Interpersonal Comparison of Utility by Measuring Neural Activity

Kaosu Matsumori[1,2,3]*, Kazuki Iijima[1,4], Yukihito Yomogida[1,5], Kenji Matsumoto[1]*
[1]Brain Science Institute, Tamagawa University; [2]Department of Social Psychology, The University of Tokyo; [3]Japan Society for the Promotion of Science; [4]National Institute of Mental Health, National Center of Neurology and Psychiatry; [5]National Institute of Neuroscience, National Center of Neurology and Psychiatry

Aggregating welfare across individuals to reach collective decisions is one of the most fundamental problems in our society. Interpersonal comparison of utility is pivotal and inevitable for welfare aggregation, because if each person's utility is not interpersonally comparable, there is no rational aggregation procedure that simultaneously satisfies even some very mild conditions for validity[1–3] (Arrow's impossibility theorem). However, scientific methods for interpersonal comparison of utility have thus far not been available[1,2,4,5]. Here, we have developed a method for interpersonal comparison of utility based on brain signals, by measuring the neural activity of participants performing gambling tasks. We found that activity in the medial frontal region was correlated with changes in expected utility, and that, for the same amount of money, the activity evoked was larger for participants with lower household incomes than for those with higher household incomes. Furthermore, we found that the ratio of neural signals from lower-income participants to those of higher-income participants coincided with estimates of their psychological pleasure by "impartial spectators", i.e. disinterested third-party participants satisfying specific conditions[6,7]. Finally, we derived a decision rule based on aggregated welfare from our experimental data, and confirmed that it was applicable to a distribution problem. These findings suggest that our proposed method for interpersonal comparison of utility enables scientifically reasonable welfare aggregation by escaping from Arrow's impossibility and has implications for the fair distribution of economic goods. Our method can be further applied for evidence-based policy making in nations that use cost-benefit analyses or optimal taxation theory for policy evaluation[8,9].

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[1,2,4,5]. This stems from the history of concepts of utility.

Concepts of utility and interpersonal comparison of utility

Bentham's original sense of utility (psychological utility) is psychological pleasure[10]. Since such psychological utility cannot be directly measured from behaviour, a new concept of utility (decision utility), which can be operationally defined by the agent's choice behaviours, emerged as a surrogate for psychological utility at the beginning of the 20th century[11,12] (Supplementary Information section 1).

Among currently considered decision utilities, "expected utility" is the most informative[13] (Fig. 1a, b, Extended Data Fig. 1). Information from expected utility is fully preserved even after a linear transformation with parameters for different unit width (α_i) and different zero point (β_i) among individuals ($u_i' = \alpha_i u_i + \beta_i$, where i is for individuals). Expected utility can be expressed on a scale from 0 for the worst option to 1 for the best option (0-1 normalisation) without losing any information. This characteristic indicates that expected utility (and other decision utilities) is only *intra*personally, but not *inter*personally, comparable by definition (Fig. 1d, e).

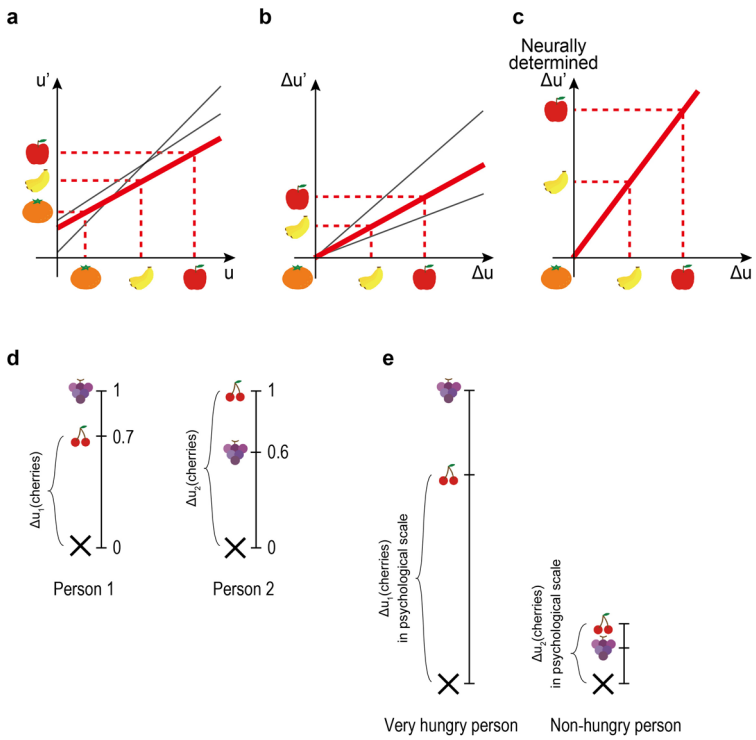


Fig. 1 | Concepts of utility.
a, Utility as an interval scale (expected utility). Each axis represents utilities in which the utility difference between an apple and a banana is the same as the utility difference between a banana and an orange. Because only the distances between the options are meaningful 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.
b, Utility difference as a ratio scale. If the expected 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.
c, 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.
d, As the lengths of the scales for expected utility have no meaning, a 0-1 normalization is applied to express the utilities of the two individuals. Δu_1(cherries) = 0.7 and Δu_2(cherries) = 1, where $\Delta u(\cdot) = u(\cdot) - u(\text{none})$. Then, Δu_1(cherries) < Δu_2(cherries), although this comparison has no empirical meaning.
e, A modified case of **d**. 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. Then, 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 hungry person than that for non-hungry person in contrast to **d**.

Since the concept of decision utility was popularized, interpersonal comparisons of utility have been regarded as neither objective nor scientific[4,5] for nearly a century. 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[2].

Analyses based on decision utility have shown that a competitive equilibrium based on an ideal market can realize a Pareto efficient state (in this state, no one's utility can increase any further without decreasing another's utility) (Extended Data Fig. 2). 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[1–3] (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[8,14] (social cost-benefit analysis). This assumption is unrealistic and contrary to intuitions about interpersonal comparisons[2,8]. 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[4,5]. 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[15] (Supplementary

Information section 1). Satisfying this condition must serve as a starting point for subsequent scientific inquiry[15]. 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[16,17], because it is often pointed out that ordinary people make interpersonal comparisons of utility with relative ease and apparent success[17,18].

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' intuitions. Finally, we applied the interpersonal comparison method to an actual distribution problem.

Neural representation of utility

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 of all, every economic choice behaviour 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. Therefore, 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"[19,20] (Supplementary Information section 2). 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 that encode prediction error of decision utility[21–23] (i.e., the difference between the utility of an obtained outcome and the expected utility of a lottery; Δu).

To examine neural representation of utility, we identified a representation which (a) correlates with the prediction error of decision utility (Δu), not with that of reward amount, (b) is not normalised to an arbitrary range such as 0-1 mentioned above, and (c) 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 Methods in detail).

Because utility for moderate amounts of money is approximately linear intrapersonally[24–26], 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 normalised utility function for food tickets using the fractile method, with considering probability weighting function[23] (see Methods, Extended Data Fig. 3). 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.

