Title: Decomposing the neurocomputational mechanisms of deontological moral preferences

Authors:
Yoonseo Zoh[1], Soyeon Kim[2], Hackjin Kim[3], M.J. Crockett[1,4], Woo-Young Ahn[2,5,6]

[1]Department of Psychology, Princeton University, Princeton, USA
[2]Department of Psychology, Seoul National University, Seoul, Korea
[3]School of Psychology, Korea University, Seoul, Korea
[4]University Center for Human Values, Princeton University, Princeton, USA
[5]Department of Brain and Cognitive Sciences, Seoul National University, Seoul, Korea
[6]AI Institute, Seoul National University, Seoul, Korea

Corresponding author:
Woo-Young Ahn, Ph.D.
Department of Psychology
Seoul National University
Seoul, Korea 08826
E-mail: wahn55@snu.ac.kr

Conflict of Interest: The authors declare no conflicting interests.

Significance statement

Moral dilemmas often pit harm to one person against the welfare of many. While protecting the individual is typically framed as deontological preference, it remains unclear whether such reasoning reflects a single process or multiple mechanisms. Combining computational modeling with neuroimaging, we show that concern for harming an individual separates into two distinct dimensions: minimizing maximum harm to an individual and setting a threshold for acceptable harm. These dimensions drove distinct patterns of moral choice and engaged separable neural systems, with the mentalizing network prioritizing the worst-off and the valuation network enforcing harm limits. Our findings demonstrate that deontological preference is multidimensional, uncovering the neurocomputational mechanisms by which people weigh consideration for individuals against the welfare of the group.

ABSTRACT

Research on the neurocomputational mechanisms of moral judgment has typically focused on contrasting *utilitarian* preferences to impartially maximize aggregate welfare and *deontological* preferences that judge the morality of actions based on rules. However, there has been little work to decompose the cognitive subcomponents of deontological preferences. Here, we investigated the neurocomputational mechanisms underlying two types of deontological preferences (Rawlsian and Kantian) and their contrast with utilitarian preferences in an incentivized moral dilemma task. Participants repeatedly decided how to allocate harm between a single individual (“the one”) and a group of 3-4 individuals (“the group”). The task distinguished preferences for Rawlsian, Kantian and utilitarian strategies by quantifying trade-offs among active harm, concern for the worst-off individual, and overall utility. Behaviorally, participants favored the Rawlsian strategy, preferring to impose more harm overall rather than disproportionately harm the one individual. Computational modeling revealed two dissociable dimensions of individual variability in Rawlsian preferences: i) minimizing the maximum amount of harm delivered to a single person and ii) subjective threshold of acceptable amount of harm imposed on one person. Combination of univariate and multivariate fMRI analyses revealed the engagement of distinct brain regions in these two dimensions of Rawlsian preferences, which respectively mapped onto activity in mentalizing and valuation networks. Our results reveal the neurocomputational mechanisms guiding tradeoffs between the welfare of one versus a larger group, and highlight distinct roles for the mentalizing and valuation networks in shaping Rawlsian moral preferences.

INTRODUCTION

Many common moral dilemmas raise the question of whether it is justifiable to impose sacrifices on some people to increase overall welfare (Baron, 1994). Moral judgments on this issue mark critical distinctions between different moral principles and have occupied a central place in the moral psychology literature (Greene et al., 2004; Crockett et al., 2013; Cushman et al., 2013; Everett et al., 2016). For example, the classic "trolley dilemma" of whether to sacrifice one person to save five put deontological and utilitarian principles in dispute by presenting the option of taking a harmful action, which is categorically prohibited, as the only means to achieve the better outcome (Foot, 1967; Thomson, 1985). Observations of people's responses to such sacrificial dilemmas led to the dual process model, which mapped the tension between deontology and utilitarianism onto intuitive and deliberative processes (Greene et al., 2004; Greene, 2007; Greene et al., 2008; Koenigs et al., 2007; Patil et al., 2021).

The dominant responses against harming one person in a sacrificial dilemma reflect a categorical (deontic) prescription in Kantian ethics, which, however, only accounts for a fraction of deontological theories. Actual moral decisions made by lay people are unlikely to be fully captured by the unconditional prohibition imposed by rule-based Kantian ethics. There is more to the deontological principle concerning justice and fairness and its tension with consequentialism. It is possible that fairness-based deontological ethics focused on mutual understanding and commitment to cooperation may better explain people's moral decisions and their psychological underpinnings from a normative standpoint (Baumard & Sheskin, 2015; Everett et al., 2016; Baumard, 2016; Levine et al., 2023).

The Rawlsian approach is another deontological principle which reflects concerns for justice by promoting inviolable individual rights and fair divisions of benefits and burdens. John Rawls, in opposition to utilitarianism, argued *“justice denies that the loss of freedom for some is made right by a greater good shared by others”* (Rawls, 1971). The Rawlsian principle requires that the prospects of the worst-off be maximized, ensuring agreement among all parties from an impartial standpoint, and that this evaluation take precedence over considerations of efficiency. While a Rawlsian approach to deontological ethics has received comparatively less attention in understanding our moral decisions, scenarios that pit utilitarian against Rawlsian principles have shown an overall preference for Rawlsian approach, suggesting its accordance with commonly held moral views of lay people (Baron & Jurney, 1993; Baron, 1995; Konow, 2001; Sandel, 2009; Baumard & Sheskin, 2015; Everett et al., 2016). For example, when a suggested policy incurs costs on some individuals, people object to implementing the policy even though doing so will lead to an overall increase in welfare (Baron & Jurney, 1993).

Despite the common observation of preference for the Rawlsian approach, however, we know little about the motivations or computations driving decisions that conflict with moral principles outside of strict deontic rules or welfare maximization.Also, what underlies individual differences in preference of a decision that is aligned with Rawlsian principle over other normative moral principles is poorly understood. Consequently, mapping between moral decisions predicted by normative ethical principles and their underlying psychological dimensions has likely been incomplete.

One clue in understanding people’s preferences for the Rawlsian approach comes from research on the singularity effect. Previous work on empathy showed people are strongly motivated to focus on protecting the mishap of a single individual, but such motivation fails to extend to a large number of people in need (Kogut & Ritov, 2005; Lee & Feely, 2016; Small, Lowenstein & Slovic, 2007; Schelling, 1968; Vastfjall et al., 2014). For example, an iconic photo of a single child suffering from the refugee crisis evoked great attention and donation of money, which, in contrast, failed to be elicited by statistics of hundreds of thousands death toll (Slovic et al., 2017). Even in the scenarios where the victim is not identifiable, people cared about saving the individual victim much more than saving a group of victims, suggesting people's insensitivity to the absolute number of lives at risk (Fetherstonhaugh et al., 1997; Slovic, 2010; Kogut & Ritov, 2005; Wiss et al., 2015).

A neuroimaging study has suggested such preferences toward protecting a single, individual victim are imposed by the unique characteristic of our mentalizing ability, being able to represent the mind of other individuals only one at a time (Ye et al. 2020). It is possible that Rawlsian moral preference (which emphasizes improving the situation of the worst-off person) may share a common psychological underpinning with the singularity effect, as people weight the worst-off individuals' suffering more than that of others. Mentalizing biased toward the worst-off individual, but not the group as a whole, may guide a strong motivation to contribute to helping one person, even at the expense of aggregate utility.

