of Neuroscience*, 31 (2011), 7527.

Caplin, Andrew, Dean, Mark, Glimcher, Paul W., and Rutledge, Robb B., "Measuring Beliefs and Rewards: A Neuroeconomic Approach*," *The Quarterly Journal of Economics*, 125 (2010), 923–960.

Cardinal, Rudolf N., Pennicott, David R., Lakmali, C., Sugathapala, Robbins, Trevor W., and Everitt, Barry J., "Impulsive Choice Induced in Rats by Lesions of the Nucleus Accumbens Core," *Science*, 292 (2001), 2499.

Cassidy, Ryan Michael, and Tong, Qingchun, "Hunger and Satiety Gauge Reward Sensitivity," *Frontiers in Endocrinology*, 8 (2017), 104.

Chang, Hasok, "Inventing Temperature: Measurement and Scientific Progress," Oxford Studies in the Philosophy of Science (New York, Oxford University Press, 2004).

Daniels, Norman, "Wide Reflective Equilibrium and Theory Acceptance in Ethics," *The Journal of Philosophy*, 76 (1979), 256–282 (Journal of Philosophy, Inc.).

Davidson, Donald, "Problems of Rationality," (Oxford, Oxford University Press, 2004).

Díaz-Francés, Eloísa, and Rubio, Francisco, "On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables," *Statistical Papers*, 54 (2013), 309–323.

Elster, Jon, and Roemer, John E., eds., "Interpersonal Comparisons of Well-Being," Studies in Rationality and Social Change (Cambridge, Cambridge University Press, 1991).

Frank, Michael J., Seeberger, Lauren C., and O'Reilly, Randall C., "By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism," *Science*, 306 (2004), 1940.

Gaspar, Patricia, Berger, Brigitte, Febvret, Aude, Vigny, Annette, and Henry, Jean Pierre, "Catecholamine innervation of the human cerebral cortex as revealed by comparative immunohistochemistry of tyrosine hydroxylase and dopamine-beta-hydroxylase," *Journal of Comparative Neurology*, 279 (1989), 249–271 (John Wiley & Sons, Ltd).

Gläscher, Jan, "Visualization of Group Inference Data in Functional Neuroimaging," *Neuroinformatics*, 7 (2009), 73–82.

Glimcher, Paul W., and Fehr, Ernst, "Neuroeconomics: Decision making and the brain," (Academic Press, 2013).

Goldman, Alvin I., "Simulation and Interpersonal Utility," *Ethics*, 105 (1995), 709–726 (University of Chicago Press).

Gottfried, Jay A., O'Doherty, John, and Dolan, Raymond J., "Encoding Predictive Reward Value in Human Amygdala and Orbitofrontal Cortex," *Science*, 301 (2003), 1104.

Harsanyi, John C., "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, 63 (1955), 309–321 (University of Chicago Press).

——, "Morality and the theory of rational behavior," *Social research*, (1977), 623–656 (JSTOR).

Hirose, Iwao, "Egalitarianism," (Routledge, 2014).

HM Treasury, "The green book: Central government guidance on appraisal and evaluation," *London: HM Treasury*, (2018).

Hume, David, "A treatise of human nature," (Courier Corporation, 2003).

Izuma, Keise, Saito, Daisuke N., and Sadato, Norihiro, "Processing of Social and Monetary Rewards in the Human Striatum," *Neuron*, 58 (2008), 284–294.

Kable, Joseph W, and Glimcher, Paul W, "The neural correlates of subjective value during intertemporal choice," *Nature Neuroscience*, 10 (2007), 1625–1633.

Kahneman, Daniel, and Tversky, Amos, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47 (1979), 263–91.

Keysers, Christian, Gazzola, Valeria, and Wagenmakers, Eric-Jan, "Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence," *Nature Neuroscience*, 23 (2020), 788–799.

Kobayashi, Shunsuke, Pinto de Carvalho, Ofelia, and Schultz, Wolfram, "Adaptation of Reward Sensitivity in Orbitofrontal Neurons," *The Journal of Neuroscience*, 30 (2010), 534.

Krantz, David, Luce, Duncan, Suppes, Patrick, and Tversky, Amos, "Foundations of Measurement, Vol. I: Additive and Polynomial Representations," (New York Academic Press, 1971).

Kreps, David M., "Microeconomic Foundations I: Choice and Competitive Markets," (Princeton University Press, 2012).

Lebreton, Maël, Bavard, Sophie, Daunizeau, Jean, and Palminteri, Stefano, "Assessing inter-individual differences with task-related functional neuroimaging," *Nature Human Behaviour*, 3 (2019), 897–905.

Lee, Michael D., and Wagenmakers, Eric-Jan, "Bayesian cognitive modeling: A practical course.," Bayesian cognitive modeling: A practical course. (New York, NY, US, Cambridge University Press, 2013).

Levy, Dino J, and Glimcher, Paul W, "The root of all value: a neural common currency for choice," *Decision making*, 22 (2012), 1027–1038.

Lewis, David A., Foote, Stephen L., Goldstein, Menek, and Morrison, John H., "The dopaminergic innervation of monkey prefrontal cortex: a tyrosine hydroxylase immunohistochemical study," *Brain Research*, 449 (1988), 225–243.

Louie, Kenway, Glimcher, Paul W, and Webb, Ryan, "Adaptive neural coding: from biological to behavioral decision-making," *Neuroeconomics*, 5 (2015), 91–99.

Luce, R. Duncan, "Individual choice behavior.," Individual choice behavior. (Oxford, England, John Wiley, 1959).

Luce, Robert Duncan, "Utility of gains and losses: Measurement-theoretical and experimental approaches.," Utility of gains and losses: Measurement-theoretical and experimental approaches. (Mahwah, NJ, US, Lawrence Erlbaum Associates Publishers, 2000).

Matsumoto, Madoka, Matsumoto, Kenji, Abe, Hiroshi, and Tanaka, Keiji, "Medial prefrontal cell activity signaling prediction errors of action values," *Nature Neuroscience*, 10 (2007), 647–656.

McClure, Samuel M., Laibson, David I., Loewenstein, George, and Cohen, Jonathan D., "Separate Neural Systems Value Immediate and Delayed Monetary Rewards," *Science*, 306 (2004), 503.

Moscati, Ivan, "Measuring Utility: From the Marginal Revolution to Behavioral Economics," Oxford Studies in History of Economics (New York, Oxford University Press, 2018).

Müller-Lyer, F. C., "Optische urteilstauschungen," *Archiv fur Anatomie und Physiologie, Physiologische Abteilung*, 2 (1889), 263–270.

Murayama, Kou, Matsumoto, Madoka, Izuma, Keise, Sugiura, Ayaka, Ryan, Richard M., Deci, Edward L., and Matsumoto, Kenji, "How Self-Determined Choice Facilitates Performance: A Key Role of the Ventromedial Prefrontal Cortex," *Cerebral Cortex*, 25 (2015), 1241–1251.

Narens, Louis, and Skyrms, Brian, "The Pursuit of Happiness: Philosophical and Psychological Foundations of Utility," (Oxford, Oxford University Press, 2020).

von Neumann, John, and Morgenstern, Oskar, "Theory of Games and Economic Behavior," (Princeton University Press, 2007).

Oppenheimer, Daniel M., Meyvis, Tom, and Davidenko, Nicolas, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of Experimental Social Psychology*, 45 (2009), 867–872.

Ratcliff, Roger, Smith, Philip L., Brown, Scott D., and McKoon, Gail, "Diffusion Decision Model: Current Issues and History," *Trends in Cognitive Sciences*, 20 (2016), 260–281.

Rawls, John, "A Theory of Justice," (Harvard University Press, 1971).

Robbins, Lionel, "An essay on the nature and significance of economic science," (Ludwig von Mises Institute, 1932).

——, "Interpersonal Comparisons of Utility: A Comment," *The Economic Journal*, 48 (1938), 635–641 ([Royal Economic Society, Wiley]).

Roberts, Kevin W. S., "Interpersonal Comparability and Social Choice Theory," *The Review of Economic Studies*, 47 (1980), 421–439 ([Oxford University Press, Review of Economic Studies, Ltd.]).

Rossi, Mauro, “Simulation theory and interpersonal utility comparisons reconsidered,” *Synthese*, 191 (2014), 1185–1210.

Saez, Emmanuel, “Using elasticities to derive optimal income tax rates,” *The review of economic studies*, 68 (2001), 205–229 (Wiley-Blackwell).