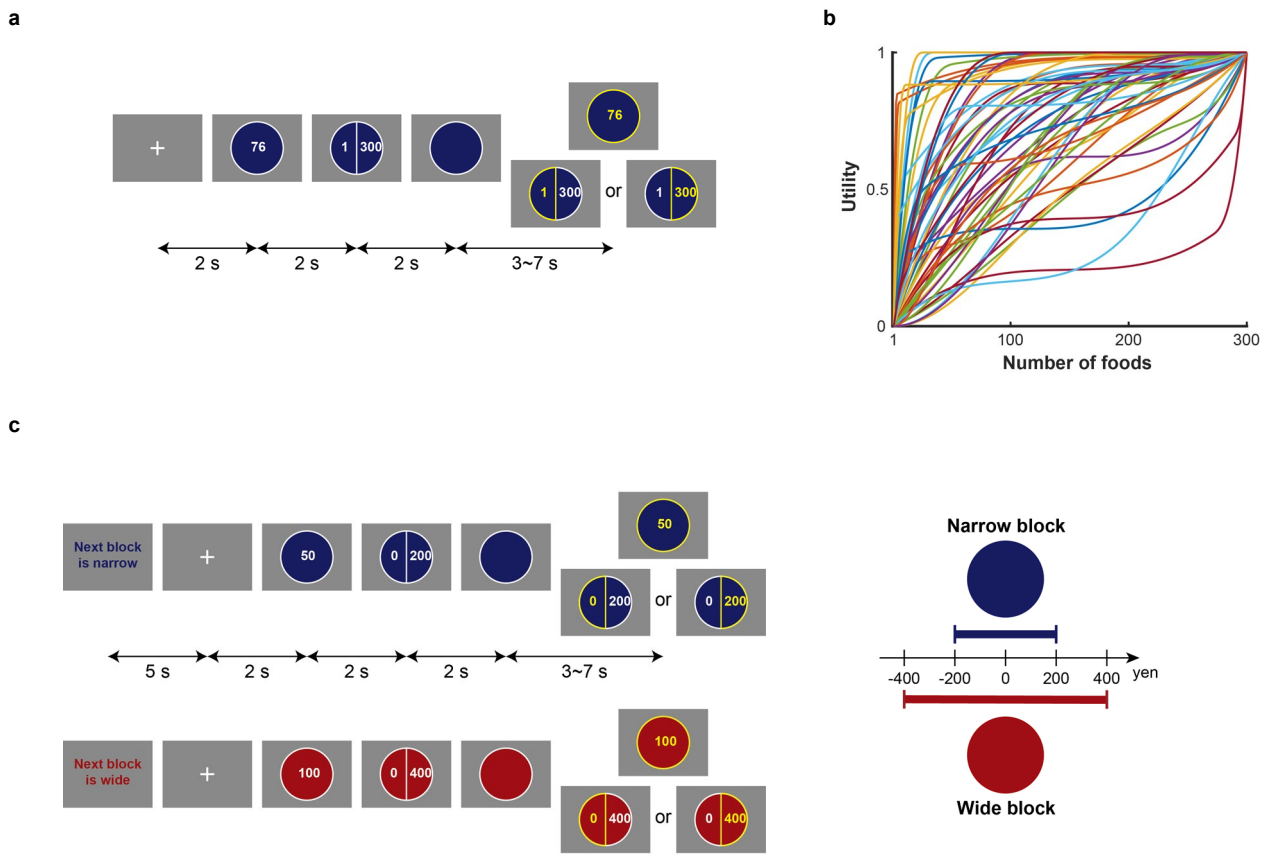


Fig. 2 | Gambling tasks.
a, Food gambling task. Participants were required to choose between a sure (riskless) offer and a lottery to get indicated amounts of snacks as reward.
b, The experimentally obtained utility functions for food amounts. 0-1 normalization was applied to the utility functions.
c, 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.

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 Methods), 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 behaviours were 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' behaviours.

Next, we analysed 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[19,20], 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.

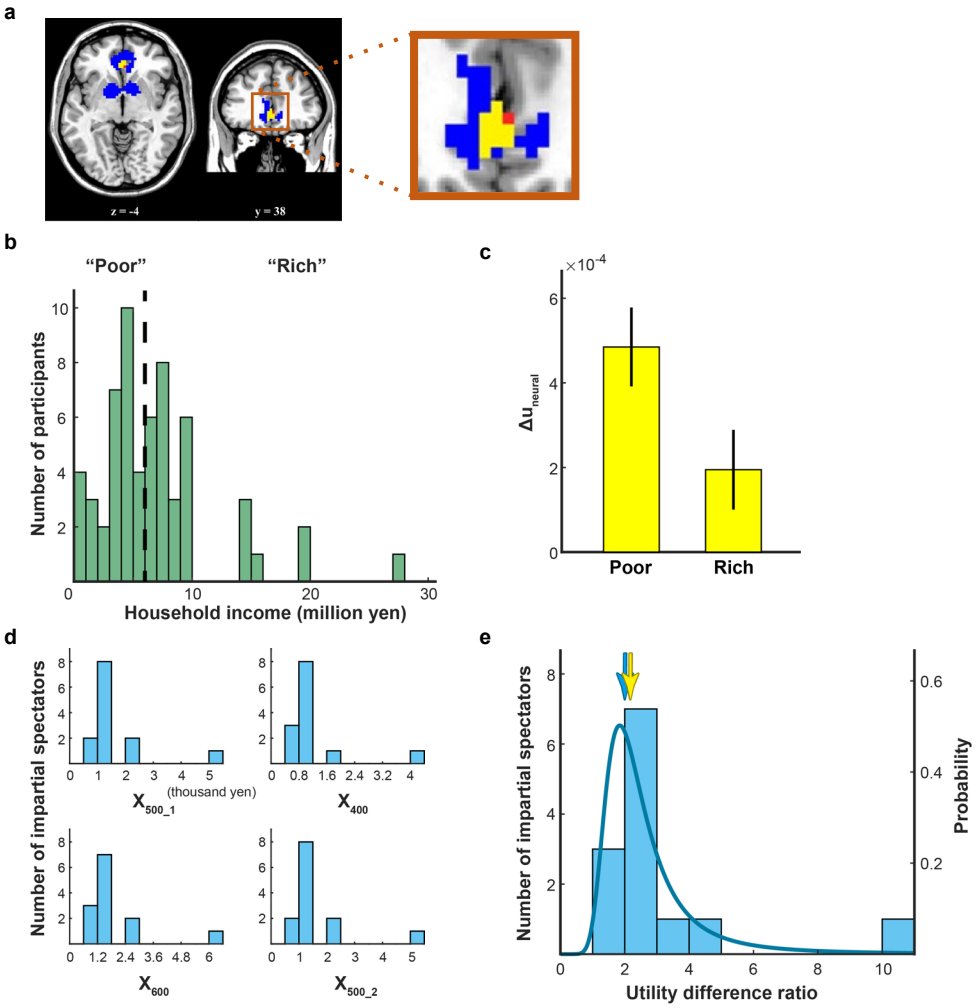


Fig. 3 | Interpersonal comparison of neural measures for utility.

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 named the "utility region".

b, Distribution of annual household incomes of the participants for the gambling tasks. 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).

c, Mean percent signal changes of the utility region per 1 yen (Δu_{neural}) in "poor" and "rich" groups. The poor group showed significantly greater values than the rich group ($t(58) = 2.19$, $p = 0.016$ one-sided, $BF_{+0} = 4.233$). Error bars indicate SEM across group members.

d, The distributions of impartial spectators' estimates of the amounts of money needed to please the poor group equivalently to the rich group, when the participants were asked 4 times (for the rich 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.

e, 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 **d** and its fit curve on the Fieller-Hinkley distribution. The Δu_{neural} ratio of the poor group to the rich group coincided with the impartial spectators' estimates of the utility difference ratio ($BF_{10} = 0.182$). Yellow and light blue arrows indicate the Δu_{neural} ratio and the median of impartial spectators' estimates, respectively.

A utility representation that can objectively determine the distance between the best and worst options requires brain activities whose representation of utility is not normalised across different task conditions. In order to examine whether the activity of the ACC/vmPFC region that correlated with utility prediction errors in the food gambling task normalises its response, we conducted the monetary gambling task in 2 different contexts (Fig. 2c). 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 contexts 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 (¥)). This allowed us to examine whether brain activity was influenced by the range of reward amounts between blocks.

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). A preliminary exploration of the data indicated that there were two outliers among the participants (>3.5SD from the mean value), possibly because of the relatively small number of trials split between the two blocks. Grubbs tests for the outliers were significant (Ps < 0.01 two-sided). Accordingly, we excluded these participants from the block splitting analysis. We found moderate evidence for the absence of normalisation among blocks ($t(57) = 0.19$, $P = 0.42$ one-sided, $BF_{+0} = 0.169$, median posterior $\delta = 0.025$ with 90% CI [-0.185, 0.235]) (GLM3) (Extended Data Fig. 4).

