神経信号測定による効用の個人間比較 [a]

Kaosu Matsumori[b, c, d, e] **Kazuki Iijima**[a, f, g] **Yukihito Yomogida**[a, h, i]
Kenji Matsumoto[a, j]

Abstract

Aggregating welfare across individuals is a fundamental problem in our society. There is no rational aggregation procedure that satisfies even some very mild conditions without interpersonally comparable utility (Arrow's impossibility theorem). However, scientific methods for interpersonal comparison of utility have thus far not been available. Here, we have developed such a method based on brain signals. We found that medial frontal activity was correlated with changes in expected utility. The ratio of lower-income and higher-income participants' neural signals coincided with estimates of their psychological pleasure by "impartial spectators". We used the aggregated welfare from our experimental data to derive an optimal decision rule. These findings suggest that our interpersonal comparison method enables scientifically reasonable welfare aggregation by escaping from Arrow's impossibility.

Keywords: Neuroeconomics, fMRI, Arrow's impossibility, Distributive justice
JEL classification: D60, D71, D87

[a] This 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.
[b] Brain Science Institute, Tamagawa University; Machida, Tokyo, 194-8610, Japan.
[c] Department of Social Psychology, University of Tokyo; Bunkyo-ku, Tokyo, 113-8656, Japan.
[d] Japan Society for the Promotion of Science; Chiyoda-ku, Tokyo, 102-0083, Japan.
[e] E-mail: kaosu.matsumori@gmail.com
[f] National Institute of Mental Health, National Center of Neurology and Psychiatry; Kodaira, Tokyo, 187-8553, Japan
[g] E-mail: iijima.kazuki@gmail.com
[h] National Institute of Neuroscience, National Center of Neurology and Psychiatry; Kodaira, Tokyo, 187-8502, Japan.
[i] E-mail: yukihitoyomogida@gmail.com
[j] E-mail: matsumot@lab.tamagawa.ac.jp

1. Introduction

Aggregating welfare across individuals is a fundamental problem in our society. There is no rational aggregation procedure that satisfies even some very mild conditions without interpersonally comparable utility (Arrow's impossibility theorem) (Arrow, 1963; Sen, 2018). However, scientific methods for interpersonal comparison of utility have thus far not been available. It is important to extend the informational basis of welfare economics by developing an appropriate interpersonal comparison method. We argue an appropriate method for interpersonal comparison of utility has to satisfy the following two conditions. First, the interpersonal comparison method must rest upon scientific demonstration rather than upon ethical principle (Robbins, 1932). This condition warrants the objectivity and reproducibility of results about interpersonal comparison of utility. Second, the interpersonal comparison method must effectively capture human intuitions about utility and be validated by their coherence, as was done for the fluid expansion-based temperature scale in its initial stages (Chang, 2004). Satisfying this condition must serve as a starting point for subsequent scientific inquiry (Chang, 2004). One popular idea among both economists and philosophers is that, in order to solve the problem of interpersonal comparisons of utility, we have to look at how ordinary people make such comparisons in everyday life, because it is often pointed out that ordinary people make interpersonal comparisons of utility with relative ease and apparent success (Rossi, 2014).

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

2. Results

2.1. Neural representation of utility

To examine neural representation of utility, 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. 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) using general linear model (GLM) analyses (GLM1-4).

Because utility for moderate amounts of money is approximately linear intrapersonally (Wakker & Deneffe, 1996), 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. We determined each participant's 0-1 normalised utility function for food tickets using the fractile method, with considering probability weighting function (Stauffer et al., 2014). The slope of the utility function for most of the participants decreased as the amount of food increased (Figure 1A), indicating that their marginal utilities diminished in the range from 1 to 300 food tickets.

Next, we analysed the fMRI data to identify brain regions whose activation correlated with utility. 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 Figure 1B, $t(55) = 4.79$, 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.

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. 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). We found moderate evidence for the absence of normalisation among blocks ($t(57) = 0.19$, $P = 0.42$ one-sided, $BF_{+0} = 0.169$) (GLM3).