Schultz, Wolfram, Dayan, Peter, and Montague, P. Read, “A Neural Substrate of Prediction and Reward,” *Science*, 275 (1997), 1593.

Sen, Amartya, “The Idea of Justice,” (Belknap Press of Harvard University Press, 2009).

——, “Collective choice and social welfare,” (Cambridge, Massachusetts, Harvard University Press, 2018).

Seymour, Ben, Daw, Nathaniel, Dayan, Peter, Singer, Tania, and Dolan, Ray, “Differential Encoding of Losses and Gains in the Human Striatum,” *The Journal of Neuroscience*, 27 (2007), 4826.

Smith, Adam, “The Theory of Moral Sentiments,” (1759).

Stauffer, William R., Lak, Armin, and Schultz, Wolfram, “Dopamine Reward Prediction Error Responses Reflect Marginal Utility,” *Current Biology*, 24 (2014), 2491–2500 (Elsevier).

Stott, Henry P., “Cumulative prospect theory’s functional menagerie,” *Journal of Risk and Uncertainty*, 32 (2006), 101–130.

Sutton, Richard S., and Barto, Andrew G., “Reinforcement learning: An introduction,” (MIT press, 2018).

Tal, Eran, “Old and New Problems in Philosophy of Measurement,” *Philosophy Compass*, 8 (2013), 1159–1173 (John Wiley & Sons, Ltd).

——, “Measurement in Science,” in *Stanford Encyclopedia of Philosophy*, Edward Zalta, ed. (2015).

Tversky, Amos, and Kahneman, Daniel, “Advances in prospect theory: Cumulative representation of uncertainty,” *Journal of Risk and Uncertainty*, 5 (1992), 297–323.

Wakker, Peter, and Deneffe, Daniel, “Eliciting von Neumann-Morgenstern Utilities When Probabilities Are Distorted or Unknown,” *Management Science*, 42 (1996), 1131–1150.

Waldner, Ilmar, “The Empirical Meaningfulness of Interpersonal Utility Comparisons,” *Journal of Philosophy*, 69 (1972), 87 (Journal of Philosophy Inc).

Woo, Choong-Wan, Chang, Luke J, Lindquist, Martin A, and Wager, Tor D, “Building better biomarkers: brain models in translational neuroimaging,” *Nature Neuroscience*, 20 (2017), 365–377.

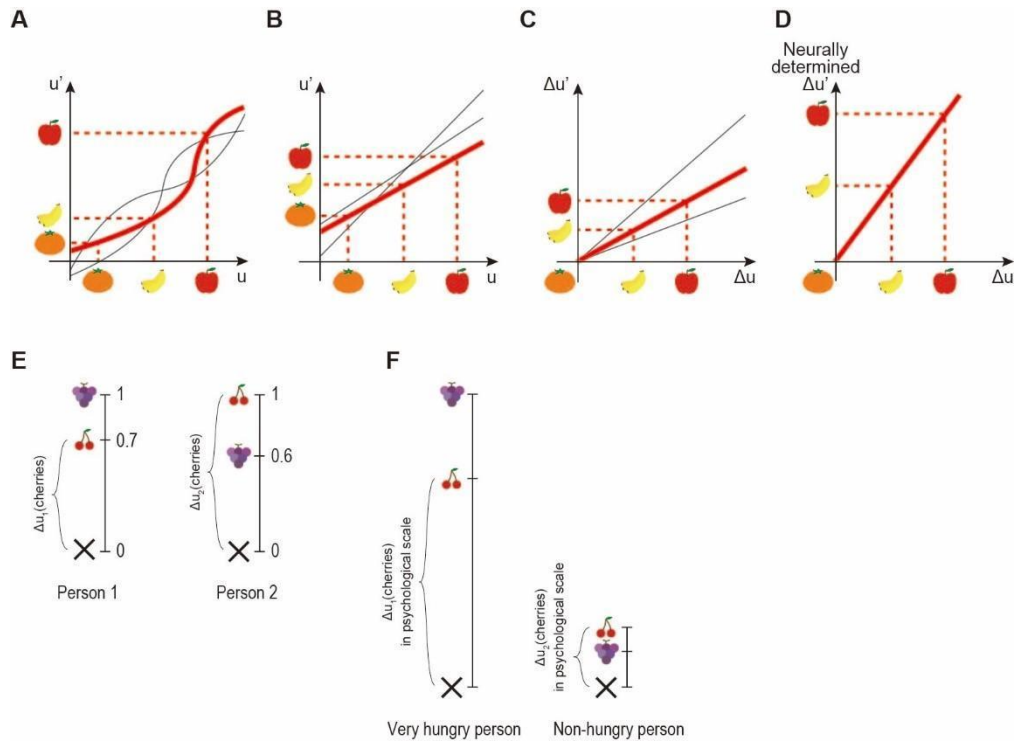


Fig. 1. Concepts of utility.

(**A**) Utility as an ordinal scale (ordinal utility). Each axis represents the preference for an apple over a banana and a banana over an orange. Because only the order among the options is meaningful in this scale, any monotonically increasing transformation is permitted without losing any information. A representative (red line) and two variants (gray thin lines) are shown. (**B**) Utility as an interval scale (cardinal utility). Each axis represents utilities in which the utility difference between an apple and a banana is the same as the utility difference between a banana and an orange. Because only the distances between the options are meaningful in this scale, any positive linear transformation ($u' = \alpha u + \beta$, $\alpha > 0$) is permitted without losing any information. A representative (red line) and two variants (gray thin lines) are shown. (**C**) Utility difference as a ratio scale. If the cardinal utility was given, the utility difference (Δu) is a ratio scale by subtracting the reference option (an orange in the figure). Any positive scalar multiplication ($\Delta u' = \alpha \Delta u$, $\alpha > 0$) is permitted without losing any information. (**D**) Given a neuroscientific measurement method, a single scale for Δu can be determined by using the axis of neural activity correlating with Δu in

each individual. (**E**) As the lengths of the scales for cardinal utility have no meaning, a 0-1 rescaling is applied to express the utilities of the two individuals. Δu_1(cherries) = 0.7 and Δu_2(cherries) = 1, where $\Delta u(\cdot) = u(\cdot) - u(\text{none})$. Thus, $\Delta u_1(\text{cherries}) < \Delta u_2(\text{cherries})$, although this comparison has no empirical meaning. (**F**) A modified case of **E**. Suppose Person 1 is very hungry and Person 2 is not hungry at all. We all intuitively recognize that a hungry person is more pleased by receiving cherries than a non-hungry person. Thus, we can give a specific meaning to the lengths of the scales to match our intuition by making the hungry person's scale longer and the non-hungry person's scale shorter, providing apparently larger psychological utility of cherries for the hungry person than for the non-hungry person in contrast to **E**.