The utility representation for use of interpersonal comparison must represent utility irrespective of varying reward type. We sought brain regions which represent such utility by finding the overlapped region between the neural correlates of utility prediction error for food and money (Fig. 3a). For food, the results of the food gambling task (GLM1) were considered. For money, 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] ($t(59) = 9.02$, $t(59) = 7.65$) in the striatum and at [-6, 38, -4] ($t(59) =6.38$) in the vmPFC) (FWE corrected Ps < 0.001 one-sided) (blue and yellow areas in Fig. 3a). The overlapped region was found in the ACC/vmPFC, hereafter called the "utility region" in this paper (yellow area in Fig. 3a). We confirmed that the activity of "utility region" correlated with utility

prediction errors without discriminating between narrow and wide blocks. Thus, the "utility region" was correlated with the prediction error of utility rather than that of reward amount, in the non-normalised scale, irrespective of reward type.

Interpersonal comparison of utility based on neural signals

Next, we compared the PSCs of the "utility region" (hereafter referred to as 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). We found moderate evidence for a difference in Δu_{neural} ($t(58) = 2.19$, $p = 0.016$ one-sided, $BF_{+0} = 4.233$, median posterior $\delta = 0.486$ with 90% CI [0.081, 0.912]) (Fig. 3c), while the variances of utility prediction errors were not different between the groups (see Supplementary Information section 3, Extended Data Fig. 5). The posterior means of Δu_{neural} across participants, $E[\mu_{PSCpoor}]$ and $E[\mu_{PSCrich}]$, were $4.654 * 10^{-4}$ and $2.143 * 10^{-4}$, respectively ($E[\mu_{PSCpoor}]/E[\mu_{PSCrich}]$ was 2.171). Thus, we were able to objectively determine an *inter*personally comparable scale of utility difference by using Δu_{neural} (Fig. 1c). The scale of utility difference would not be interpersonally comparable if it was based on choice behaviour alone.

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

In order to validate our scientific interpersonal comparison method using Δu_{neural}, we asked an additional 15 participants to complete an impartial spectator task. Two participants were excluded from the following analysis, because they did not pass the instructional manipulation check (see Methods). We considered the remaining 13 participants "impartial spectators" for this task, because they were impartial and empathetic to the MRI participants[6,7] (see Methods and Supplementary Information section 3). 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 Methods). Their answers are shown in Fig. 3d. 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. 3e).

The utility difference ratio of poor group to rich group based on the impartial spectator participants' intuitive estimates was not different from that based on Δu_{neural} with moderate evidence (BF_{10} = 0.182, median posterior δ = 0.033 with 95% CI [-0.449, 0.240]) (Fig. 3e, Extended Data Fig. 6). Thus, the interpersonal comparison method based on MRI signals from the brain was coherent to impartial spectators' intuition.

Application to a distribution problem

We applied the interpersonal comparison method based on Δu_{neural} to a distribution problem concerning whether to distribute ¥1,500 to every poor participant or to every rich participant from the social planner's perspective (Fig. 4a). Here, we took the decision rule which maximizes utilitarian social welfare because the utilitarian social welfare function is the most common when utility differences are interpersonally comparable[27]. According to the ratio of posterior means of Δu_{neural} ($E[\mu_{PSCpoor}]/E[\mu_{PSCrich}] = 2.171$) described above, for the problem of choosing between the policy of distributing ¥k to every poor group participant and the policy of distributing ¥m to every rich group participant, the optimal decision rule, for an expected utilitarian social welfare maximizer, was

$$\begin{cases} \text{giving ¥k to the poor group} & (\text{if } m/k < 2.171) \\ \text{either is fine} & (\text{if } m/k = 2.171) \\ \text{giving ¥m to the rich group} & (\text{otherwise}). \end{cases}$$

Now, m/k (= 1,500/1,500 = 1) was less than the experimentally obtained value (2.171) for the distribution problem. Therefore, we decided to allocate ¥1,500 to each participant of the poor group, and actually conducted it for the person who agreed to receive additional compensation. This made the social welfare increase 2.171 times as much as giving the same amount of money to the rich group participants. Moreover, the decision rule tells us that the social welfare increase by giving ¥3,000 to the rich group would still be less than that by giving ¥1,500 to the poor, but giving ¥4,000 to the rich should exceed it (Fig. 4b-c).

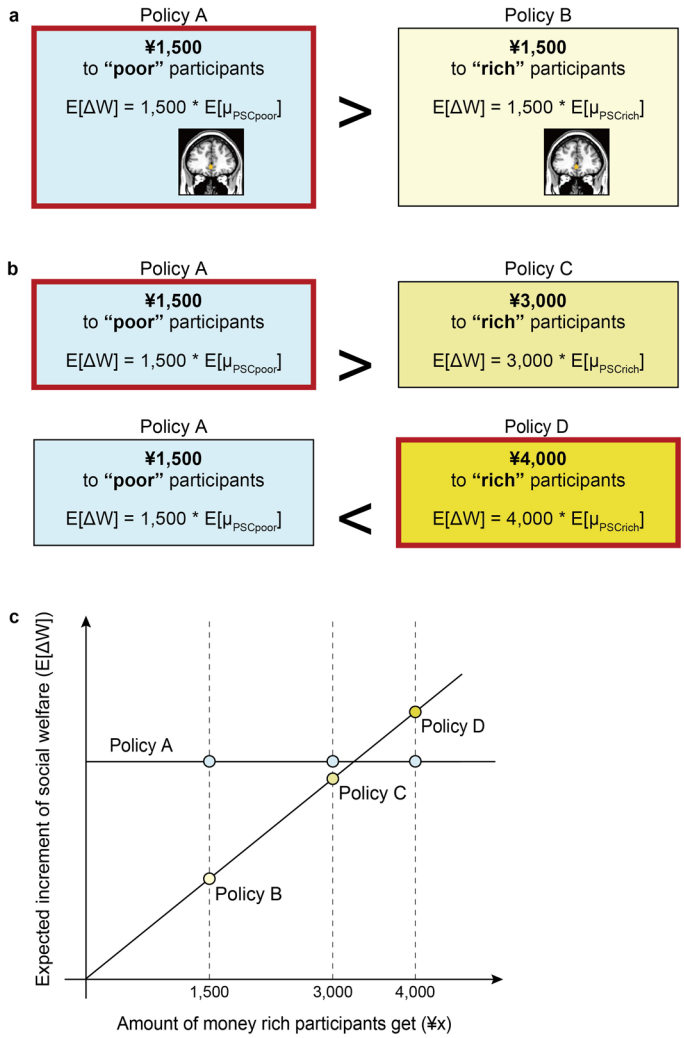


Fig. 4 | Application to a distribution problem.
a, Two distribution policies, A and B, were compared. A was chosen, because it increased expected utilitarian social welfare (E[ΔW]) more than B.
b, Policies C and D are additionally assumed and compared with A. C would not be chosen, but D would be chosen based on their E[ΔW] values.
c, Graphic illustration of expected increments of social welfare by distributing ¥1,500 to "poor" participants and by distributing ¥x to "rich" participants. The compared policies (A vs. B, A vs. C, and A vs. D) are plotted.

Conclusions

We succeeded in developing a method for interpersonal comparison of utility based on the activity from a specific brain region (ACC/vmPFC), which satisfied the 2 conditions required for interpersonal comparison of utility, resting upon scientific demonstration[4,5] and being validated by coherence to impartial spectators' estimation[15] (Supplementary Information section 3). The proposed method is prominent in the academic sense that it enables us to escape from Arrow's impossibility[1,2]. Moreover, it can be adopted for practical applications, such as determining distributional weight for cost-benefit analysis[8] or the ratio of marginal utility for taxpayers to the marginal value of public funds[9] in nations that use cost-benefit analysis or optimal taxation theory for policy evaluation. Thus, we demonstrated a major advance in social science using a neuroscientific study[28]. 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[15] and for philosophical reflections on appropriate social structures[29,30].