Previous work also demonstrated the role of mentalizing in Rawlsian preference. Rawlsian strategy of maximizing the 'worst-off' position as reflected in its '*maximin*' principle in distributive decisions (maximize the prospect with respect to the least advantaged position) was shown to correlate with the activity in mentalizing regions including RTPJ (Kameda et al., 2014). Notably, this Rawlsian preference in distributive decisions for others was concordant with avoiding the worst possible outcome in gambling decisions for self at the behavioral and neural levels, suggesting a domain-general avoidance of the worst-off outcome driven by its magnified representation in mentalizing networks (Kameda et al., 2014). However, the role of mentalizing networks in decisions to protect the worst-off individual from harm has not been tested, and it remains unknown whether mentalizing networks are also implicated in computations in pitting Rawlsian principle (maximizing the minimum welfare; maximin) against utilitarian principle (minimizing the aggregated harm) in allocation of harm.

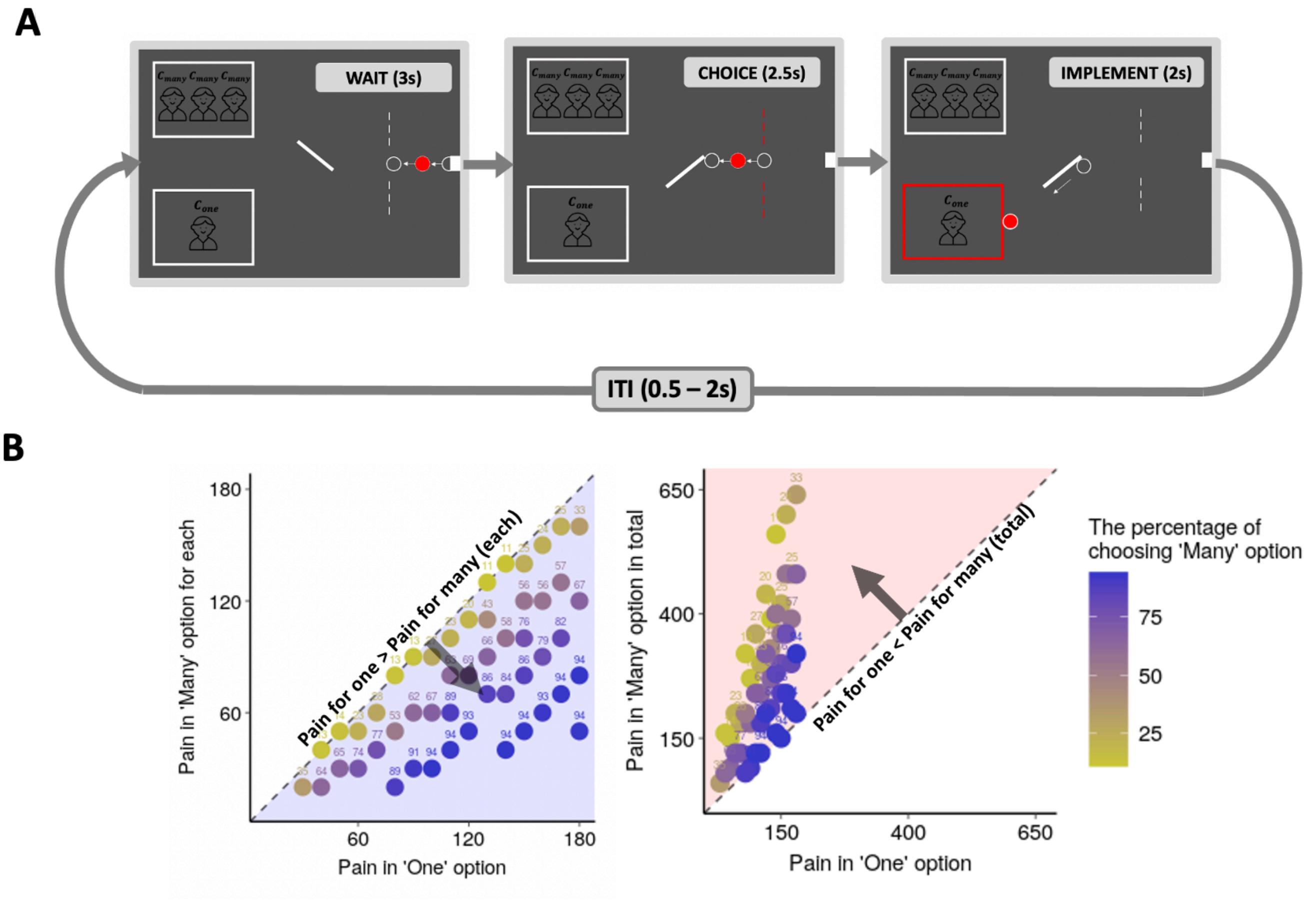
Another psychological dimension that might be critical in explaining the psychological basis of Rawlsian preference is the representation of agreeability. Rawlsian principle is grounded on the contractualist idea that morality of an action is founded upon mutual agreement of all involved parties (Parfit, 2011; Scanlon, 2000; Rawls, 1971; Habermas, 1990). Rawls, by introducing the idea of original position where no one knows in advance where they will end up being situated in the society, aimed to establish a theory of justice that is beneficial and fair to all - and therefore everyone would reasonably agree to choose (Rawls, 1971). While it is unlikely that people apply an impartial reasoning in a strict sense as Rawls conceived, people may spontaneously adopt the idea of agreeability and determine whether the option is acceptable to all in guiding their moral decisions (Misyak et al., 2014). Therefore, we hypothesized that the representation of agreeability might be another important dimension of Rawlsian moral preference that is separable from the biased mentalizing towards the worst-off individual.

Consistent with this hypothesis, prior studies have found evidence of contractualist moral intuitions in lay people's moral reasoning (Misyak et al., 2014; Levine et al., 2023; Levine et al., 2020). We examined whether such contractualist moral intuition composes a meaningful psychological dimension of Rawlsian moral preference. In psychological space, the idea of agreeability may be translated into a signature of differing interpretations across individuals on what is a fair outcome to all. While this may not be easy to capture with behavioral data alone, at a computational level, we expected it to be quantifiable as subjective thresholds of what a decision-maker considers a reasonable distribution of the benefits and burdens from an impartial standpoint. Therefore, we sought to identify a computational signature of contractualist intuition in Rawlsian moral preference as well as its neural instantiation, which we posited to be psychologically and neurally distinct from the heightened concern for an individual in the worst-off position.

To this end, we deployed a novel fMRI paradigm where participants had to repeatedly decide how to allocate harm between a single individual ("the one") and a group of 3-4 individuals ("the group") (Fig. 1A). We operationalized harm as amount of time spent completing a cold pressor task (Bayer et al., 2005). We manipulated the amount of harm in "the one" option to be always the same or bigger than the amount of harm allocated to each individual in "the group" option (C_{one}), but was always the same or smaller than the total amount of harm involved in "the group" option ($C_{group}*N_{group}$)

(Fig. 1B). There were 50 unique combinations of trials, each repeated three times under different default conditions (no default, default allocation to the one, default allocation to the group) making 150 trials in total.