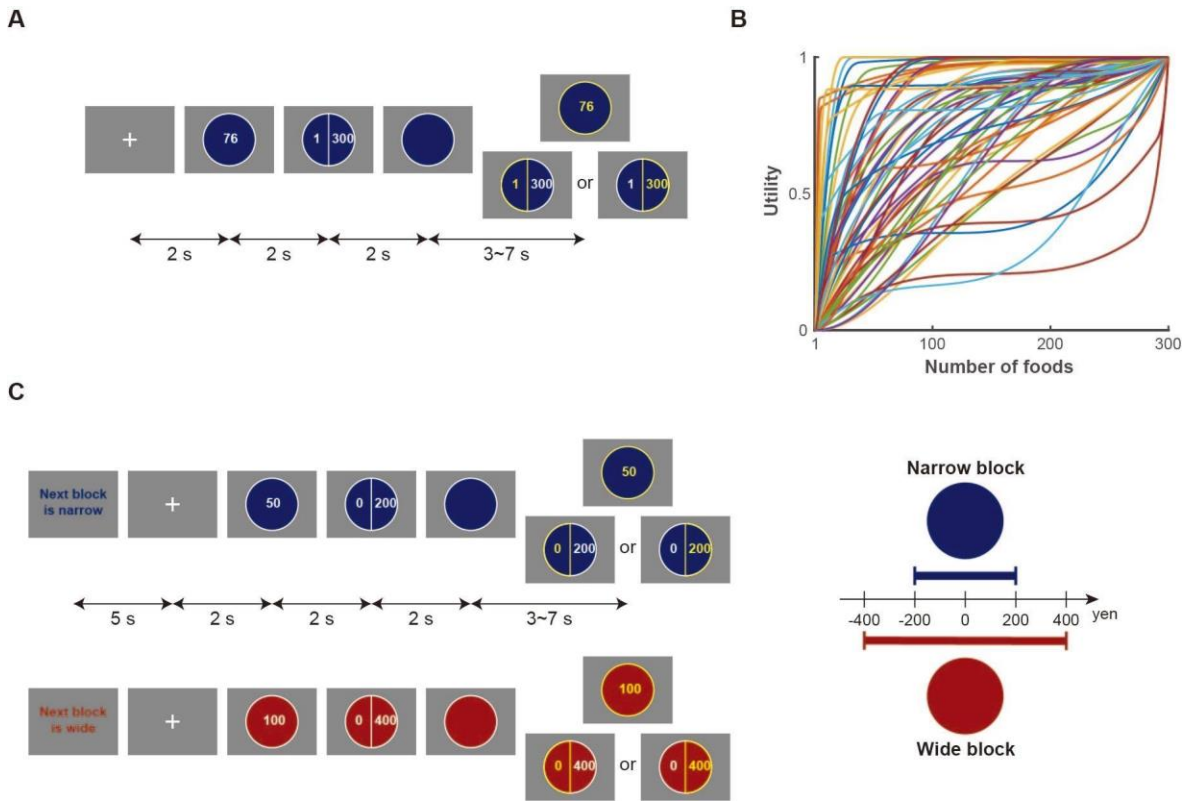


Fig. 2. Gambling tasks.

(**A**) Food gambling task. Participants were required to choose between a sure (riskless) offer and a lottery to get indicated amounts of snacks as reward. (**B**) The experimentally obtained utility functions for food amounts. 0-1 rescaling was applied to the utility functions. (**C**) Monetary gambling task. Participants got money as a reward. This task consisted of alternating narrow and wide blocks with the corresponding offer ranges (-¥200 - +¥200 and -¥400 - +¥400) as illustrated at the right side.

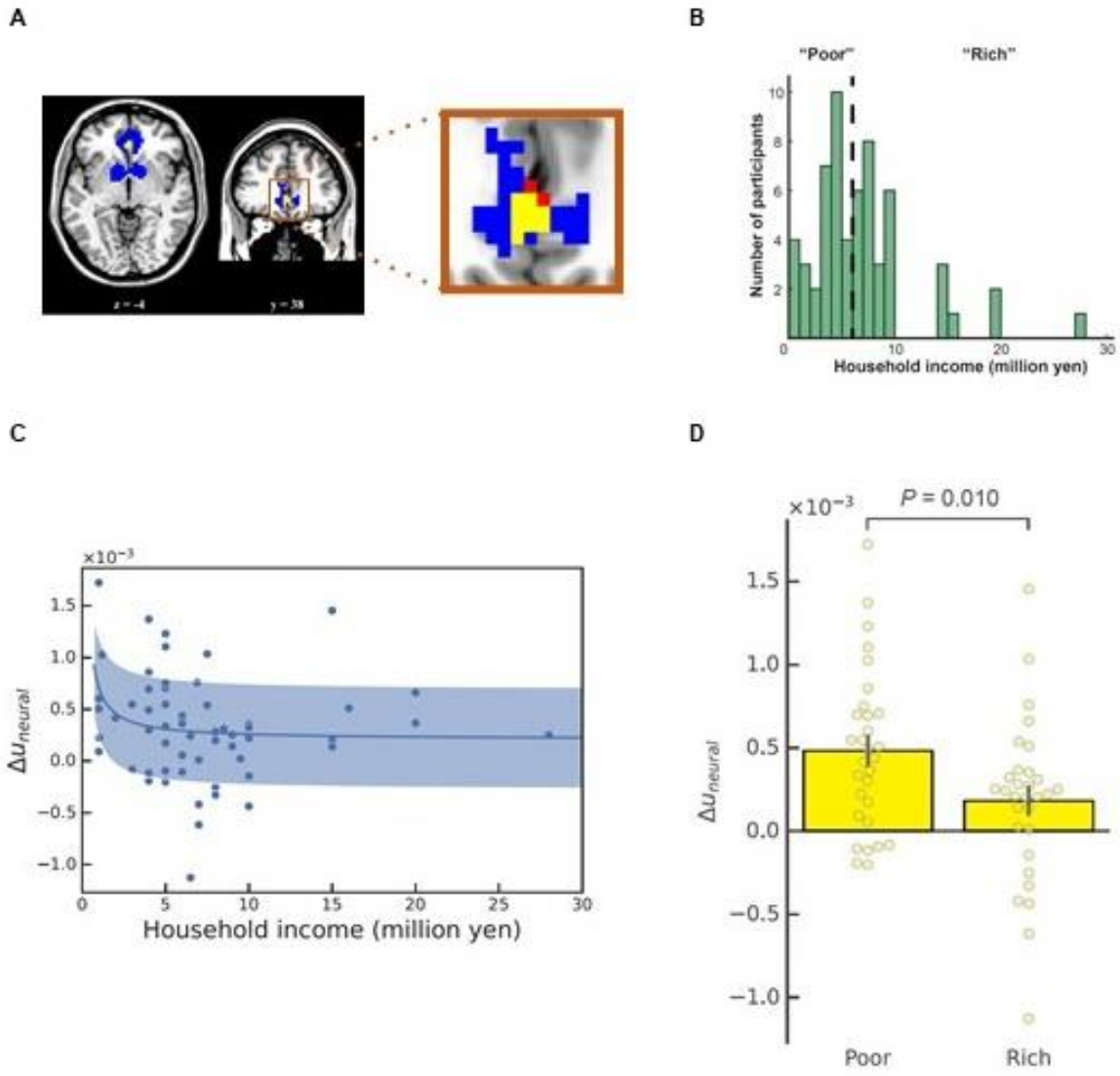


Fig. 3. Interpersonal comparison of neural measures for utility.

(**A**) Neural representations of utility prediction error. The activated regions are shown for the food gambling task (red), monetary gambling task (blue), and their overlapped region (yellow) ($P < 0.05$, FWE corrected). The area in the frame is magnified at right. The overlapped region in the anterior cingulate cortex and its adjacent ventromedial prefrontal cortex (ACC/vmPFC) is called the "utility region". The percent signal change in the utility

region per 1 yen is defined as Δu_{neural}. (**B**) Distribution of annual household incomes of the participants for the gambling tasks. (**C**) Δu_{neural} of the participants are plotted against their household incomes. Δu_{neural} was inversely proportional to income ($\beta_{income^{-1}} = 0.247 \pm 0.132$ (mean ± SE), $t(56) = 1.874$, $P = 0.033$ one-sided) after regressing out age and sex. The shaded area indicates SD. (**D**) When the participants were split between the lower half ("poor" group) and the upper half ("rich" group) (indicated by the vertical dashed line at 6 million yen in **B**), the poor group showed significantly greater Δu_{neural} than the rich group ($t(58) = 2.39$, $P = 0.010$ one-sided, $BF_{+0} = 5.723$). Error bars indicate SEM across group members.

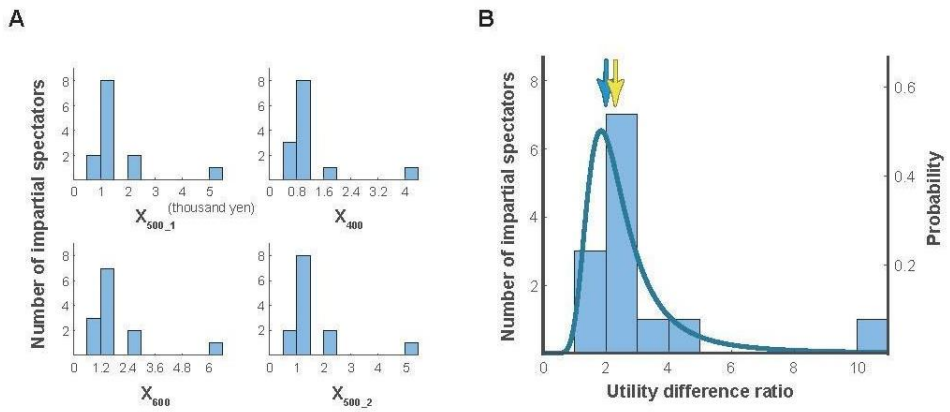


Fig. 4. Validation by impartial spectators' estimation.