References

1. Arrow, K. J. *Social choice and individual values*. (Yale university press, 1963).
2. Sen, A. *Collective choice and social welfare*. (Harvard University Press, 2018).
3. Sen, A. *The Idea of Justice*. (Harvard University Press, 2009).
4. Robbins, L. *An essay on the nature and significance of economic science*. (Ludwig von Mises Institute, 1932).
5. Robbins, L. Interpersonal Comparisons of Utility: A Comment. *Econ. J.* **48**, 635–641 (1938).
6. Smith, A. *The Theory of Moral Sentiments*. (1759).
7. Harsanyi, J. C. Morality and the theory of rational behavior. *Soc. Res.* 623–656 (1977).
8. Boardman, A., Greenberg, D., Vining, A. & Weimer, D. *Cost-Benefit Analysis*. (2018).
9. Saez, E. Using elasticities to derive optimal income tax rates. *Rev. Econ. Stud.* **68**, 205–229 (2001).
10. Bentham, J. *An Introduction to the Principles of Morals and Legislation*. vol. 45 527

(1789).

11. Kreps, D. M. *Microeconomic Foundations I: Choice and Competitive Markets*. (Princeton University Press, 2012).
12. Moscati, I. *Measuring Utility: From the Marginal Revolution to Behavioral Economics*. (Oxford University Press, 2018). doi:10.1093/oso/9780199372768.001.0001.
13. von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behavior*. (Princeton University Press, 2007).
14. HM Treasury. The green book: Central government guidance on appraisal and evaluation. *Lond. HM Treas.* (2018).
15. Chang, H. *Inventing Temperature: Measurement and Scientific Progress*. (Oxford University Press, 2004). doi:10.1093/0195171276.001.0001.
16. Harsanyi, J. C. Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *J. Polit. Econ.* **63**, 309–321 (1955).
17. Rossi, M. Simulation theory and interpersonal utility comparisons reconsidered. *Synthese* **191**, 1185–1210 (2014).
18. Davidson, D. *Problems of Rationality*. (Oxford University Press, 2004). doi:10.1093/0198237545.001.0001.
19. Glimcher, P. W. & Fehr, E. *Neuroeconomics: Decision making and the brain*. (Academic Press, 2013).
20. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage* **76**, 412–427 (2013).
21. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275**, 1593 (1997).
22. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction*. (MIT press, 2018).
23. Stauffer, W. R., Lak, A. & Schultz, W. Dopamine Reward Prediction Error Responses Reflect Marginal Utility. *Curr. Biol.* **24**, 2491–2500 (2014).
24. Wakker, P. & Deneffe, D. Eliciting von Neumann-Morgenstern Utilities When Probabilities Are Distorted or Unknown. *Manag. Sci.* **42**, 1131–1150 (1996).
25. Booij, A. S. & Kuilen, G. A parameter-free analysis of the utility of money for the

general population under prospect theory. *J. Econ. Psychol.* **30**, 651–666 (2009).

26. Abdellaoui, M., Bleichrodt, H. & L'Haridon, O. A tractable method to measure utility and loss aversion under prospect theory. *J. Risk Uncertain.* **36**, 245–266 (2008).
27. d'Aspremont, C. & Gevers, L. Equity and the Informational Basis of Collective Choice. *Rev. Econ. Stud.* **44**, 199–209 (1977).
28. Camerer, C., Loewenstein, G. & Prelec, D. Neuroeconomics: How Neuroscience Can Inform Economics. *J. Econ. Lit.* **43**, 9–64 (2005).
29. Rawls, J. *A Theory of Justice*. (Harvard University Press, 1971). doi:10.2307/j.ctvjf9z6v.
30. Daniels, N. Wide Reflective Equilibrium and Theory Acceptance in Ethics. *J. Philos.* **76**, 256–282 (1979).

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[24–26]. 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 (Extended Data Fig. 3b) (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 analysed, whereas all trials were used to analyse the behavioural 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.

455 **Behavioural analyses: Measuring CEs**

456 To measure CEs in the food gambling and probability weighting tasks, we used PEST[23,31]
457 (Extended Data Fig. 3a).

458 We assessed the number of tickets or the amount of money in the sure offer that was
459 subjectively equivalent to the value associated with each lottery offer. The rules
460 governing the PEST procedure were adapted from Luce (2000). As we noted above, each
461 trial consisted of three numbers (sure number (SU), lottery win number (LW), and lottery
462 lose number (LL)). Each PEST sequence consisted of several trials during which one
463 constant lottery offer (constant LW and LL) was presented as a choice option against the
464 sure offer (variable SU). On the initial trial of a PEST sequence, SU was determined by
465 rounding (LW + LL)/2 to the nearest integer. Based on the participant's choice between
466 the sure offer and the lottery, SU was adjusted on the subsequent trial. If the participant
467 chose the lottery offer on trial t, then SU was increased by ε on trial t + 1. However, if the
468 participant chose the sure offer on trial t, SU was reduced by ε on trial t + 1. Initially, ε
469 was determined by rounding (LW + LL)/4 to the nearest integer. ε was adjusted according
470 to the following rules. Rule 1: after the third trial of a PEST sequence, every time the
471 participant switched from one option to the other, the size of ε was halved. Rule 2: if
472 updated SU based on ε was equal to LW or LL, the size of ε was halved and then SU was
473 updated. Thus, the procedure converged by locating subsequent sure offers on either side
474 of the true indifference value and reducing ε until the interval containing the indifference
475 value was small (1 for tickets and 10 for money). When the updated SU was the same as
476 the previous SU, the PEST procedure terminated.

477

478 **Constructing Utility Functions in the food gambling task**

479 We determined each participant's utility function for tickets in the range between 1 and
480 300 tickets using the fractile method based on 50/50 gambles[23] (one example fractile
481 procedure is shown in Extended Data Fig. 3b). To do so, we first measured the CE of a
482 50/50 gamble between 1 and 300 tickets (x_1) using PEST. We then used x_1 as an outcome
483 to construct two new gambles (1 ticket, p = 0.5 and x_1 tickets, p = 0.5; x_1 tickets, p = 0.5
484 and 300 tickets, p = 0.5) and measured their CEs (x_2 and x_3). Finally, we measured the
485 CEs of a 50/50 gamble between 1 and x_2 tickets (x_4), that between x_2 and x_1 tickets (x_5),
486 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)[23], 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[32] and cumulative prospect theory[33], because these two theories coincide on the domain of two-outcome lotteries[34].

The expected utility of a lottery, with considering probability weighting function, can be calculated as $E[lottery] = w(p = 0.5) * u(LW) + (1 - w(p = 0.5)) * u(LL)$.

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 $2^{0.5}/2$)[35] (Supplementary Information section 3, Extended Data Fig. 5). 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 $p < 0.05$)[35,36].