Figure 1 Experimental design and choice behavior


A
C_{many} C_{many} C_{many}
WAIT (3s)
C_{one}
CHOICE (2.5s)
IMPLEMENT (2s)
ITI (0.5 – 2s)
B
Pain in 'Many' option for each
180
120
60
Pain for one > Pain for many (each)
Pain in 'One' option
Pain in 'Many' option in total
650
400
150
Pain for one < Pain for many (total)
The percentage of choosing 'Many' option
75
50
25

A. Schematic of harm allocation task Participants repeatedly chose how to allocate harm between a single individual and a group of 3-4 individuals. Here, the amount of harm was specified as the length of the time getting cold pressor task and shown on the top of each person icon representing individuals. A trial starts with a red ball proceeding across the screen toward two different options. The direction of the lever in the middle indicates the default option that will be implemented if the participant took no action (WAIT stage). After the ball crosses the dotted line, the participant may press the button to switch the lever and direct the ball toward the other option (CHOICE stage). Once the ball hits the lever, the participant can no longer switch it. And the ball continues to move toward the option the participant has chosen to assign the harm and then the box surrounding the chosen option changes color once the ball hits it (IMPLEMENT stage).

B. Preference for Rawlsian option as a function of harm for one vs. group. Across trials, the amount of harm delivered to the one person and the group of people varied. Each dot represents a unique combination of trial parameters. **The left** plot compares pain delivered to the 'one' option versus each individual in the 'many' option, while **the right** plot compares pain delivered to the 'one' option versus the total pain across all individuals in the 'many' option. By design, minimizing the total amount of harm always required delivering at least an equal or greater amount of harm to the one than to each member of the group. Participants therefore had to choose between selecting a utilitarian option (minimize the total amount of harm delivered) or Rawlsian option (minimizing the maximum amount of harm delivered to a single individual). Preference for the Rawlsian option increased with increasing harm delivered to the one individual. The gray dotted line indicates equivalent amounts of harm for the one and the group.

With this task, we were able to capture the degree to which individuals take into account different normative considerations in harm allocation decisions. In particular, the task enabled us to characterize how much one upholds a Rawlsian strategy over a utilitarian strategy when they are pitted against each other. Furthermore, we were able to distinguish Rawlsian from Kantian preferences by enabling participants on a subset of trials to either make an active choice or allow a default outcome to occur, revealing the degree to which participants preferred to avoid performing harmful actions irrespective of outcomes. By applying computational models to choice behavior, we examined computations involved in this trade-off as well as whether they can be ascribed to distinct psychological dimensions of Rawlsian preferences. We hypothesized : i) preferences for maximin strategy and ii) representation of agreeability.

We characterized the neural basis of these two dimensions using univariate and multivariate methods. Importantly, the motivation to protect a worst-off individual manifested as 'maximin' strategy is directional in a sense that employing more of the strategy means the individual prefers Rawlsian principle over utilitarian principle. However, the dimension of agreeability has no such directional orientation, because giving more or less to a single individual does not necessarily mean being more committed to Rawlsian principle but represents a different idea of agreeability one holds. Therefore, for the 'maximin' dimension, we applied a univariate model-based approach for identifying brain regions that scale with the recruitment of maximin strategy. On the other hand, for probing the neural instantiation of agreeability dimension, multivariate approach of inter-subject representational similarity analysis

(IS-RSA) was applied to examine which brain region shows shared responses among participants who have similar idea of agreeability. For the latter, we specifically examined whether participants with similar profiles of agreeability threshold also exhibit similar neural representations of harm concerning an individual recipient during moral decision-making. Our analysis revealed the neural encoding patterns of harm that are similar among participants who shared a similar level of agreeability, independent of the actual decisions made.

The utilization of both univariate and multivariate neuroimaging analysis techniques enabled us to investigate the neural mappings associated with two distinct dimensions of Rawlsian moral principles. Our research findings shed light on the psychological dimensions that underlie Rawlsian moral preferences, offering new perspectives into the moral intuitions that shape individuals' judgments when weighing different moral principles.

RESULTS

Rawlsian strategy was favored in harm allocation decisions

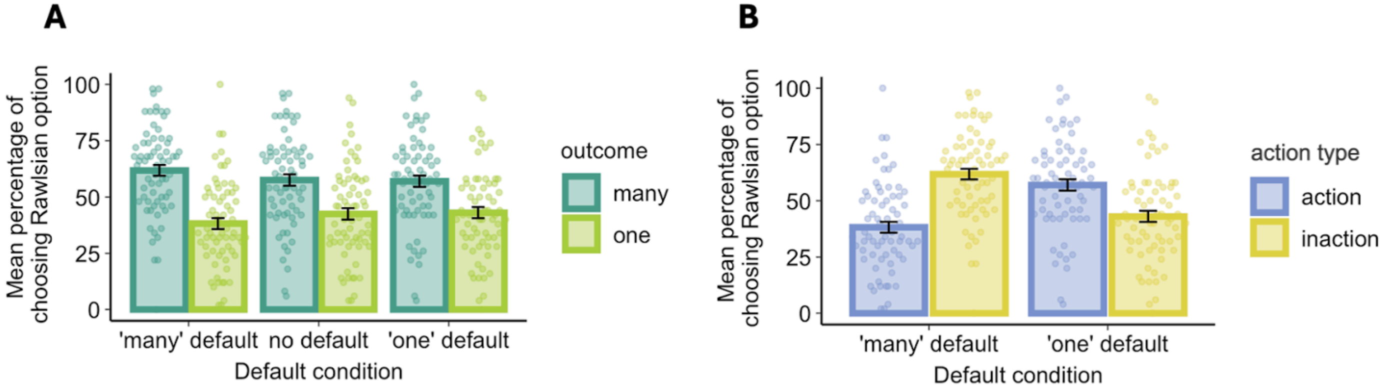
Our task design was able to distinguish three different normative moral preferences: Rawlsian, Kantian, and utilitarian. Rawlsian and utilitarian preferences were distinguished by choices between “the one” and “the group” option. “The one” option entailed less harm overall than “the group” option. Therefore, the choice between two options pit the “maximin” (maximizing the minimum welfare - in the context of our task, this means minimizing the maximum amount of harm) Rawlsian strategy against the utilitarian strategy of maximizing the overall welfare, which means minimizing the overall amount of harm in our task, with the choice of “the group” indicating one favored the Rawlsian strategy over utilitarian strategy.

We first examined the proportion of trials on which participants chose the Rawlsian strategy over utilitarian strategy. Most participants favored Rawlsian option (“the group”), with the proportion of Rawlsian choice significantly different from chance

(*M* = 0.59, *SD* = 0.20, *t*(67) = 3.68, *p* < .001). Moreover, when comparing the total amount of harm participants chose to assign to “the one” vs “the group”, participants chose to give 68.36 sec (*SD* = 37.19) of more harm overall to “the group” on average in order to save an individual from being disproportionately targeted.