(**A**) The distributions of impartial spectators' estimates of the amounts of money needed to please the rich group equivalently to the poor group, when the participants were asked 4 times (for the poor group participants' receipt of ¥500, ¥400, ¥600, and ¥500) in the impartial spectator task. X_{500_1}, X_{400}, X_{600}, and X_{500_2} are the estimates for the corresponding questions. (**B**) The distribution of the impartial spectators' estimates of the utility difference ratio of the poor group to the rich group based on the combined data of **A** and its fit curve on the Fieller-Hinkley distribution. Yellow and light blue arrows indicate the Δu_{neural} ratio and the median of impartial spectators' estimates, respectively.

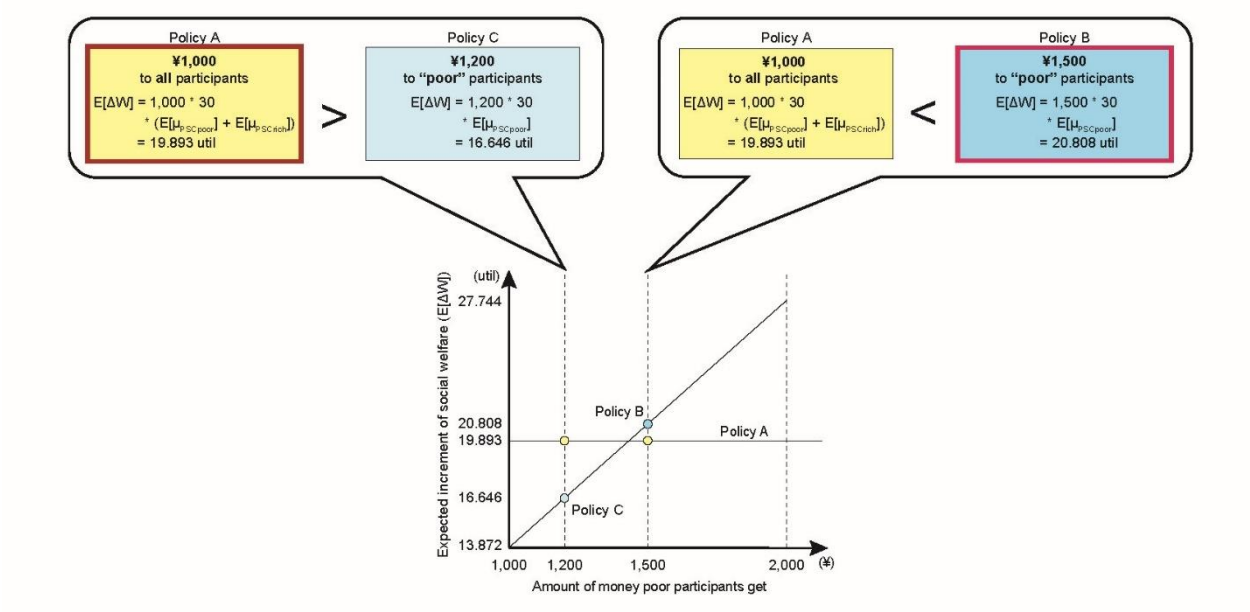


Fig. 5. Application to a distribution problem.

Two distribution policies, A and B, were compared (right balloon). Policy B was chosen, because it increased expected utilitarian social welfare (E[ΔW]) more than policy A. Policies A and C are also compared (left balloon), and policy A would be chosen over policy C using the same rationale. The plot illustrates expected increments of social welfare by distributing ¥1,000 to all participants and by distributing ¥x to "poor" participants, including points for the compared policies A, B, and C.

For Online Publication

APPENDIX: Methods

Gambling experiment: Participants

63 healthy right-handed student volunteers who live with their parents completed this experiment. All participants were native Japanese speakers recruited from the student pool of Tamagawa University. No participant had a history of neurological or psychiatric disease, and written informed consent was obtained from all participants prior to participation. The study was approved by the ethical committee of Tamagawa University in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. One participant was excluded from the analysis because of a technical problem with data storage. In addition, in the monetary gambling task, 2 participants were excluded from the analysis because of excessive head motion (> 3 mm) during the task, resulting in the final sample of 60 (35 females and 25 males; age: 20.13 ± 0.98 years, range: 18-22 years). In the food gambling task, 3 participants (2 were the same participants as those excluded in the monetary gambling task) were excluded from the analysis because of excessive head motion, and 3 other participants were excluded from the analysis because of insufficient motivation to obtain the food, judged from < 5 certainty equivalents (CEs) at largest, resulting in the final sample of 56 (33 females and 23 males; age: 20.09 ± 0.98 years, range: 18-22 years).

Procedures

Participants conducted the monetary gambling task and the food gambling task in the MRI scanner. Linear utility functions were assumed for money, and non-linear utility functions were assumed for food. Afterward, participants completed a probability weighting task outside the MRI scanner to measure subjective probability for a 50/50 gambling situation.

Monetary gambling task

In this task, each participant was required to choose between a sure (riskless) offer and a lottery (Fig. 2C). Participants were instructed that they would gain or lose money in each trial depending on the outcome of the gamble. Each fMRI session consisted of a narrow block (20 trials) in which the amount of money was in the range ¥-200~200 and a wide block (20 trials) in which the amount of money was in the range ¥-400~400. Participants were explicitly informed about the range of possible monetary outcomes in each block, and could discern the current block according to the color of the disk on which the amount of money was displayed. Each trial began with a fixation cross at the center of the monitor (2 s). Next, the sure offer was presented (2 s). There were five sure offers each in the narrow block (¥-100, -50, 0, 50, 100) and the wide block (¥-200, -100, 0, 100, 200). Next, the sure offer disappeared and the lottery offer was presented (2 s). Each lottery offer was a gamble entailing a 50% chance of gaining (or losing) some amount of money, and a 50% chance of neither gain nor loss of money (¥0). There were four amounts each for the lottery in the narrow block (¥-200, -100, 100, 200) and the wide block (¥-400, -200, 200, 400). Next, the lottery offer disappeared and participants chose between the sure and lottery offers. Each participant was required to press a left or right button with the index or middle finger of the right hand within 30 s to choose the corresponding offer. The trial was aborted if the participant failed to respond within the time window ("error trials"). The association between buttons and offers was counterbalanced across participants. The chosen offer was presented on the monitor again immediately after pressing the button. Next, the amount of money earned was displayed in flashing yellow text. The delay from button press to feedback was jittered between 3 and 7 s (average 5 s). Inter-trial intervals were jittered between 3 and 7 s (average 5 s). All 5 * 4 = 20 possible combinations of sure and lottery offers were presented in each block, in randomized order. Each participant was engaged in this task for 3 sessions of 40 trials except the error trials. The order and color of the 2 blocks were counterbalanced across participants. The block-start cue ("Next block is narrow" or "Next block is wide") was presented in the beginning of each block (5 s). Stimulus presentation and timing data collection for all stimuli and response events were

performed using Matlab (MathWorks, Natick, MA, USA) and Psychtoolbox (www.psychtoolbox.org).

Previous studies have shown that utility for money is approximately linear for moderate amounts of money (Wakker and Deneffe 1996; Abdellaoui, Bleichrodt, and L'Haridon 2008; Booij and Kuilen 2009). Because we used small amounts of money, we regarded the utility function for money as linear. To dissociate the brain regions that represent utility from the amount of reward, we used a food gambling task.