Bayesian t-test results are reported using the two-tailed Bayes factor BF_{10} that represents $P(\text{data}|\ H_1: \mu_{wpoor} \neq \mu_{wrich})/P(\text{data}|\ H_0: \mu_{wpoor} = \mu_{wrich})$, 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)[23]. 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 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 normalisation 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 analyse 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

551 hypothesis focusing on the activations of the reward-related regions, the second-level
552 analysis was performed for the voxels within the explicit anatomical mask that is
553 manually drawn to include the whole medial prefrontal cortex, whole caudate head and
554 body, whole putamen, and whole pallidum. Unless otherwise stated, the statistical
555 threshold was set at $p < 0.05$, family-wise error (FWE) corrected, $k = 10$.
556 First, we analysed the data from the food gambling task to determine brain regions which
557 correlate with estimated utility (GLM1) and amount of food (GLM2). GLM1 included 5
558 parametric regressors (1) the utility of the sure offer, (2) the expected utility of the lottery
559 offer weighted by subjective probability ($w(p = 0.5)$), (3) the utility of the chosen option
560 weighted by subjective probability, (4) the utility of sure feedback, and (5) the utility
561 prediction error of lottery feedback weighted by subjective probability, at each event
562 onset. Onsets of each trial per se and six motion parameters were treated as regressors of
563 no interest. Similarly, GLM2 included 5 parametric regressors (1) the amount of the sure
564 offer, (2) the expected amount of the lottery offer weighted by subjective probability, (3)
565 the expected amount of the chosen option weighted by subjective probability, (4) the
566 amount of sure feedback, and (5) the amount prediction error of lottery feedback weighted
567 by subjective probability, at each event onset. In addition, the same regressors of no
568 interest as those in GLM1 were included. We were able to identify the brain region whose
569 activity correlated with the utility prediction error (from GLM1), but not with the amount
570 prediction error (from GLM2).
571 Second, in order to determine whether the neural representation of utility prediction error
572 identified by GLM1 was normalised by the change in width of reward across blocks, we
573 analysed the data from the monetary gambling task (GLM3). GLM3 included the
574 following parametric regressors (1) the monetary amount of the sure offer, (2) the
575 expected monetary amount of the lottery offer weighted by subjective probability, (3) the
576 monetary amount of the chosen option weighted by subjective probability, (4) the
577 monetary amount of sure feedback, (5) the amount prediction error of lottery feedback
578 weighted by subjective probability in the narrow block, and (6) the amount prediction
579 error of lottery feedback weighted by subjective probability in the wide block, at each
580 event onset. In addition, the same regressors of no interest as those in GLM1 were
581 included. We found that the neural representation of utility prediction error (= amount
582 prediction error in monetary gambling task) identified by the analysis with GLM1 was

not normalised 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/)[37] 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 $2^{0.5}/2$)[35]. Outliers were detected using the repeated Grubbs test[38]. 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 90% CI using an H_1 without restriction of $\mu_{PSCnarrow} \geq \mu_{PSCwide}$ in order not to bias estimates in the expected direction.

The PSCs of the "utility region" (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 $P(\text{data}|\ H_{+}: \mu_{PSCpoor} > \mu_{PSCrich})/P(\text{data}|\ H_{0}: \mu_{PSCpoor} = \mu_{PSCrich})$, 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 90% CI using an H_1 without restriction of $\mu_{PSCpoor} \geq \mu_{PSCrich}$ in order not to bias estimates in the expected direction.

615 **Impartial spectator experiment: Participants**

616 15 healthy student volunteers who live with their parents participated in this experiment.
617 All participants were native Japanese speakers who were recruited from the same
618 university pool as the MRI participants. None had participated in the gambling
619 experiment. No participant had a history of neurological or psychiatric disease, and
620 written informed consent was obtained from all participants prior to participation. Two
621 participants did not pass an instructional manipulation check[39] for the task in the initial
622 instruction. They were excluded from analyses, resulting in the final sample of 13 (7
623 females and 6 males; age: 20.85 ± 1.14 years, range: 19-23 years). The study was
624 approved by the ethical committee of Tamagawa University in accordance with the
625 ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.

626

627 **Procedures**

628 Participants completed the impartial spectator task outside the MRI scanner.

629

630 **Impartial spectator task**

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

643

644 **Validation of the interpersonal comparison of utility**

645 Impartial spectators' estimates of the utility difference ratio of the poor group to the rich
646 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 $P(\text{data}|\ H_1: PSC_{ratio} \neq IS_{ratio})/P(\text{data}|\ H_0: PSC_{ratio} = IS_{ratio})$. Effect size estimates are also reported as median posterior Cohen's δ with 95% CI. The model is illustrated in Extended Data Fig. 6.

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[40] 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.

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

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

$\sigma_{PSC} \sim$ 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($\mu_{PSCpoor}$, σ_{PSC})

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

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

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

Application of the interpersonal comparison of utility

The policy that distributes ¥k to each participant of the poor group and ¥m 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[27,41].

$$W(k,m) = \sum_{i=1}^{N} u_i(k,m)$$

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

$$\Delta W(k,m) = kN_{poor}\mu_{PSCpoor} + mN_{rich}\mu_{PSCrich}$$

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.

$$E[\Delta W(k,m)] = \int \big(kN_{poor}\mu_{PSCpoor} + mN_{rich}\mu_{PSCrich}\big)\, P\big(\mu_{PSCpoor}\big)P(\mu_{PSCrich})d\mu_{PSCpoor}d\mu_{PSCrich}$$

$$= kN_{poor}\int \mu_{PSCpoor}\, P\big(\mu_{PSCpoor}\big)d\mu_{PSCpoor} + mN_{rich}\int \mu_{PSCrich}\, P(\mu_{PSCrich})d\mu_{PSCrich}$$

$$= kN_{poor}E\big[\mu_{PSCpoor}\big] + mN_{rich}E[\mu_{PSCrich}]$$

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 each participant of the poor group and the policy of distributing ¥m to each participant of the rich group. In the case that these two policies generate the same amount of social welfare increase, the following equation holds.

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

$$kN_{poor}E\big[\mu_{PSCpoor}\big] = mN_{rich}E[\mu_{PSCrich}]$$

$$\frac{m}{k} = \frac{N_{poor}E[\mu_{PSCpoor}]}{N_{rich}E[\mu_{PSCrich}]} = \frac{E[\mu_{PSCpoor}]}{E[\mu_{PSCrich}]}$$

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

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

This decision rule was applied to choose between the policy of distributing ¥1,500 to the poor group and the policy of distributing ¥1,500 to the rich group.

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

Data availability

The PSC data and impartial spectators' behavioural data supporting this study are available in the OSF repository, https://osf.io/bdwx4/. The gambling participants' behavioural data supporting this study are available from the corresponding author upon reasonable request. The anatomical mask and statistical fMRI maps associated with this study are available in the NeuroVault repository,

https://neurovault.org/collections/YVHREMBI/.

Code availability

The custom R and Stan code supporting the results of this study is available in the OSF repository, https://osf.io/bdwx4/. The custom Matlab code that supports the findings of this study is available from the corresponding author upon reasonable request.

Methods references

31. Luce, R. D. *Utility of gains and losses: Measurement-theoretical and experimental approaches.* xii, 331 (Lawrence Erlbaum Associates Publishers, 2000).
32. Kahneman, D. & Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47**, 263–91 (1979).
33. Tversky, A. & Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).
34. Abdellaoui, M., Bleichrodt, H. & Paraschiv, C. Loss Aversion Under Prospect Theory: A Parameter-Free Measurement. *Manag. Sci.* **53**, 1659–1674 (2007).
35. Keysers, C., Gazzola, V. & Wagenmakers, E.-J. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nat. Neurosci.* **23**, 788–799 (2020).
36. Lee, M. D. & Wagenmakers, E.-J. *Bayesian cognitive modeling: A practical course.* xiii, 264 (Cambridge University Press, 2013). doi:10.1017/CBO9781139087759.
37. Gläscher, J. Visualization of Group Inference Data in Functional Neuroimaging. *Neuroinformatics* **7**, 73–82 (2009).
38. Grubbs, F. E. Sample criteria for testing outlying observations. *Ann. Math. Stat.* **21**, 27–58 (1950).
39. Oppenheimer, D. M., Meyvis, T. & Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* **45**, 867–872 (2009).
40. Díaz-Francés, E. & Rubio, F. On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. *Stat. Pap.* **54**, 309–323 (2013).
41. Roberts, K. W. S. Interpersonal Comparability and Social Choice Theory. *Rev. Econ.*

Stud. **47**, 421–439 (1980).

Acknowledgements 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.

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. analysed 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.