Next, we examined the prevalence of Kantian preferences. Kantian ethics posits that certain actions should be absolutely prohibited. In the context of our task, making an active choice of who gets harmed is wrong, because harming is morally not permissible at any cost. Therefore, Kantian preferences should bias participants toward “inaction” or allowing the default option to be implemented without any active choice. In our task design, we repeated the same choice set across three conditions that differed in terms of default option (“one-default”, “many-default”, and “no-default”), and thus, we were able to compare participants’ choices across conditions and test whether participants indeed displayed Kantian preference toward allowing a default option.

Figure 2. Rawlsian preference dominate Utilitarian and Kantian preferences



A In the no default condition, most participants preferred the Rawlsian option over the Utilitarian one. We observed a significant increase in preference for the default option when the Rawlsian option was presented as the default choice. That is, Rawlsian option was more likely to be chosen when it was presented as default, compared to when there was no default option or when Ralwisan option was presented as alternative option, which provides partial support for the moral preference predicted by Kantian ethics. **B** The impact of default condition, however, was qualified by participants’ choice preference. There was a strong preference for the Rawlsian option regardless of whether it was presented as a default or the alternative option.

We do not find evidence for an overall Kantian preference, as participants were not more likely to choose the default option over the active choice across conditions (Figure 2: the difference in probability of switching to alternative option between "one-default" and "many-default" condition: $t(67) = 3.99$; $p < .001$). Instead, we observed that the preference for the default option was evident only when the default option aligned with Rawlsian preferences, as participants were more likely to choose the Rawlsian option when it was presented as the default (Figure 2, difference in choice rate for Rawlsian option across three conditions: $F(1.45, 97.05) = 14.56$; $p < .001$: difference in choice rate for Rawlsian option between "many-default" and "no-default condition: $t(67) = 4.03$, $p < .001$; difference in choice rate for Rawlsian option between "many-default" and "one-default" condition: $t(67) = 4.13$, $p < .001$). This suggests that Rawlsian option was strongly favored over the utilitarian option irrespective of whether it was presented as a default or alternative. In other words, Rawlsian preference to avoid disproportionately harming the worst-off individual overruled consideration of "inaction" of active harm as prescribed by Kantian ethics.

Computational model of harm allocation decisions

We next built a set of computational models to examine what computations go into harm allocation decisions and the sources of individual variability in Rawlsian moral preferences. Our models were theoretically motivated with respect to two dimensions of Rawlsian moral preference we hypothesized. Firstly, we modeled the motivation to protect a worst-off individual, or 'maximin' strategy, where the utility comparison is made for the worst-off outcome for an individual in two options. This Rawlsian strategy was directly pitted against utilitarian strategy of "minimizing aggregate harm" in our task, where the utility comparison is made between the total amount of harm between two options. Therefore, we considered a weighting parameter, noted as α (alpha), indicating how much participants gave weight to the welfare of "the one" versus "the group".

Next, we considered how people might take into account agreeability in harm allocation decision, which we hypothesized to be a distinct dimension of Rawlsian moral preference from the maximin. Therefore, we considered a parameter that reflects the representation of agreeability and is subject to the employment of 'maximin' strategy.

This parameter, noted as " φ" (phi), expressed a participant's subjective threshold of (reasonably) acceptable amount of harm to impose on one person. We conducted parameter estimation and model comparison using hierarchical Bayesian approach (Berger, 2013; Gelman et al., 2004; Lee, 2011). The model that best explained our choice data as identified the smallest leave-one-out information criterion (LOOIC) value (Vehtari, Gelman, & Gabry, 2017) included three free parameters - a maximin parameter that has value between 0 and 1, a constant term of agreeability parameter, and an inverse temperature parameter. This model accurately predicted choice behavior in our task, outperforming a range of alternative models (Supplementary figure 1, correlation between model prediction and observed choice proportion of Rawlsian decision : r = 0.99, p <.001).

Our best model posits the difference in subjective utility between the Rawlsian option and utilitarian option is computed as weighted differences between two strategies. The outperformance of this model over the ones which constrained the weight to either 0 or 1 (alpha constrained to 1; difference in expected predictive accuracy : -2144.2, *SE* = 195.3; alpha constrained to 0; difference in expected predictive accuracy : –3438.5, *SE* = 191.4) indicates that people balanced the utilitarian and Rawlsian moral considerations in harm allocation decisions, rather than only relying on one strategy. Additionally, inclusion of an agreeability parameter markedly increased the model fit to data (difference in expected predictive accuracy : -270.8, *SE* = 43.8), suggesting the representation of agreeability indeed comprises a psychologically distinct dimension of Rawlsian moral preference. The agreeability parameter, by being adjoined to maximin computation, was able to capture what participants considered a reasonably acceptable amount of harm to allocate more or less to an individual.

Next, we looked into the estimated value of alpha and phi, two free parameters capturing individual variability in Rawlsian moral preference. The estimated value of parameter α at the group level was 0.95 (*SD* = 0.01). This confirms that participants heavily relied on Rawlsian "maximin" strategy more than utilitarian strategy in their choice, placing a higher priority on saving an individual from the worst off position than to maximizing the total utility. On the other hand, unlike alpha parameter which had lower variance with strong preference toward Rawlsian strategy, individual estimates of phi parameter showed a wide range of values (the estimated value at the group level : *M* = -10.03, *SD* = 2.85), from negative to positive, suggesting substantial individual variability in subjective threshold of acceptable amount of more or less burden for an individual. Importantly, there was a weak, non-significant correlation between estimated parameter value of α and φ (Supplementary figure 2, r (66) = .13, p = .296), suggesting the dissociability of those two parameters in representing two dimensions of individual variability in Rawlsian preference. Both parameters showed good recoverability in parameter recovery test, indicating they reliably contribute to account for cognitive processes underlying choice behavior in the task (Supplementary table 3, α: r = 0.98, p < .001; φ : r = 0.92, p < .001).

Neural sensitivity to the maximin computation predicts Rawlsian moral preference

As we found two distinct computations involved in Rawlsian preferences, we proceeded to examine their neural substrates. While our modeling results suggest that adherence to maximin strategy and representation of agreeability are dissociable dimensions of Rawlsian preference at algorithmic level, it is possible that processing the computation of two dimensions are subserved by either shared or distinct domain of neural processes at the the level of implementation (Marr, 1982). If two dimensions engage common neural substrates, it may suggest two dimensions, while their algorithms differ, share the same underlying social motivation or goal at a higher level (Lockwood et al., 2020). On the other hand, if they engage distinct neural substrates, it would further suggest that they are psychologically distinct dimensions at both algorithmic and implementational level composing individual variability of Rawlsian moral preference. With divergent characteristics of two dimensions in terms of directionality in stipulating Rawlsian preference, we adopted both univariate and multivariate approaches for neuroimaging analysis.

In terms of “maximin” dimension where more employment of maximin strategy does scale with Rawlsian preference, we applied a univariate approach to identify brain regions of which the activity tracks decision value in favor of Rawlsian option computed by maximin strategy. Additionally, we tested whether individual variability in maximin dimension of Rawlsian preference entails differential neural sensitivity to the difference in anticipated harm for an individual in the worst-off position. Based on the previous work highlighting the role of mentalizing in the motivation to protect the single individual (Ye et al. 2020; Kameda et al., 2014), we hypothesized encoding of maximin computation as well as individual differences in the degree to which one upholds this strategy is linked with brain activity in mentalizing network (dmPFC, TPJ, and precuneus; Saxe, 2006; Koster-Hale & Saxe, 2013; Jamali et al., 2021; Schurz et al., 2014; Schurz et al., 2021).