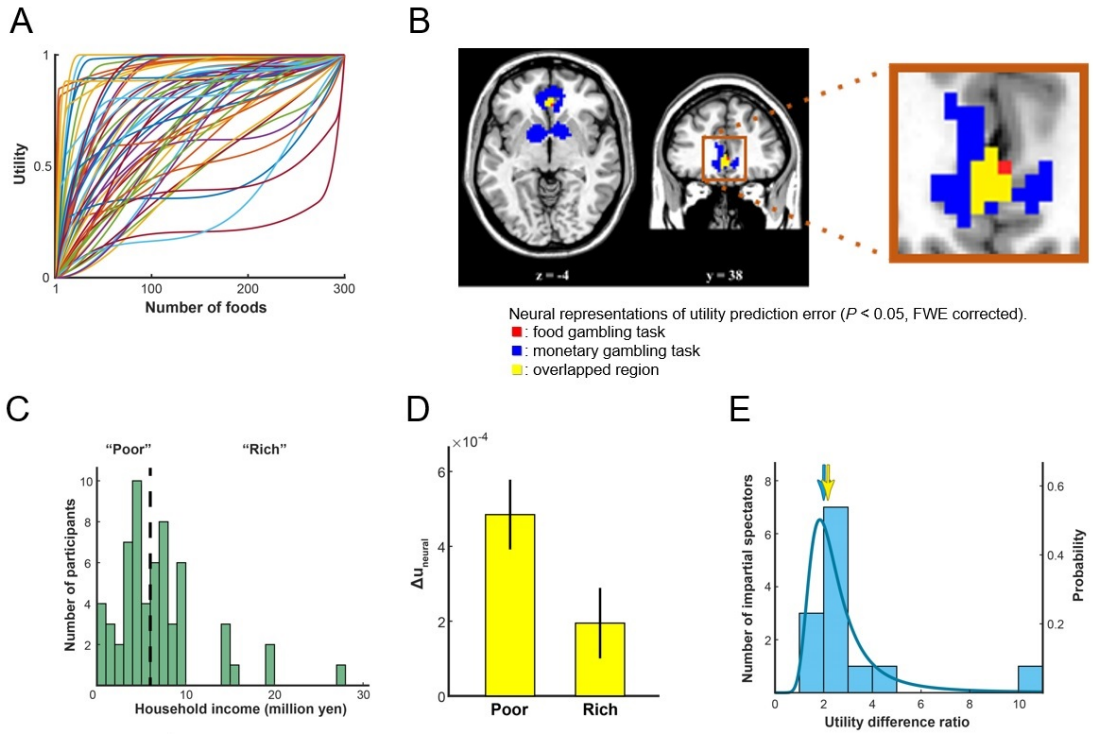

Figure 1. Neural representation of utility

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 (Figure 1B). 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 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 Figure 1B). The overlapped region was found in the ACC/vmPFC, hereafter called the "utility region" (yellow area in Figure 1B). 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

2.2. 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) (Figure 1C). We found a significant difference in Δu_{neural} ($t(58) = 2.19$, $p =$

0.016 one-sided, $BF_{+0} = 4.233$) (Figure 1D), while the subjective probabilities were not different between the groups. 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). According to the multiple regression analysis for Δu_{neural}, the term for income was significant ($P < 0.05$ one-sided), while the terms for sex and age were not. Thus, we were able to objectively determine an *inter*personally comparable scale of utility difference by using Δu_{neural}. The scale of utility difference would not be interpersonally comparable if it was based on choice behaviour alone.

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

In order to validate our scientific interpersonal comparison method using Δu_{neural}, we developed a behavioural task to measure impartial spectators' estimates of the utility difference ratio of the poor group and the rich group. An additional 15 participants completed the impartial spectator task. Two participants were excluded from the following analysis, because they did not pass the instructional manipulation check. We considered the remaining 13 participants "impartial spectators" for this task, because they were impartial and empathetic to the MRI participants (Smith, 1759). 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. From the results, impartial spectator participants' intuitive estimates of the utility difference ratio of poor group to rich group were calculated (median, 2.000) (Figure 1E). The utility difference ratio of poor to rich group based on the impartial spectators' estimates coincided with that based on Δu_{neural} with moderate evidence ($BF_{10} = 0.182$) (Figure 1E).

2.4. 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. 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 (d'Aspremont & Gevers, 1977). According to the ratio of posterior means of Δu_{neural} described above, for the problem of choosing between the policy of distributing ¥k to every poor participant and the policy of distributing ¥m to every rich participant, the optimal decision rule, for an expected utilitarian social welfare maximizer, was

$$\begin{cases} \text{giving ¥k to the poor group} & (\text{if } m/k < 2.171) \\ \text{either is fine} & (\text{if } m/k = 2.171) \\ \text{giving ¥m to the rich group} & (\text{otherwise}). \end{cases} \quad (1)$$

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.

3. Discussion

We developed a scientific method for interpersonal comparison of utility based on the activity in a specific brain region (ACC/vmPFC). The proposed method is prominent in the academic sense that it enables us to escape from Arrow's impossibility. Moreover, it can be applied for evidence-based policy making in nations that use cost-benefit analyses or optimal taxation theory for policy evaluation. The present study itself does not directly draw normative conclusions (Hume's law) on distributive justice. However, we believe that it raises a critical epistemic issue as a starting point for subsequent scientific inquiries to standardize a measurement method for utility as an interpersonally comparable quantity (Chang, 2004) and for philosophical reflections on social structures (Rawls, 1971).

Reference

Arrow, K. J. (1963). *Social choice and individual values* (2nd ed.). Yale university press.

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

d'Aspremont, C., & Gevers, L. (1977). Equity and the Informational Basis of Collective Choice. *The Review of Economic Studies*, *44*(2), 199–209.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

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

Rossi, M. (2014). Simulation theory and interpersonal utility comparisons reconsidered. *Synthese*, *191*(6), 1185–1210.

Sen, A. (2018). *Collective choice and social welfare*. Harvard University Press.

Smith, A. (1759). *The Theory of Moral Sentiments*.

Stauffer, W. R., Lak, A., & Schultz, W. (2014). Dopamine Reward Prediction Error Responses Reflect Marginal Utility. *Current Biology*, *24*(21), 2491–2500.

Wakker, P., & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern Utilities When Probabilities Are Distorted or Unknown. *Management Science*, *42*(8), 1131–1150.