Additional information

Supplementary information

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

Supplementary Information

Supplementary Information section 1: detailed introduction of utility concepts

Bentham's original sense of utility (psychological utility) is psychological pleasure[10]. For example, if one person's pleasure when eating an apple is greater than when eating an orange, then her psychological utility for the apple is greater than her utility 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[17,18]. 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 by behaviour. Therefore, along with the rise of logical positivism and behaviourism, the new concept of utility (decision utility), which can be operationally defined by the agent's behaviour, emerged as a surrogate for the 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[11,12].

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 by agent's choice behaviour. Thus, decision utility was considered an ordinal scale because only the relationship "prefer to" was available (ordinal utility) (Extended Data Fig. 1).

Then, 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[1,13]. 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[15,42]. 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[15]. Temperature scales which are based on the expansion of fluids have two free parameters for unit width (α) and zero point (β), and are unique up to positive linear transformation ($x' = \alpha x + \beta$, $\alpha > 0$) mathematically. The Celsius and Fahrenheit scales have different unit intervals and zero points, and they are

commensurable with a linear transformation. (Subsequent developments of thermodynamics even made temperature a ratio scale (Kelvin scale)).

von Neumann and Morgenstern developed the concept of the utility function on lotteries as an interval scale[13,43] (cardinal utility) (Fig. 1a). 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. 1a). Cardinal utility is also unique up to linear transformation, and any scale based on cardinal utility is commensurable with a linear transformation. 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[23,44–47].

However, this kind of utility is only based on each individual's choice behaviour 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.1a, b, d, e). 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). 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 they have been taboo for nearly a century since the concept of decision utility was popularized[4,5]. 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[2].

Supplementary Information section 2: introduction for neuroscientific studies

A natural and scientific extension of decision utility would be to use neural data[48]. First of all, every economic choice behaviour that reveals decision utility consists of voluntary muscle movements, which are generated by the signals originating in the brain. Moreover, mental states, including psychological utility, are realized by brain states. Therefore, psychological utility should reflect the activities of the responsible brain areas.

Neuroscientific studies using rodents, monkeys, and humans have repeatedly shown that damage to specific brain regions, such as the ventral/medial prefrontal cortices, the ventral striatum, and the midbrain dopaminergic nuclei, causes impairments in choice behaviours[49–52]. Also, a recent study reported that electrical stimulation to the ventral prefrontal cortex caused changes in choice behaviours of monkeys[53]. Furthermore, functional neuroimaging studies on human participants have shown that the activities of these regions represent the value of various rewards (e.g. food, money, and social praise) on a single common scale[19,20,54,55]. These areas are called the reward-related brain regions or value system. Moreover, the activities of the reward-related regions are less responsive to food when not hungry[56,57], and less responsive to monetary rewards when the reward is delayed[58,59], 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 activities in the reward-related regions. One study using monkey subjects showed that neuronal activity in the midbrain dopaminergic nuclei, which project to the reward-related regions, encoded utility difference (Δu)[23]. 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 behaviour alone (Fig. 1c). 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).

Some previous studies have reported that the neural responses to rewards were normalised across contexts in some reward-related regions[60]. Such responses should provide relative scales to represent utilities. However, an absolute scaled representation of utilities is demanded for coherent comparison of utilities. Therefore, neural representations for interpersonal comparison of utility require brain activities that identify the utility differences on an absolute scale. Some previous studies actually reported that certain brain regions represented values without normalisation of their activity to the specific context[61–64].

Supplementary Information section 3: supplementary discussion

In order to extract the brain activity correlated with utility, we measured functional MR

878 signals from participants with different household incomes while they were performing
879 monetary and food gambling tasks. From a separate group of participants, we measured
880 estimates of the relative pleasure of the MRI participants with higher household incomes
881 and those with lower household incomes to validate the comparability of the brain activity
882 between the "rich" and "poor" MRI participants. We obtained the following findings.
883 First, we showed that the activity of ACC/vmPFC correlated with the prediction error of
884 utility rather than that of reward amount. Second, we showed that the activity of this
885 region was independent of specific contexts (i.e., the range of reward amounts that
886 changed between blocks). Third, we showed that Δu_{neural} was larger for the poor group
887 than for the rich group. Fourth, we showed that the ratio of Δu_{neural} coincided with the
888 impartial spectators' intuitive estimates. Finally, we applied the interpersonal comparison
889 method to solve an actual distribution problem and derived a social welfare maximizing
890 decision rule.
891 Our finding that the activity of ACC/vmPFC correlated with the prediction error of utility
892 rather than that of objective amount of reward is consistent with many previous studies
893 that indicate vmPFC encodes subjective value[19,20]. However, in contrast to these previous
894 studies, the striatum was not extracted in our study. This may be due to differences
895 between our methods and theirs. We extracted brain activity that correlated, during the
896 food gambling task in particular, with utility defined by a standard economic method (i.e.
897 the fractile method). Therefore, it is possible that the ACC/vmPFC represents the
898 prediction error of utility but the striatum does not. This interpretation is consistent with
899 the finding that striatum activation was apparently correlated with the prediction error of
900 utility in the monetary gambling task; however, it was not distinguishable from the
901 prediction error of reward amount because of their linearity[65]. Moreover, some previous
902 studies also suggested that the ACC/vmPFC represents more subjective information than
903 the striatum[66,67].
904 Stauffer and colleagues reported that midbrain dopamine neurons encoded the prediction
905 error of utility[23]. Because the dopamine neurons are known to directly project to the
906 ACC/vmPFC[68,69], such dopaminergic inputs may underlie the information about
907 prediction error of utility in ACC/vmPFC we reported in the present study.
908 If the ACC/vmPFC activity represents information that can determine the utility
909 difference between options depending on individual mental states, it should discriminate

910 the utility differences without normalisation when the reward range changes across blocks.
911 Previous studies reported that subjective value signals exhibit range-dependent
912 normalisation that contributes to efficient coding of a wide range of possible reward
913 values by neurons with limited dynamic range[60,61,70–72]. On the other hand, certain brain
914 regions represented subjective value without normalisation which contributes to
915 preserving transitivity across the wide range of possible reward values[61–64]. In the present
916 study, we found that Δu_{neural} did not exhibit range-dependent normalisation in the
917 monetary gambling task, in which the blocks with narrow (¥-200~200) and wide (¥-
918 400~400) ranges of rewards alternated every 20 trials. This suggests that the ACC/vmPFC
919 contributes to preserving the transitivity across different ranges of reward values. Thus,
920 the ACC/vmPFC seems to play an important role in determining the scale of utility
921 difference, although we cannot rule out the possibility that the ACC/vmPFC exhibits
922 range-dependent normalisation when the reward ranges are outside of the conditions in
923 our experiment or each block lasts much longer.
924 A recent study argued that inter-individual (or inter-group) differences in fMRI studies
925 are difficult to interpret if the variances of explanatory variables were different between
926 groups[73]. However, that is not the case in the present study. In the monetary gambling
927 task, the variances of the explanatory variable (i.e., prediction error of utility) were not
928 different between poor and rich groups, because all offers were common among
929 participants, and risk attitudes (i.e., the value of probability weighting function when p =
930 0.5) were not different between the poor and rich groups (Extended Data Fig. 5).
931 Therefore, we would like to propose ACC/vmPFC activity as a potent candidate for an
932 objective signal that enables the interpersonal comparison of utility.
933 Our method for interpersonal comparison of utility rests upon scientific demonstration[4,5],
934 and it was validated by human intuition[7,17,74]. Two conditions for impartial spectators
935 have been proposed: (1) impartiality requirement: spectators should be third parties whose
936 interests are not affected by their judgements about observees (2) empathy requirement:
937 spectators should have sufficient information and motivation to infer observees'
938 pleasure[6,7]. We recruited independent participants for the impartial spectator task from
939 the same university students pool as the MRI participants, and they were incentivized to
940 conduct the task seriously (see Methods). Therefore, our participants were considered
941 impartial and empathetic to the MRI participants, which matched the requirements for

942 impartial spectators originally proposed by Adam Smith[6,7].