To first identify brain regions involved in maximin computation, we built a general linear model (GLM) that reveals parametric responses at choice onset, independently from participants' choices, to the objective amounts of harm maximally assigned to an individual that would result from choosing "the one", relative to "the group" (C_{one} > C_{group}). This essentially reflects the "maximin" computation of maximizing the minimum welfare (minimizing the maximum harm) imposed on an individual as predicted by Rawlsian principle. A positive effect of this parametric contrast indicates the voxels encoding the extent of evidence in favor of Rawlsian option arising from maximin computation. Consistent with our prediction, this contrast revealed an effect in mentalizing network, including right TPJ, left middle frontal gyrus, and right medial frontal gyrus, suggesting mentalizing is a key process involved in upholding maximin strategy (Figure 3).

Having found that maximin computation is encoded in the activity of brain regions in mentalizing network, we examined the hypothesis that varying neural sensitivity in those regions to evidence in favor of Rawlsian option predicts individual differences in maximin dimension of Rawlsian moral preference. To this end, we added individual differences in moral preferences estimated from our winning model, parameter α, as second (group) level regressor onto the parametric contrast of difference between the objective amounts of harm assigned to an individual in "the one" option compared to "the group" (C_{one} > C_{group}). This analysis showed a positive link between Rawlsian moral preference and neural responses to maximin computation in mentalizing network including the regions of inferior frontal gyrus, superior and inferior parietal lobule, and superior frontal gyrus. The effect extended to other regions involved in social processing as well, including bilateral insula and left middle frontal gyrus. Interestingly, this analysis also revealed negative link between Rawlsian preference and activity in regions implicated in valuation including vmPFC and medial frontal gyrus, indicating neural responses to maximin computation in valuation network scaled negatively with participants' Rawlsian preference (Figure 3). Functional connectivity analyses provided further suggestive evidence that individual differences in Rawlsian preference may be explained by the degree to which one incorporates concerns for the worst-off individual into integrative values of moral decision through mentalizing (See Supplemental Analysis 1).

Figure 3. Rawlsian preference is associated with differential neural sensitivity to the maximin computation

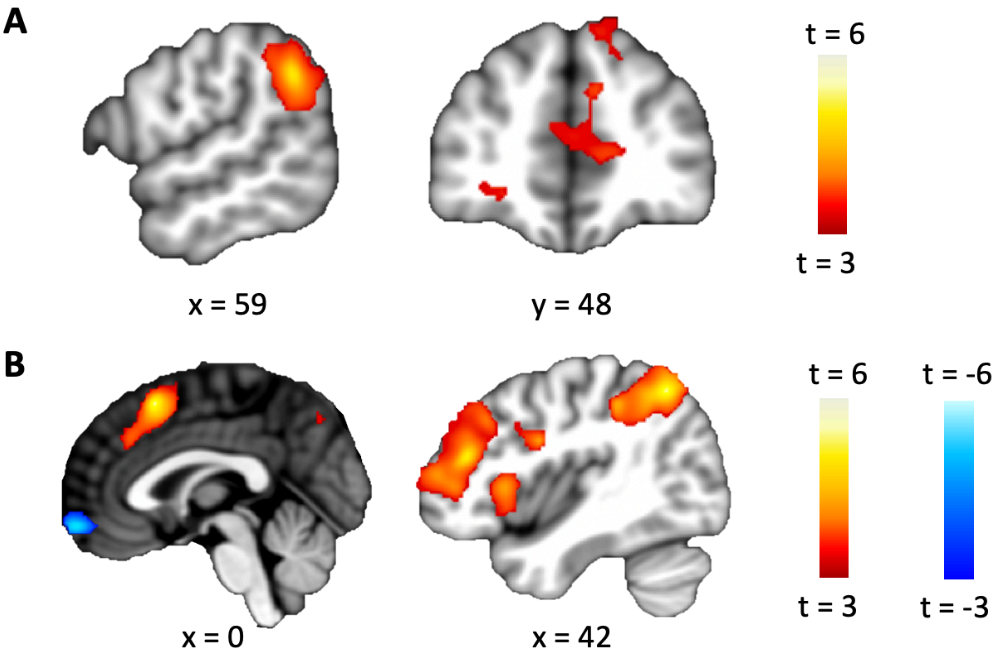


Fig 3A The brain regions of right TPJ / supramarginal gyrus, left middle frontal gyrus, and right medial frontal gyrus parametrically responded to the harm for one relative to an individual in the group (C_{one} > C_{group}), which is the input to the maximin computation in Rawlsian strategy.

Fig 3B Rawlsian moral preference is predicted by increased neural sensitivity in brain regions including bilateral insula, the inferior frontal gyrus, inferior and superior parietal lobule, right Insula, and left middle frontal gyrus. In addition, Rawlsian moral preference was associated with decreased neural sensitivity in vmPFC to the maximin computation. The whole-brain maps show brain regions where parametric response to relative harm for “the one” vs individual in “the group” correlated with the estimated parameter value of Rawlsian preference alpha from our model.

Neural Signature of contractualist moral intuition

Next, we sought to examine the neural substrates of “agreeability” dimension in Rawlsian preference. Importantly, unlike “maximin” dimension where more employment of maximin strategy does scale with Rawlsian preference, individual differences on “agreeability” dimension is lacking such directional implication as it is concerned with representation of different ideas one considers agreeable to give more or less to an individual than the rest. Therefore, the univariate approach which we applied to identify neural substrates of “maximin” computation would be less informative for addressing this question. Instead, we tested the hypothesis that participants who share similar ideas of agreeability are also likely to share brain responses to encoding of harm ascribed to an individual. This reflects the idea of agreeability as the subjective threshold of what people deem as mutually agreeable, which we confirmed with our best computational model. We further predicted that varying deviations from such threshold are likely to evoke neural activity that are similar among individuals who represent the equivalent or similar threshold.

To test this possibility, we employed IS-RSA (Inter-subject representational similarity analysis) approach which maps the representational distance across individuals in behavioral measure onto brain activity. By linking individual differences in agreeability parameter we estimated from our model to brain data, it allowed us to identify neural substrates of “agreeability” dimension of Rawlsian preference. Specifically, we examined which brain regions display shared brain patterns during encoding of harm ascribed to the worst-off individual predicted by similarity in agreeability parameter from our model. To this end, we ran dyadic regression models across voxels, where for every pair of participants in our data, neural similarity in encoding the amount of harm to the worst-off individual was regressed onto similarity in agreeability parameter (phi). Critically, we controlled the similarity in maximin parameter (alpha) in our regression model. This allowed us to examine similar neural responses to harm that is specifically attributed to the similarity in agreeability above and beyond participants' decisions.