Food gambling task

In this task, each participant was required to choose between a sure (riskless) offer and a lottery with the same appearance as the monetary gambling task (Fig. 2A). Each participant could get "tickets to be exchanged for snacks" in this task. Before the fMRI scan, each participant chose their preferred snack (rice cracker or chocolate). Participants were instructed that at the end of the experiment, one trial would be randomly selected and the number of tickets to exchange for the preferred snack would be determined according to the outcome of their actual decision. One ticket could be exchanged for one piece of the snack, so the number of tickets they obtained was the maximum number of their preferred snack they could receive. Each trial began with a fixation cross at the center of the monitor (2 s). Then, the sure offer was presented (2 s). The possible number of tickets in sure offers was 1~300. Then the lottery offer was presented (2 s). Each lottery offer was a gamble entailing a 50% chance of gaining one of two numbers of tickets. The possible number of tickets in lottery offers was 1~300. Then, the lottery offer disappeared and participants chose between the sure and lottery offers. Each participant was required to press the left or right button with the index or middle finger of the right hand within 30 s to choose the corresponding offer. The trial was aborted if the participant failed to respond within the time-window ("error trials"). The association between buttons and offers was the same as in the monetary gambling task. The chosen offer was presented on the monitor again immediately after pressing the button. Then, the numerical outcome from the chosen offer was displayed in flashing yellow text as in the monetary gambling task. The delay from

button press to feedback was jittered between 3 and 7 s (average 5 s). Inter-trial intervals were jittered between 3 and 7 s (average 5 s). The color of the disk (red or blue) was counterbalanced across participants. In this task, we measured certainty equivalents (CEs) based on the fractile method with three phases (Supplementary Fig. 1B) (see Constructing Utility Functions in food gambling task in detail). In each phase, CE measurement trials and a random trial (all three numbers were sampled from a uniform distribution, unif{1, 300}) were randomly permuted and this procedure was repeated until all CEs converged (see Measuring CEs in detail). Each participant was engaged in this task for 3~5 sessions of 40 trials, and so completed a minimum of 120 trials. The fMRI data from these 120 trials were analyzed, whereas all trials were used to analyze the behavioral data. The fMRI session for this task was terminated when the sum of completed trials from the food task and error trials from both monetary and food tasks reached 200. Stimulus presentation and timing data collection for all stimuli and response events were performed using Matlab (MathWorks, Natick, MA, USA) and the Psychtoolbox (www.psychtoolbox.org).

Probability weighting task

This task was very similar to the previous tasks. Participants were instructed that at the end of the experiment, one trial would be randomly selected and a monetary payment would be made according to their actual decision. We measured the CE of a 50/50 gamble (p = 0.5 for each outcome) between ¥2,000 and ¥2,500 using Parameter Estimation through Sequential Testing (PEST) (see Measuring CEs in detail). CE measurement trials and a random trial (all three numbers were sampled from the uniform distribution, 10 * unif{200, 250}) were randomly permuted and this procedure was repeated until the CE converged. Each participant engaged in this task for typically 30-40 trials. Stimulus presentation and timing data collection for all stimuli and response events were performed using Matlab (MathWorks, Natick, MA, USA) and the Psychtoolbox (www.psychtoolbox.org).

Questionnaire

After the gambling experiment, each participant was asked about their understanding of the tasks and their household income.

MRI data acquisition

Functional imaging was conducted using a 3T Trio A Tim MRI scanner (Siemens, Germany) and gradient echo T2*-weighted echo-planar images (EPI) were acquired with blood oxygenation level-dependent contrasts. 42 contiguously interleaved transverse EPI-image slices were acquired in each volume (slice thickness, 3 mm; no gap; repetition time, 2,500 ms; echo time, 25 ms; flip angle, 90°; field of view, 192 mm^2; matrix, 64 × 64). Slice orientation was tilted −30° from the AC-PC line. For each participant, data were acquired in 3 scanning sessions for the monetary gambling task and 3-5 scanning sessions for the food gambling task. We discarded without analysis the initial 3 dummy volumes per session to allow for the T1 equilibrium effect. A high-resolution anatomical T1-weighted image was also acquired for each participant.

Behavioral analyses: Measuring CEs

To measure CEs in the food gambling and probability weighting tasks, we used PEST (Luce 2000; Stauffer, Lak, and Schultz 2014) (Supplementary Fig. 1A).

We assessed the number of tickets or the amount of money in the sure offer that was subjectively equivalent to the value associated with each lottery offer. The rules governing the PEST procedure were adapted from Luce (2000). As we noted above, each trial consisted of three numbers (sure number (SU), lottery win number (LW), and lottery lose number (LL)). Each PEST sequence consisted of several trials during which one constant lottery offer (constant LW and LL) was presented as a choice option against the sure offer (variable SU). On the initial trial of a PEST sequence, SU was determined by rounding (LW + LL)/2 to the nearest integer. Based on the participant's choice between the sure offer and the lottery, SU was adjusted on the subsequent trial. If the participant chose the lottery offer on trial t, then SU was increased by ε on trial t + 1. However, if the participant chose the sure offer on trial t, SU was reduced by ε on trial t + 1. Initially, ε was determined

by rounding (LW + LL)/4 to the nearest integer. ε was adjusted according to the following rules. Rule 1: after the third trial of a PEST sequence, every time the participant switched from one option to the other, the size of ε was halved. Rule 2: if updated SU based on ε was equal to LW or LL, the size of ε was halved and then SU was updated. Thus, the procedure converged by locating subsequent sure offers on either side of the true indifference value and reducing ε until the interval containing the indifference value was small (1 for tickets and 10 for money). When the updated SU was the same as the previous SU, the PEST procedure terminated.

Constructing Utility Functions in the food gambling task

We determined each participant's utility function for tickets in the range between 1 and 300 tickets using the fractile method based on 50/50 gambles (Stauffer, Lak, and Schultz 2014) (one example fractile procedure is shown in Supplementary Fig. 1B). To do so, we first measured the CE of a 50/50 gamble between 1 and 300 tickets (x_1) using PEST. We then used x_1 as an outcome to construct two new gambles (1 ticket, p = 0.5 and x_1 tickets, p = 0.5; x_1 tickets, p = 0.5 and 300 tickets, p = 0.5) and measured their CEs (x_2 and x_3). Finally, we measured the CEs of a 50/50 gamble between 1 and x_2 tickets (x_4), that between x_2 and x_1 tickets (x_5), that between x_1 and x_3 tickets (x_6) and that between x_3 and 300 tickets (x_7). We fitted the CE data using local data interpolation (i.e., splines, MATLAB SLM tool) (Stauffer, Lak, and Schultz 2014), after determining the probability weighting function (see below). This procedure fits cubic functions on consecutive segments of the data. We used two polynomial pieces (three knots) for our fittings. The left side knot was located at 1, the right side knot was located at 300, and the middle knot position was determined by minimizing the difference between empirical data and the fitted curve (candidate positions were 2, 3, ..., 299). We restricted the fitting to have a positive slope over the whole range of the outcomes, and we required the function to be weakly increasing based on the fundamental economic assumption that more of a good does not decrease total utility. This procedure provided an estimation of the shape of the participant's utility function in the range of 1-300 tickets.

Estimating w(p = 0.5) in probability weighting task

On the basis of the assumption that utility for moderate amounts of money is approximately linear, we set $u(¥x) = \alpha x + \beta$.

This means that the value of the probability weighting function w when p = 0.5 can be measured by using the CE in the probability weighting task, according to the following equations.

$w(p = 0.5) * u(¥2500) + (1 - w(p = 0.5)) * u(¥2000) = u(¥CE)$

$w(p = 0.5) * (\alpha * 2500 + \beta) + (1 - w(p = 0.5)) * (\alpha * 2000 + \beta) = \alpha * CE + \beta$

$w(p = 0.5) * 2500 + (1 - w(p = 0.5)) * 2000 = CE$

$w(p = 0.5) = (CE - 2000)/500$

Note that these equations hold in both original prospect theory (Kahneman and Tversky 1979) and cumulative prospect theory (Tversky and Kahneman 1992), because these two theories coincide on the domain of two-outcome lotteries (Abdellaoui, Bleichrodt, and Paraschiv 2007).

The expected utility of a lottery with probability weighting function can be calculated as

$E[lottery] = w(p = 0.5) * u(LW) + (1 - w(p = 0.5)) * u(LL)$ if $LL \geq 0$,

$E[lottery] = w(p = 0.5) * u(LL) + (1 - w(p = 0.5)) * u(LW)$ if $LW \leq 0$.

The values of w(p = 0.5) were compared between the participants whose household incomes were in the lower half among all participants (the “poor” group) and the upper half (the “rich” group) using the classical t-test and Bayesian t-test with default effect size priors (Cauchy scale $2^{0.5}/2$) (Keysers, Gazzola, and Wagenmakers 2020) (Supplementary Fig. 3). BFs were computed using the Savage–Dickey density ratio method. Note that BF < 1/3 is considered moderate evidence for the null hypothesis, and BF > 3 is considered moderate evidence for the alternative hypothesis (roughly similar to $p < 0.05$) (Lee and Wagenmakers 2013; Keysers, Gazzola, and Wagenmakers 2020). Bayesian t-test results are reported using the two-tailed Bayes factor BF_{10} that represents $P(data| H_1: \mu_{wpoor} \neq \mu_{wrich})/P(data| H_0: \mu_{wpoor} = \mu_{wrich})$, where μ_{wpoor} and μ_{wrich} indicate the mean of w(p = 0.5) from the poor group and that from the rich group, respectively. Effect size estimates are

reported as median posterior Cohen's δ with 95% CI. Bayesian hypothesis tests in this paper were performed using R (https://www.R-project.org/) and Stan (https://mc-stan.org/), and all Markov chain Monte Carlo settings were 3 chains with 100,000 samples each, with a burn-in of 10,000 and a thinning factor of 1. For all parameters, $\hat{R}$s were lower than 1.1.