943 Scientific measurement methods are necessary for the interpersonal comparison of utility.
944 If an interpersonal comparison method is solely based on the impartial spectators'
945 intuition, it is unlikely to work well when judgments are polluted by systematic biases[75].
946 Without scientific measurement methods, we cannot determine whether systematic biases
947 are operating. Furthermore, non-scientific interpersonal comparison methods based on
948 empathy may work in a small group but not in a large society, because empathy is too
949 information-intensive to reach people living in different cultures[76,77]. Our new
950 interpersonal comparison method based on a scientific measurement (human brain
951 imaging) may serve to extend empathy and be applicable to a larger society.

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

958 In the present study, we found a neural representation of utility prediction error. Because
959 utility functions are usually defined on states but not on processes in a certain context,
960 some readers may think that the neural representation of utility is also a function on the
961 states. However, previous studies have shown that the activity of the reward system
962 including vmPFC is modulated not only by outcome states[20] but also by processes like
963 opportunity equality[66]. Thus, it is not likely that utility-related information in the brain is
964 limited to the states usually considered in utility functions.

965 In the present study, we applied univariate analyses to read utility signals, because they
966 were not only more straightforward than multivariate analyses but also seemed more
967 efficient for detecting signals that are not normalised across contexts according to a
968 previous study[62]. However, there may be a multivariate analysis that improves the
969 accuracy of reading out utility signals[78].

970 In the present study, we focused on utility difference, not on utility level, in comparing
971 utility interpersonally, because it could be treated quantitatively in the experiment. By
972 comparing utility differences interpersonally, we actually conducted a redistribution to
973 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[79]). Utility level is also important to realize a full version of interpersonal comparison of utility[27,41]. Future studies should conduct interpersonal comparison of utility level, which enables distributions according to variable principles discussed in modern justice theories other than utilitarianism[27,41,80].

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 pool 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 the limited participants in the present study.

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[15,42,81]. This is the case for constructing an interpersonally comparable utility as well. We validated the interpersonal comparison method by showing coherence between objective measures (i.e. fMRI signals) and human intuition[7,17,74], as was done for the temperature scale in the initial stage[15] (Supplementary Information section 1). We believe that this serves as a starting point for subsequent scientific inquiries to standardize the measurement method for utility as an interpersonally comparable quantity.

Supplementary references

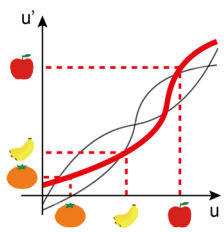
42. Tal, E. Measurement in Science. in *Stanford Encyclopedia of Philosophy* (ed. Zalta, E.) (2015).
43. Krantz, D., Luce, D., Suppes, P. & Tversky, A. *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*. (New York Academic Press, 1971).
44. Luce, R. D. *Individual choice behavior.* xii, 153 (John Wiley, 1959).
45. Waldner, I. The Empirical Meaningfulness of Interpersonal Utility Comparisons. *J. Philos.* **69**, 87 (1972).
46. Stott, H. P. Cumulative prospect theory's functional menagerie. *J. Risk Uncertain.* **32**, 101–130 (2006).
47. Ratcliff, R., Smith, P. L., Brown, S. D. & McKoon, G. Diffusion Decision Model: Current Issues and History. *Trends Cogn. Sci.* **20**, 260–281 (2016).

48. Narens, L. & Skyrms, B. *The Pursuit of Happiness: Philosophical and Psychological Foundations of Utility*. (Oxford University Press, 2020). doi:10.1093/oso/9780198856450.001.0001.
49. Camille, N., Griffiths, C. A., Vo, K., Fellows, L. K. & Kable, J. W. Ventromedial Frontal Lobe Damage Disrupts Value Maximization in Humans. *J. Neurosci.* **31**, 7527 (2011).
50. Butter, C. M., McDonald, J. A. & Snyder, D. R. Orality, Preference Behavior, and Reinforcement Value of Nonfood Object in Monkeys with Orbital Frontal Lesions. *Science* **164**, 1306 (1969).
51. Cardinal, R. N. *et al.* Impulsive Choice Induced in Rats by Lesions of the Nucleus Accumbens Core. *Science* **292**, 2499 (2001).
52. Frank, M. J., Seeberger, L. C. & O'Reilly, R. C. By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science* **306**, 1940 (2004).
53. Ballesta, S., Shi, W., Conen, K. E. & Padoa-Schioppa, C. Values encoded in orbitofrontal cortex are causally related to economic choices. *Nature* **588**, 450–453 (2020).
54. Levy, D. J. & Glimcher, P. W. The root of all value: a neural common currency for choice. *Decis. Mak.* **22**, 1027–1038 (2012).
55. Izuma, K., Saito, D. N. & Sadato, N. Processing of Social and Monetary Rewards in the Human Striatum. *Neuron* **58**, 284–294 (2008).
56. Gottfried, J. A., O'Doherty, J. & Dolan, R. J. Encoding Predictive Reward Value in Human Amygdala and Orbitofrontal Cortex. *Science* **301**, 1104 (2003).
57. Cassidy, R. M. & Tong, Q. Hunger and Satiety Gauge Reward Sensitivity. *Front. Endocrinol.* **8**, 104 (2017).
58. McClure, S. M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science* **306**, 503 (2004).
59. Kable, J. W. & Glimcher, P. W. The neural correlates of subjective value during intertemporal choice. *Nat. Neurosci.* **10**, 1625–1633 (2007).
60. Louie, K., Glimcher, P. W. & Webb, R. Adaptive neural coding: from biological to behavioral decision-making. *Neuroeconomics* **5**, 91–99 (2015).
61. Kobayashi, S., Pinto de Carvalho, O. & Schultz, W. Adaptation of Reward Sensitivity in Orbitofrontal Neurons. *J. Neurosci.* **30**, 534 (2010).

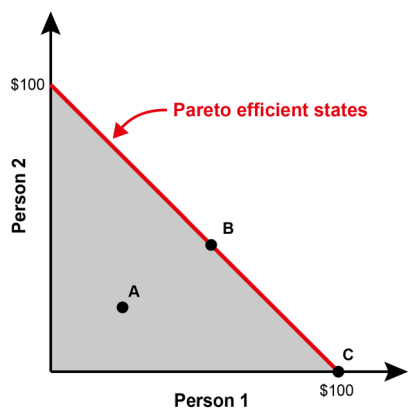
1038 62. Burke, C. J., Baddeley, M., Tobler, P. N. & Schultz, W. Partial Adaptation of
1039 Obtained and Observed Value Signals Preserves Information about Gains and Losses.
1040 *J. Neurosci.* **36**, 10016 (2016).
1041 63. Grabenhorst, F. & Rolls, E. T. Different representations of relative and absolute
1042 subjective value in the human brain. *NeuroImage* **48**, 258–268 (2009).
1043 64. Padoa-Schioppa, C. & Assad, J. A. The representation of economic value in the
1044 orbitofrontal cortex is invariant for changes of menu. *Nat. Neurosci.* **11**, 95–102
1045 (2008).
1046 65. Seymour, B., Daw, N., Dayan, P., Singer, T. & Dolan, R. Differential Encoding of
1047 Losses and Gains in the Human Striatum. *J. Neurosci.* **27**, 4826 (2007).
1048 66. Aoki, R. *et al.* Social Equality in the Number of Choice Options Is Represented in the
1049 Ventromedial Prefrontal Cortex. *J. Neurosci.* **34**, 6413 (2014).
1050 67. Murayama, K. *et al.* How Self-Determined Choice Facilitates Performance: A Key
1051 Role of the Ventromedial Prefrontal Cortex. *Cereb. Cortex* **25**, 1241–1251 (2015).
1052 68. Lewis, D. A., Foote, S. L., Goldstein, M. & Morrison, J. H. The dopaminergic
1053 innervation of monkey prefrontal cortex: a tyrosine hydroxylase
1054 immunohistochemical study. *Brain Res.* **449**, 225–243 (1988).
1055 69. Gaspar, P., Berger, B., Febvret, A., Vigny, A. & Henry, J. P. Catecholamine
1056 innervation of the human cerebral cortex as revealed by comparative
1057 immunohistochemistry of tyrosine hydroxylase and dopamine-beta-hydroxylase. *J.*
1058 *Comp. Neurol.* **279**, 249–271 (1989).
1059 70. Rangel, A. & Clithero, J. A. Value normalization in decision making: theory and
1060 evidence. *Decis. Mak.* **22**, 970–981 (2012).
1061 71. Tobler, P. N., Fiorillo, C. D. & Schultz, W. Adaptive Coding of Reward Value by
1062 Dopamine Neurons. *Science* **307**, 1642 (2005).
1063 72. Polanía, R., Woodford, M. & Ruff, C. C. Efficient coding of subjective value. *Nat.*
1064 *Neurosci.* **22**, 134–142 (2019).
1065 73. Lebreton, M., Bavard, S., Daunizeau, J. & Palminteri, S. Assessing inter-individual
1066 differences with task-related functional neuroimaging. *Nat. Hum. Behav.* **3**, 897–905
1067 (2019).
1068 74. Goldman, A. I. Simulation and Interpersonal Utility. *Ethics* **105**, 709–726 (1995).
1069 75. Müller-Lyer, F. C. Optische urteilstauschungen. *Arch. Anat. Physiol. Physiol. Abt.* **2**,