Figure 4. IS-RSA analysis

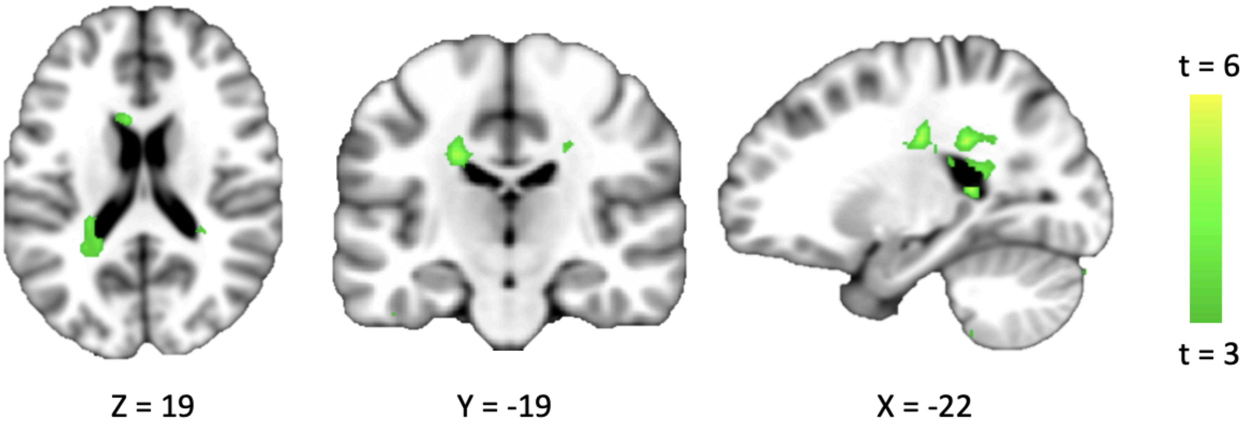


Fig 4. Brain regions identified to show increased similar multivariate patterns in encoding harm assigned to the worst-off individual among participants with greater similarity in agreeability dimension of Rawlsian moral preference. Dyadic regression models where the neural similarity in encoding the worst-off harm were regressed against the similarity in model parameters were estimated for individual voxels and the beta estimates of agreeability dimension for each individual voxel was mapped onto 3D brain space.

Our analysis revealed that the shared idea of agreeability predicted the extent to which regions in valuation network showed a similar response to the amount of harm that was disproportionately assigned to the worst-off individual. Specifically, the similarity in the idea of agreeability between participants predicted increased similarity of parametric responses to the amount of harm to the worst-off individual in regions encompassing valuation network, including striatum and cingulate gyrus (Fig 4). The results are consistent with our computational model where maximin strategy and agreeability were represented by distinct parameters in computing decision value, but further suggest that those two dimensions are also implemented in distinct domain of neural processing, with the variance on idea of agreeability uniquely mapped onto the neural encoding of harm on brain regions implicated in value-related processing.

DISCUSSION

In the current study, we investigated how people adjudicate conflicting normative moral principles when making decisions about allocations of harm. Our findings revealed distinct dimensions of Rawlsian moral preferences in this context. We developed a new paradigm in which participants allocated harm between a single individual and a group of individuals. In most trials, minimizing the total amount of harm required delivering more harm to the one than to each member of the group, thus pitting Rawlsian moral principle of promoting fairness and justice among individuals against Utilitarian principle of maximizing aggregate outcome.

Participants in our task displayed strong Rawlsian preferences overall, predominantly choosing a less optimal option in terms of overall utility to ensure that the worst-off individual does not get a disproportionate amount of harm. This result extends the previous finding that people show strong preference for Rawlsian option in distribution decision in the domain of fairness (Frohlich & Oppenheimer, 1992; Engelmann & Strobel, 2004; Charness, 2002) and highlights how Rawlsian preference, as a distinct moral preference from Kantianism (rule-based deontology) and utilitarianism, manifests in weighing harm for individual versus group.

Kantian moral preference, which makes the prediction that people will distinguish ‘action’ from ‘inaction’, was not supported in the context of our task, with the Rawlsian preference overruling its impact for participants’ choices. Influential work investigating people’s responses to sacrificial moral dilemmas showed the dominance of Kantian judgments in a scenario where it involved direct and personal harm as an instrumental means to save a greater number of people (Greene et al., 2007; Paxton, Ungar, & Greene, 2011; Greene et al., 2001; Greene et al., 2004). Moral dilemmas used in previous research are usually designed in a way that causes harm to the most disadvantaged individuals in a way that involves action of direct and personal harm. As a result, previous studies indicating a Kantian preference may actually be indicative of a

Rawlsian preference. The methods employed in the past did not allow for the distinction between these two preferences. Therefore, many of the findings that have highlighted the prevalence of Kantian preference could be interpreted as reflecting a Rawlsian preference. While Rawlsian and Kantian principles both fall under the umbrella of deontology, our work suggests that their underlying psychological profiles are distinguishable from each other.

Most participants in our study showed a moral preference that is a mixture between Rawlsian and utilitarian principle, while more heavily weighted toward Ralwisan one. The trade-off between the amount of harm allocated to an individual and group revealed the extent to which people are aligned with the two competing moral principles. Specifically, increasing the maximum amount of harm for the worst-off individual led to favoring the Rawlsian option over the utilitarian option, while increasing the number of harm recipients or the overall amount of harm for the group led to favoring Utilitarian option. Participants' moral decisions were best explained by our computational model which quantified the subjective value of choices by decomposing distinct dimensions of individual variability in Rawlsian moral preferences.

Past research on the psychological profile of Rawlsian moral preferences has mainly focused on the *'maximin'* Rawlsian strategy in its competition with Utilitarian strategy (Kameda et al., 2014). Our computational model also confirmed *'maximin'* Rawlsian strategy in the context of harm allocation, demonstrating the motivation to protect an individual from the worst-off outcome at the cost of increasing overall harm (and therefore maximizing the minimum welfare). However, we further illuminated another dimension of Rawlsian moral preference that is distinct from the employment of 'maximin' strategy. Theoretically motivated by the contractualist basis of Rawls's theory of justice (Rawls, 1971), we hypothesized that people would represent different ideas of agreeability grounded in the mutual respect of individuals' perspectives. Incorporating this agreeability dimension to our computational model identified an important source of individual variability in Rawlsian moral preferences. Capturing different subjective thresholds of what is deemed acceptable amount of harm to be imposed on one person, it significantly improved the model fit to the behavioral data.

Importantly, the individual variability in ‘maximin’ and ‘agreeability’ dimension of Rawlsian moral preference was not significantly correlated, suggesting those two dimensions play distinct roles in explaining individual variability of Rawlsian preference. We propose that this contractualist intuition, while not directly pitted against utilitarian strategy of maximizing the aggregate outcome as the ‘maximin’ dimension does, offers a distinct motivation in the endorsement of Rawlsian principle. In the context of harm allocation, the agreeability dimension reflects consideration of what is a reasonable amount of harm that an affected individual would agree to receive more or less than the rest. By representing the standpoint of an individual in an impartial manner, it is more likely that the resulting decision will distribute burdens in a way that is more fair and justifiable even to the worst-off. Therefore, agreeability dimension can lead to protecting the mutual respect and balance between individual’s rights so that no one gets disproportionately burdened or sacrificed for enhancing aggregate outcome.