Validation of participants' utility functions

We validated the shape of each participant's utility function (estimated curve for tickets) in the food task by using the random trials, in which the choice options did not depend on participants' previous choices (i.e., incentive compatible) (Stauffer, Lak, and Schultz 2014). In these trials, we assessed the percentages of trials in which participants chose the offer with larger estimated utility, and compared them with those assessed on the basis of expected reward amount (weighted by subjective probability) using the classical t-test.

Pre-processing of functional images

The following preprocessing procedures were performed using Statistical Parametric Mapping (SPM12) software (Wellcome Department of Imaging Neuroscience, London, UK, https://www.fil.ion.ucl.ac.uk/spm/) implemented in MATLAB: correction for head motion, adjustment of acquisition timing across slices, spatial normalization to the standard MNI template, and smoothing using a Gaussian kernel with a full-width at half maximum of 8 mm.

Functional MRI Data Analysis

We adopted a conventional two-level approach using SPM12 to analyze the fMRI data. A set of regressors was generated by convolving a canonical hemodynamic response function provided by SPM12 with a series of epochs. A voxel-by-voxel multiple regression analysis was applied to the preprocessed images for each participant. Statistical inference on contrasts of parameter estimates was then performed at the second-level between-participant analysis, using a one-sample t-test. Because we had an a priori hypothesis

focusing on the activations of the value system, the second-level analysis was performed for the voxels within the explicit anatomical mask that is manually drawn to include the whole medial prefrontal cortex, whole caudate head and body, whole putamen, and whole pallidum. Unless otherwise stated, the statistical threshold was set at $p < 0.05$, family-wise error (FWE) corrected, $k = 10$.

First, we analysed the data from the food gambling task to determine brain regions which correlate with estimated utility (GLM1) and amount of food (GLM2). GLM1 included 5 parametric regressors (1) the utility of the sure offer, (2) the expected utility of the lottery offer weighted by subjective probability ($w(p = 0.5)$), (3) the utility of the chosen option weighted by subjective probability, (4) the utility of sure feedback, and (5) the utility prediction error of lottery feedback weighted by subjective probability, at each event onset. Onsets of each trial per se and six motion parameters were treated as regressors of no interest. Similarly, GLM2 included 5 parametric regressors (1) the amount of the sure offer, (2) the expected amount of the lottery offer weighted by subjective probability, (3) the expected amount of the chosen option weighted by subjective probability, (4) the amount of sure feedback, and (5) the amount prediction error of lottery feedback weighted by subjective probability, at each event onset. In addition, the same regressors of no interest as those in GLM1 were included. We were able to identify the brain region whose activity correlated with the utility prediction error (from GLM1), but not with the amount prediction error (from GLM2).

Second, in order to determine whether the neural representation of utility prediction error identified by GLM1 was affected by the change in width of reward across blocks, we analyzed the data from the monetary gambling task (GLM3). GLM3 included the following parametric regressors (1) the monetary amount of the sure offer, (2) the expected monetary amount of the lottery offer weighted by subjective probability, (3) the monetary amount of the chosen option weighted by subjective probability, (4) the monetary amount of sure feedback, (5) the amount prediction error of lottery feedback weighted by subjective probability in the narrow block, and (6) the amount prediction error of lottery feedback weighted by subjective probability in the wide block, at each event onset. In addition, the

same regressors of no interest as those in GLM1 were included. We found that the neural representation of utility prediction error (= amount prediction error in monetary gambling task) identified by the analysis with GLM1 was not affected by the change in width of reward across blocks. Then, we identified all the voxels representing utility prediction errors without discriminating between the narrow and wide blocks in the monetary gambling task (GLM4). GLM4 included the same regressors as GLM3 except that the prediction error of lottery feedback for both narrow and wide blocks was modeled as a single parametric regressor. By finding the overlapped region between the neural representations of utility prediction error for money (GLM4) and food (GLM1), we determined the brain region which represents the common currency of utility. This was performed using MarsBaR (http://marsbar.sourceforge.net). All regressors were not standardized.

Region-of-interest (ROI) analyses

The rfxplot toolbox (http://rfxplot.sourceforge.net/) (Gläscher 2009) was used to extract percent signal change (PSC) at regions of interest.

For the region whose activity correlated with utility prediction errors in the food gambling task, the PSCs evoked by the utility prediction errors per ¥1 were compared between narrow and wide blocks using the classical paired t-test and Bayesian paired t-test with default effect size priors (Cauchy scale $2^{0.5}/2$) (Keysers, Gazzola, and Wagenmakers 2020). Bayesian t-test results are reported using the one-tailed Bayes factor BF_{+0} that represents $P(\text{data}|\ H_{+}: \mu_{PSCnarrow} > \mu_{PSCwide})/P(\text{data}|\ H_{0}: \mu_{PSCnarrow} = \mu_{PSCwide})$, where $\mu_{PSCnarrow}$ and $\mu_{PSCwide}$ indicate the mean of PSCs from narrow block and that from wide block, respectively. Effect size estimates are reported as median posterior Cohen's δ with 95% CI using an H_1 without restriction of $\mu_{PSCnarrow} \geq \mu_{PSCwide}$ in order not to bias estimates in the expected direction.

To evaluate the relationship between the PSCs of the “utility region” (Δu_{neural}) and each participant’s household income, we employed multiple regression analyses as shown below

$$\Delta u_{neural} = \beta_{income^{-1}} * income^{-1} + \beta_{age} * age + \beta_{sex} * sex + \beta_0 + \varepsilon.$$

All variables were standardized, and a dummy variable was used for sex (male = 0, female = 1).

Δu_{neural} were compared between poor and rich groups using the classical t-test and Bayesian t-test with default effect size priors (Cauchy scale $2^{0.5}/2$). Bayesian t-test results are reported using the one-tailed Bayes factor BF_{+0} that represents $P(data|\ H_+: \mu_{PSCpoor} > \mu_{PSCrich})/P(data|\ H_0: \mu_{PSCpoor} = \mu_{PSCrich})$, where $\mu_{PSCpoor}$ and $\mu_{PSCrich}$ indicate the mean of PSCs from the poor group and the rich group, respectively. Effect size estimates are reported as median posterior Cohen's δ with 95% CI using an H_1 without restriction of $\mu_{PSCpoor} \geq \mu_{PSCrich}$ in order not to bias estimates in the expected direction.

Impartial spectator experiment: Participants

15 healthy student volunteers who live with their parents participated in this experiment. All participants were native Japanese speakers who were recruited from the same university pool as the MRI participants. None had participated in the gambling experiment. No participant had a history of neurological or psychiatric disease, and written informed consent was obtained from all participants prior to participation. Two participants did not pass an instructional manipulation check (Oppenheimer, Meyvis, and Davidenko 2009) for the task in the initial instruction. They were excluded from analyses, resulting in the final sample of 13 (7 females and 6 males; age: 20.85 ± 1.14 years, range: 19-23 years). The study was approved by the ethical committee of Tamagawa University in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.

Procedures

Participants completed the impartial spectator task outside the MRI scanner.

Impartial spectator task

First, each participant was shown a histogram of the household income distribution of the previous participants for the gambling experiment. Then, they were instructed that "The amount of money you earn will depend on whether you perform the task seriously". After that, they answered the first question: "The sum of the pleasure of the lower half of previous participants each of whom received ¥500 would be balanced by the sum of the pleasure of the upper half, each of whom received ¥X_{500_1}. How much is ¥X_{500_1}?" In the second and third question, ¥400 and ¥600 were used as the amounts of money that the lower half of previous participants received (X_{400}, X_{600}). The order of ¥400 and ¥600 was counterbalanced across the participants. In the fourth question, ¥500 was again used as the amount of money that the lower half of previous participants received (X_{500_2}). Participants whose responses on the first and forth trial were exactly the same earned a bonus of ¥500.