263–270 (1889).
76. Elster, J. & Roemer, J. E. *Interpersonal Comparisons of Well-Being*. (Cambridge University Press, 1991). doi:10.1017/CBO9781139172387.
77. Bloom, P. *Against empathy: The case for rational compassion*. (Ecco, 2016).
78. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
79. Hume, D. *A treatise of human nature*. (Courier Corporation, 2003).
80. Hirose, I. *Egalitarianism*. (Routledge, 2014).
81. Tal, E. Old and New Problems in Philosophy of Measurement. *Philos. Compass* **8**, 1159–1173 (2013).

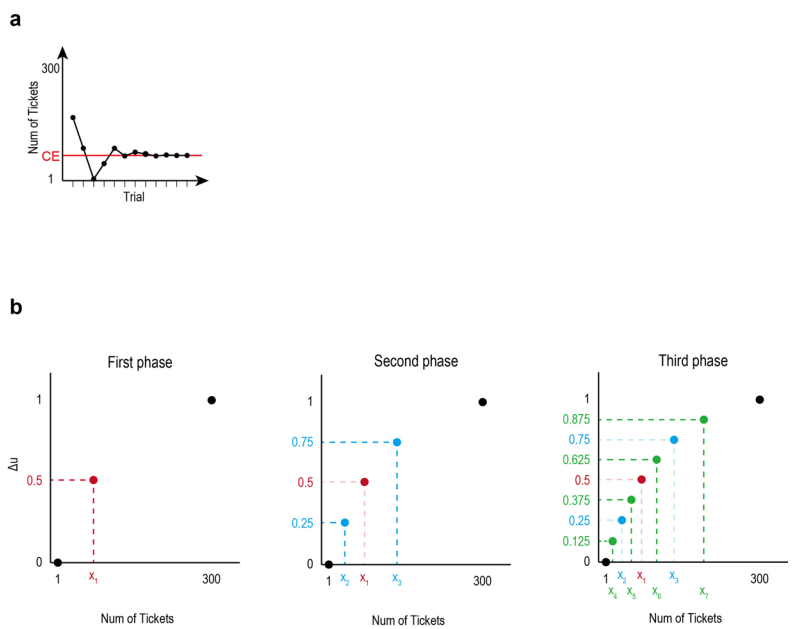
Extended data figures



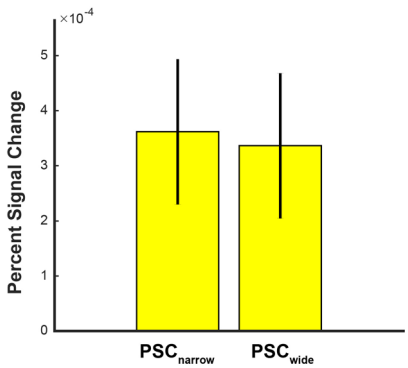
Extended Data Fig. 1 | Utility as an ordinal scale (ordinal utility).
Each axis represents the preference for an apple over a banana and a banana over an orange. Because only the order among the options is meaningful in this scale, any monotonically increasing transformation is permitted without losing any information. A representative (red line) and two variants (gray thin lines) are shown.



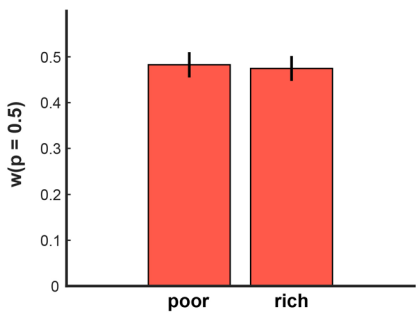
Extended Data Fig. 2 | Pareto efficiency.
Consider a situation in which Persons 1 and 2 split $100. The feasible area is shown in gray. Under the assumption that each person's utility is increasing function of money, point A is not Pareto efficient, because both persons' shares can increase from point A to B for example. Point B is Pareto efficient, because anyone's share cannot increase without decreasing the other's. Any point on the red line is in the Pareto efficient state. However, Pareto efficient states do not exclude highly unequal points like point C.



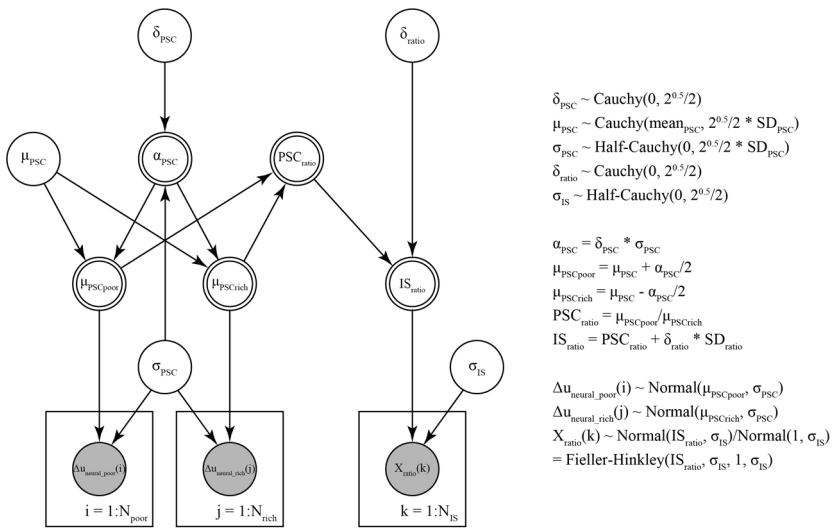
Extended Data Fig. 3 | Measuring utility function.
a, Measuring CEs using the PEST procedure. The black trace shows the numbers of tickets for a sure option compared with the 50/50 gamble between 1 and 300 tickets in the food gambling task, for illustration purposes. The lottery loss number (= 1), and 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 Methods).



Extended Data Fig. 4 | No normalization of neural signals for the utility prediction errors.
The activities of the brain region representing utility prediction errors in the food gambling task are shown. 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 with moderate evidence ($t(57) = 0.19$, $p = 0.42$ one-sided, $BF_{+0} = 0.169$, median posterior $\delta = 0.025$ with 90%CI [-0.185, 0.235]). Error bars indicate the SEM across participants.



Extended Data Fig. 5 | 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 ($t(58) = 0.207$, $p = 0.84$ two-sided, $BF_{10} = 0.268$, median posterior $\delta = 0.044$ with 95% CI [-0.418, 0.510]). Error bars indicate the SEM across group members.



Extended Data Fig. 6 | 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.