We consider a link between the agreeability dimension we have identified in our work and impartial reasoning intuitively reflected in the original position that Rawls introduced in his theory. In the original position, no one knows in advance where they will end up being situated in the society, and therefore it leads to impartially promoting a policy that even the worst-off can reasonably agree with. Previous work suggested veil of ignorance reasoning, which Rawls conceived as a way of inducing original position, counterintuitively leads to promote utilitarian maximization of aggregate outcome (Huang et al., 2019). However, we note that this is likely due to impartiality shared between the utilitarianism and contractualist thinking, rather than veil-of-ignorance reasoning that favors the greater good per se. We predict that in situations where increasing aggregate welfare involves sacrificing an individual’s welfare to the degree that the burden on the worst-off is too disproportionate to reasonably agreed from an impartial standpoint, (i.e.,directly pitting utilitarianism against Rawlsian principles), contractualist reasoning may diverge from utilitarian maximization and provide unique motivation to protect the balance between beneficiaries and the worst-off.

Our research highlights both distinctions and parallels between psychological profiles of Rawlsian preference in the domain of fairness and harm. Past works have shown that altruistic and egalitarian preferences were more apparent when making decisions involving harm than when allocating monetary gain between self and other, suggesting distinct moral preference in the domain of harm (Crockett et al., 2014; Davis et al., 2018). We suggest several possibilities as to why people may display strong Rawlsian preference specifically in the context of harm allocation. First, people might have an intuitive sense that pain is worse when experienced alone than when experienced with others, as illustrated by the common expression “misery loves company”. With the intuitive notion that harm is experienced worse alone than when with others, people may show greater aversion to targeting a specific individual in allocating harm, leading to more employment of maximin strategy. Moreover, the representation of agreeability may become more salient as one considers how the worst-off individual may also find harm received alone less agreeable, which can further motivate one to protect the worst-off from receiving disproportionately more harm than what the worst-off individual might reasonably agree.

Alternatively, in choosing the option that puts the worst-off position at risk of harm, there can be greater uncertainty about whom the one individual will be, how that individual will experience the allocated harm, and how that would affect their well-being, which may temper the preference for a utilitarian option. While we did not specify any information about the identity of harm recipients in our task, participants may have considered the worst-case scenario where the person situated in the worst-off position is particularly vulnerable to harm. Some of our participants indeed reported such concerns during debriefing interviews. Past work suggested that people behave in a more prosocial way when the impact of their decision on others’ welfare is uncertain (Kappes et al., 2019). Building on that finding, we speculate that the worst-off position via singularity effect makes uncertainty about the risk of harm more salient than it is for a group of individuals, leading to the motivation to protect the worst-off. For example, one might feel uncertain about how harm will be experienced by and impact the well-being of an individual, but such uncertainty is hard to be extended to a group of individuals whose experiences are likely to be collectively represented as average. Therefore, the unique characteristic of phenomenal and impact uncertainty about harm where it can only be drawn in an individualized manner may drive strong Rawlsian preference in the harm domain.

Our work also highlights psychological profiles of Rawlsian preference that are shared across harm and fairness domains. We found that computation of Rawlsian maximin strategy as well as its individual variability was associated with the activity of mentalizing network in the brain. This is consistent with previous work which investigated the neural basis of 'maximin' strategy in the domain of fairness and risky decision-making (Kameda et al., 2014). This also dovetails with the observation that mPFC is more engaged in response to events of a single person than a group of individuals, demonstrating the neural underpinnings of singularity effect (Ye et al. 2020). Therefore, we provide a unified explanation as to why people may dislike an outcome that does not share the greater good with the worst-off, regardless of whether the allocated outcome is less benefit or more burden. Our tendency to help the worst-off despite the cost of reduced utility is driven by taking the perspective of the worst-off and inferring what it is like to be situated in that particular position for getting to experience unjust distribution of benefits and burdens.

Finally, we investigated the neural profile of agreeability dimension of Rawlsian preference by applying a multivariate approach. With the IS-RSA approach that allowed mapping between inter-individual variation in model parameter and brain activity, we examined similar neural engagement in parametric response to the worst-off harm predicted by similarity in the idea of agreeability. We found that regions in valuation network including striatum and cingulate gyrus showed similar response to the amount of harm that was assigned to the worst-off individual. Our finding that two dimensions of Rawlsian preference can be respectively mapped onto mentalizing and valuation network highlights that they are dissociable both at the level of computation and implementation. Moreover, given the previous work suggesting the role of striatum in tracking integration of equity and utility (Hsu et al, 2008), the neural profile of agreeability dimension we identified here seems to reflect the evaluation of the worst-off individual's position from an impartial standpoint. We speculate that striatum plays a key role in maintaining reasonable balance in individuals' share of the greater good by tracking how much a share for the worst-off position deviates from one's internal threshold of agreeability.

To conclude, our study elucidates the neurocomputational process involved in harm allocation decisions adjudicating normative moral principles beyond the conflict between categorical deontic prescription and maximizing aggregate welfare. Our results demonstrate that Rawlsian moral preference, as separate from Kantian deontological preference, can be further decomposed into two distinct neurocognitive aspects. Our findings underscore the significance of broader deontological principle and highlight the nuanced nature of individuals' moral preferences within normative ethical theories. Formalizing the psychological mechanisms of our moral preferences in their diverse dimensions contributes to a more comprehensive understanding of the complex landscape of human morality, enriching our knowledge of how they shape and guide our social behaviors.

METHODS

Participants

Sixty-eight participants were recruited via a university community website as paid volunteers. Exclusion criteria were history of neurological disorders, psychiatric disorders, use of psychoactive medication or drug, and pregnancy or suspicion of pregnancy. Participants who majored in psychology or who had previously participated in social psychology studies were also excluded due to concerns that previous experience with psychology studies involving deception could increase the suspicion about the task in the current study. All participants were right-handed, had normal or corrected-to-normal vision, and fulfilled safety criteria for cold-pressor test and MRI. Sixteen participants were excluded from the fMRI analysis because six participants attended only the behavioral parts of the study due to the technical issues, four participants fell asleep in the scanner, one participant requested to exit the scanner during the first run, one participant had focal brain atrophy that was not identified prior to the scanning, four participants had excessive head movement (>3 mm) during the scanning, leaving a total of fifty-two participants whose data were analyzed for the fMRI study (18 females; 34 males; mean age 23.27 ± 3.06). We reported the behavioral results from analyzing the data of all 68 participants who completed the behavioral task (29 females; 39 males; mean age 23.43 ± 3.21).

Procedure

The procedure of the study was approved by the Institutional Review Board at Seoul National University (1907/003-007). The study consisted of a two hour-long session. At the start of the study, the experimenter explained the procedure and the purpose of the study. To make participants believe their decisions will be actually implemented and have real consequences for other individuals, we informed participants that the aim of the current experiment was two folds : 1) to examine people's preferences in making social decisions and 2) to decide parameters for a a group-decision making task in another study of the lab. While the second aim of the study was a cover story that was not carried out in reality, to increase the credibility of the cover story, we asked participants to sign the informed consent for using participant's decision as well as for waiving future participation in the study of group-decision making task.