We recruited independent participants for the impartial spectator task from the same university students pool as the MRI participants, and they were incentivized to conduct the task seriously. Two conditions for impartial spectators have been proposed: (1) impartiality requirement: spectators should be third parties whose interests are not affected by their judgements about observees (2) empathy requirement: spectators should have sufficient information and motivation to infer observees' pleasure (Smith 1759; Harsanyi 1977). Our participants were considered impartial and empathetic to the MRI participants, which matched the requirements for impartial spectators originally proposed by Adam Smith (Smith 1759; Harsanyi 1977).

Validation of the interpersonal comparison of utility

Impartial spectators' estimates of the utility difference ratio of the poor group to the rich group were calculated as $X_{ratio} = (X_{500_1}/500 + X_{400}/400 + X_{600}/600 + X_{500_2}/500)/4$. Impartial spectators' estimated values of the total (or mean) utility of the poor group and those of the rich group were assumed to follow normal distributions with a common standard deviation (σ_{IS}). Then, X_{ratio} was assumed to be generated by the following formula:

$$X_{ratio}(k) \sim \text{Normal}(IS_{ratio}, \sigma_{IS})/\text{Normal}(1, \sigma_{IS}) = \text{Fieller-Hinkley}(IS_{ratio}, \sigma_{IS}, 1, \sigma_{IS})$$

where IS_{ratio} is the mean of the estimated utility of the poor group when the mean of the estimated utility of the rich group is given as 1.

In order to check whether impartial spectators' estimation and neural signals coincide, IS_{ratio} and the Δu_{neural} ratio of the poor group to the rich group (PSC_{ratio}) were compared using a Bayesian model with default effect size priors (Cauchy scale $2^{0.5}/2$). Results are reported as BF_{10} that represents P(data| H_1: $PSC_{ratio} \neq IS_{ratio}$)/P(data| H_0: $PSC_{ratio} = IS_{ratio}$). Effect size estimates are also reported as median posterior Cohen's δ with 95% CI. The model is illustrated in Supplementary Fig. 4.

The details of the model are as follows. We first calculated the mean and standard deviation of obtained Δu_{neural} data: $mean_{PSC} = \frac{N_{poor}\Delta u_{neural_poor}+N_{rich}\Delta u_{neural_rich}}{N_{poor}+N_{rich}}$, $SD_{PSC}=\sqrt{\frac{(N_{poor}-1)\times sd(\Delta u_{neural_poor})^2+(N_{rich}-1)\times sd(\Delta u_{neural_rich})^2}{N_{poor}+N_{rich}-2}}$. We used a normal approximation of a Fieller-Hinkley distribution (Díaz-Francés and Rubio 2013) to calculate the standard deviation of Δu_{neural} ratio as $SD_{PSC_ratio} = \sqrt{\frac{mean(\Delta u_{neural_poor})^2}{mean(\Delta u_{neural_rich})^2}\left(\frac{sd(\Delta u_{neural_poor})^2}{mean(\Delta u_{neural_poor})^2}+\frac{sd(\Delta u_{neural_rich})^2}{mean(\Delta u_{neural_rich})^2}\right)}$. Then, we calculated the standard deviation of ratios of Δu_{neural} and impartial spectator data as $SD_{ratio} = \sqrt{\frac{(\frac{N_{poor}+N_{rich}}{2}-1)\times SD_{PSC_ratio}{}^2+(N_{IS}-1)\times sd(X_{ratio})^2}{\frac{N_{poor}+N_{rich}}{2}+N_{IS}-2}}$. Here, $mean(\cdot)$ and $sd(\cdot)$ indicate the sample mean and the unbiased sample variance, respectively.

We used the following priors for the effect size, mean, and standard deviation of Δu_{neural}, the effect size of the difference between Δu_{neural} data and impartial spectator data, and standard deviation of impartial spectators' intuition.

$\delta_{PSC} \sim$ Cauchy(0, $2^{0.5}/2$)

$\mu_{PSC} \sim$ Cauchy($mean_{PSC}$, $2^{0.5}/2 * SD_{PSC}$)

$\sigma_{PSC} \sim$ Half-Cauchy(0, $2^{0.5}/2 * SD_{PSC}$)

$\delta_{ratio} \sim$ Cauchy(0, $2^{0.5}/2$)

$\sigma_{IS} \sim$ Half-Cauchy(0, $2^{0.5}/2$)

Mean difference of Δu_{neural} between the poor group and rich group (α_{PSC}), PSC_{ratio}, and IS_{ratio} was defined as follows.

$\alpha_{PSC} = \delta_{PSC} * \sigma_{PSC}$

$\mu_{PSCpoor} = \mu_{PSC} + \alpha_{PSC}/2$

$\mu_{PSCrich} = \mu_{PSC} - \alpha_{PSC}/2$

$PSC_{ratio} = \mu_{PSCpoor}/\mu_{PSCrich}$

$IS_{ratio} = PSC_{ratio} + \delta_{ratio} * SD_{ratio}$

The last equation is based on the definition of Cohen's δ ($\delta_{ratio} = (IS_{ratio} - PSC_{ratio})/SD_{ratio}$).

The data were assumed to be generated by the following formula. The Δu_{neural} data were assumed to follow the normal distributions.

$\Delta u_{neural_poor}(i) \sim Normal(\mu_{PSCpoor}, \sigma_{PSC})$

$\Delta u_{neural_rich}(j) \sim Normal(\mu_{PSCrich}, \sigma_{PSC})$

Then, X_{ratio} was assumed to be generated by the following formula:

$X_{ratio}(k) \sim Normal(IS_{ratio}, \sigma_{IS})/Normal(1, \sigma_{IS}) = Fieller\text{-}Hinkley(IS_{ratio}, \sigma_{IS}, 1, \sigma_{IS})$

Application of the interpersonal comparison of utility

The policy that distributes ¥x to each participant of the poor group and ¥y to each participant of the rich group was considered from a social planner's perspective. Here, a utilitarian social welfare function was considered because it is the most common when utility differences are interpersonally comparable (d'Aspremont and Gevers 1977; Roberts 1980).

$$W(x, y) = \sum_{i=1}^{N} u_i(x, y)$$

The social welfare increase by the policy for moderate amounts of money was expressed as follows.

$$\Delta W(x, y) = xN_{poor}\mu_{PSCpoor} + yN_{rich}\mu_{PSCrich}$$

Given posterior of $P(\mu_{PSCpoor})$ and $P(\mu_{PSCrich})$ in the Bayesian t-test model that has H_1 without restriction of $\mu_{PSCpoor} \geq \mu_{PSCrich}$, the expected social welfare increase was expressed as follows.

$$E[\Delta W(x,y)] = \int \left(xN_{poor}\mu_{PSCpoor} + yN_{rich}\mu_{PSCrich}\right) P\left(\mu_{PSCpoor}\right)P(\mu_{PSCrich})d\mu_{PSCpoor}d\mu_{PSCrich}$$

$$= xN_{poor}\int \mu_{PSCpoor}\, P\left(\mu_{PSCpoor}\right)d\mu_{PSCpoor} + yN_{rich}\int \mu_{PSCrich}\, P(\mu_{PSCrich})d\mu_{PSCrich}$$

$$= xN_{poor}E\left[\mu_{PSCpoor}\right] + yN_{rich}E[\mu_{PSCrich}]$$

The expected social welfare can be maximized by choosing the option with maximum E[ΔW].

This was used to determine the optimal decision rule for the problem of choosing between the policy of distributing ¥k to all participants and the policy of distributing ¥(k + ρ) to each participant of the poor group. In the case that these two policies generate the same amount of social welfare increase, the following equation holds.

$$E[\Delta W(k,k)] = E[\Delta W(k+\rho,0)]$$

$$k(N_{poor}E\left[\mu_{PSCpoor}\right] + N_{rich}E[\mu_{PSCrich}]) = (k+\rho)N_{poor}E[\mu_{PSCpoor}]$$

$$k(E\left[\mu_{PSCpoor}\right] + E[\mu_{PSCrich}]) = (k+\rho)E[\mu_{PSCpoor}]$$

$$k/\rho = \frac{E[\mu_{PSCpoor}]}{E[\mu_{PSCrich}]}$$

Therefore, the optimal decision rule, for expected utilitarian social welfare maximizer, is

$$\begin{cases} \text{giving ¥}k \text{ to all participants} & (\text{if } k/\rho < E[\mu_{PSCpoor}]/E[\mu_{PSCrich}]) \\ \text{either is fine} & (\text{if } k/\rho = E[\mu_{PSCpoor}]/E[\mu_{PSCrich}]) \\ \text{giving ¥}(k+\rho)\text{to the poor group} & (\text{otherwise}). \end{cases}$$

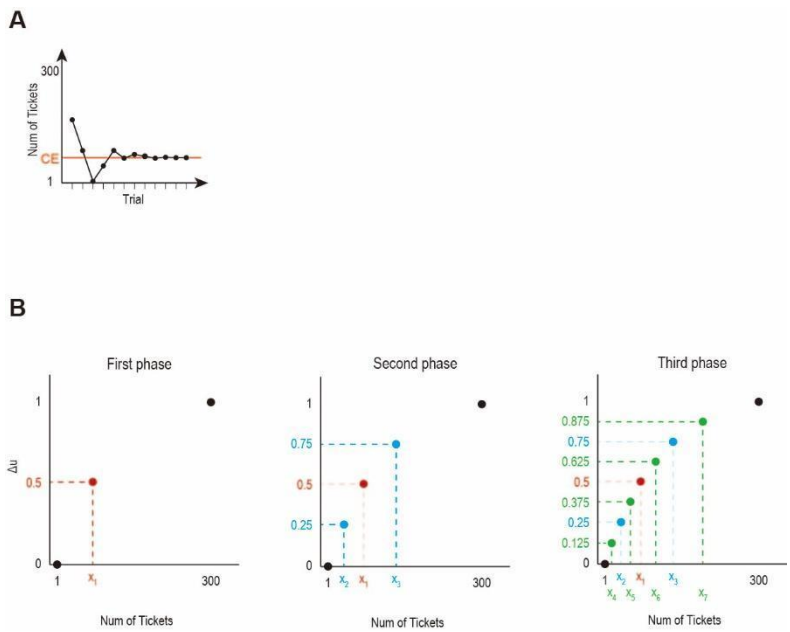
This decision rule was applied to choose between the policy of distributing ¥1,000 to all participants and the policy of distributing ¥1,500 to participants in the poor group only.

Based on neural data, we gave an additional ¥1,500 to all the "poor" group participants who agreed to receive additional compensation (N = 13).

Author contributions Ka.M. conceived the idea, and Ka.M., K.I. and Ke.M. jointly designed the study. Ka.M., Y.Y. and Ke.M. created the experimental materials. Ka.M., K.I. and Ke.M. performed the research. Ka.M. analyzed the data. Ka.M. and Ke.M. jointly wrote the paper. Ka.M., K.I., Y.Y. and Ke.M. provided critical comments.

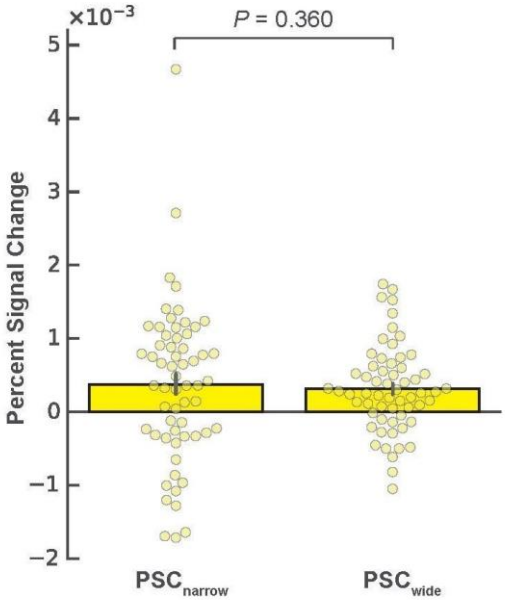
Competing interests The authors declare no competing interests.

Correspondence and requests for materials should be addressed to Ka.M. and Ke.M.



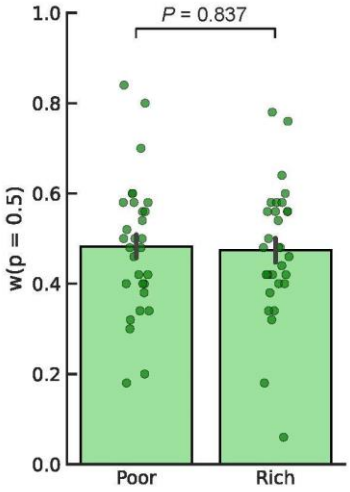
Supplementary Fig. 1.

Measuring utility function. (**A**) Measuring CEs using the PEST procedure. The black trace shows the numbers of tickets for a sure option compared with the 50/50 gamble between 1 and 300 tickets in the food gambling task, for illustration purposes. The lottery loss number (= 1), and lottery win number (= 300) remained unchanged throughout a PEST sequence, whereas the sure number was adjusted on subsequent trials based on the participant's choice between the sure offer and lottery (see Measuring CEs). The procedure lasted until convergence. (**B**) Iterative fractile method for measuring utility in the food gambling task. First, the CE of the 50/50 gamble between 1 and 300 tickets (x1) was measured by using PEST (red point). Then, x_1 was used as an outcome to construct two new gambles (1 ticket, p = 0.5 and x_1 tickets, p = 0.5; x_1 tickets, p = 0.5 and 300 tickets, p = 0.5) and measured their CEs (x_2 and x_3, light blue points). Finally, CEs of 50/50 gamble between 1 and x_2 tickets (x_4), that between x_2 and x_1 tickets (x_5), that between x_1 and x_3 tickets (x_6) and that between x_3 and 300 tickets (x_7) were measured (light green points). It is illustrated for the case of w(p = 0.5) = 0.5 in this figure. If w(p = 0.5) = 0.4, the $\Delta u(x_1)$ is 0.4 rather than 0.5 and so on. w(p = 0.5) was determined by the probability weighting task (see APPENDIX).



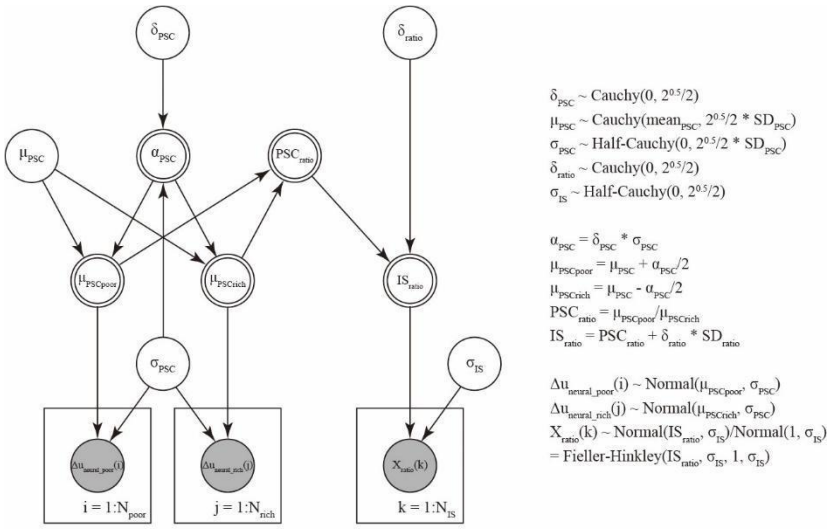
Supplementary Fig. 2.

Neural signals for the utility prediction errors in two reward-range conditions. The activity of the brain region representing utility prediction errors in the food gambling task are shown. There was moderate evidence that the mean percent signal changes for utility prediction errors per 1 yen were not rescaled between the narrow and wide blocks in the monetary gambling task ($t(59) = 0.36$, $P = 0.360$ one-sided, $BF_{+0} = 0.193$, median posterior $\delta = 0.044$ with 95% CI [-0.201, 0.290]). Error bars indicate the SEM across participants.



Supplementary Fig. 3.

Probability weighting of poor and rich groups. Subjective probabilities at p = 0.5 (w(p = 0.5)) measured from the poor and rich group participants in the probability weighting task are shown. The values were not different between the poor and rich groups with moderate evidence (t(58) = 0.207, *P* = 0.837 two-sided, BF_{10} = 0.268, median posterior δ = 0.044 with 95% CI [-0.418, 0.510]). Error bars indicate the SEM across group members.



Supplementary Fig. 4.

The graphical model to assess coincidence between neural signals and impartial spectators' estimations. Shaded and unshaded nodes represent unobserved variables and observed variables, respectively. Single-bordered and double-bordered nodes represent probabilistic and deterministic variables, respectively.