After providing the informed consent, a protocol for saliva collection was administered for data to be reported for the purpose of a different study. Participants then completed a battery of trait questionnaires, which was followed by the instruction about the harm allocation task. Participants were told that at the end of the task, one trial will be randomly selected and their decision in that trial will be implemented to a group of participants in another study in the lab. We emphasized that there are no right or wrong answers, and participants shall make choices based on their own preference. In order to minimize the concerns about reputation or reciprocity and their impact on participants' choices, we informed participants that their choices and identity would be kept confidential to participants in another study of the lab getting affected by the participant's decision.

Before the scanning, participants first completed a 20 sec of cold pressor test at 4ºC water. We specifically chose to administer cold pressor test for 20 sec given it is not long enough to develop numbness with respect to increase of the time. Also, this was shorter than any pair of options presented in the task, allowing us to examine how people make decisions that have impact on other individuals to some extent beyond their own experience.This procedure allowed participants to experience pain involved in a cold pressor test before the harm allocation task where they were supposed to believe that their decision of allocating the amount of harm as time of getting this test involving pain will be actually implemented to other individuals. Following the administration of the cold pressor task, the participant completed the harm allocation task in the fMRI scanner.

After completing the scanning, participants completed a short debriefing questionnaire asking their experience and beliefs about the experimental setup, including a question of how much they believed that their choices would be actually implemented to other individuals (rated from 1 = 'not at all' to 5 = 'fully'). Participants overall reported high levels of belief that the cover story of our study would be real (*M* = 3.96, s.e.m. = 0.16). Participants also reported high confidence that their choices and identity would remain confidential (choices : *M* = 4.80, s.e.m. = 0.07, identity : *M* = 4.77, s.e.m. = 0.07), suggesting concerns about reputation or reciprocity had little or no influence over participants' choices in the task. All the participants were fully debriefed about the deception before departing the laboratory.

Harm allocation task.

On every trial, participants had to choose between allocating harm to a single individual ("the one") and a group of individuals ("the group"). "The one" option contained more harm (longer time of getting the cold pressor test) in terms of what was maximally given to an individual, while "the group" option contained more harm in total. Specifically, the amount of harm in "the one" option (C_{one}) was always the same or bigger than the amount of harm allocated to each individual in "the group" option (C_{group}), but was always the same or smaller than the total amount of time involved in "the group" option ($C_{group}*N_{group}$), thus satisfying the following property : $C_{group} \leq C_{one} \leq C_{group}*N_{group}$. This allowed us to examine how participants pit the "maximin" Ralwsian strategy against the utilitarian strategy of maximizing the overall amount of welfare (or minimizing the overall amount of harm). Here, Rawlsian strategy indicates maximizing the welfare of the worst-off position of getting the most amount of harm, so choosing ‘the group’ to minimize the maximum harm at the cost of imposing more harm overall. On the other hand, utilitarian strategy concerns bringing the better outcome in the aggregated level, and therefore choosing “the one” to minimize the overall amount of harm.

In both “the one” option and “the group” option, the amount of time ranged from 30s to 180s in increments of 10; with the number of individuals in “the group” option alternated between 3 and 4. The fifty choice pairs satisfying aforementioned conditions were predetermined and repeated across three conditions which differed in the default option that was implicated when participants were not pressing any button. The default option was indicated by the direction of the lever in the middle of the screen. Varying the default option across the same set of trials allowed us to examine the participant’s preference for the default option as predicted by Kantian theory. Duplicating the choice pairs three times resulted in 150 trials (50 trial pairs x 3 conditions) evenly distributed and delivered across three scanning runs lasting approximately 8 minutes each. We randomized the presentation order of the trial as well as the location of “the one” and “the group” options for every trial.

After 3s of the wait phase where the choice pair and the default condition are revealed, participants had a maximum of 2.5 s to choose either option by pressing a button box with their right index or middle finger. Button presses in selection of an alternative option resulted in shifting the lever, directing the ball to move toward the chosen option and hit the box surrounding the option in 0.5s. The selected option was highlighted for 1.5s. The end of each trial was followed by a jittered fixation cross (0.5 - 2s).

Computational modeling and model comparison

Model fitting and comparison were performed using hierarchical Bayesian estimation (Berger, 2013; Gelman et al., 2004; Lee, 2011) using the package RStan (Stan Development Team 2020) in R (R Core Team 2020). The individual parameters were treated as a random sample drawn from a group distribution and individual and group parameters were estimated simultaneously in a mutually constraining manner. We employed weakly informative priors for group-level parameters and non-centered parameterization (Matt trick) method for optimizing the sampling process (Ahn, Haines, & Zhang, 2017). The posterior distribution of parameters were simulated by means of four independent Markov Chain Monte Carlo (MCMCs) sampling chains. Each chain consisted of 1,000 MCMC samples after discarding 1,000 warm-up samples, for a total of 4,000 posterior samples. We compared the goodness of fit of different models using LOOIC (Leave-One-Out Information Criterion) as a metric (Vehtari et al., 2017). The model with the lowest LOOIC score was selected as our winning model and used for subsequent analysis. Posterior predictive checks were carried out by simulating choice data based on the parameters derived from the winning model.

The weighting parameter α (alpha) indicated how much participants employed Rawlsian strategy or utilitarian strategy. It can take any value between 0 and 1, and when $\alpha = 0$, it corresponds to pure utilitarian preferences and calculates the difference in utility between two options solely based on the difference in the total amount of harm, and therefore always end up choosing “the one” option that has lower overall harm. On the other hand, when $\alpha = 1$, it takes ‘maximin’ Rawlsian strategy, and the subjective utility of two options will rely on the difference between the amount of harm that each individual gets and thus going for “the group” option that maximizes the minimum welfare for an individual. We compared the model that allowed this parameter to freely have value between 0 and 1 to the models where the weighting parameter is constrained to either 0 or 1 and tested whether people balanced maximin and utilitarian strategy or relied on only one of them in harm allocation decisions.

Next, we considered agreeability parameter (phi), which we assumed to be a separate dimension of Rawlsian moral preference. It is possible that people represent the threshold of agreeability either as the ratio or difference between the amount of harm that individuals get. Therefore, we tested which type of agreeability parameter (constant or ratio term) leads to better fit to the data. If the model with a ratio term excels in model fit, it would suggest that people represent the agreeable amount of differential harm between individuals in a multiplicative manner (e.g., people think it is reasonable for the worst-off individual to receive twice as more harm than the rest). On the other hand, if the model with a constant term excels, it would indicate that people represent the agreeable amount of differential harm between individuals either as a decremental or incremental amount.

Our model comparison result indicated that the model that explained our choie data best was a model with three free parameters which posits the difference in subjective utility between Rawlsian option and Utilitarian option is as following:

$$U(one) = -(1-\alpha)\,C_{one} - \alpha\,(C_{one} + \phi)$$
$$U(group) = -(1-\alpha)\,C_{group}\,N_{group} - \alpha\,C_{group}$$

$$\Delta U = (1-\alpha)\,(\Delta\,total) - \alpha\,(\Delta\,single + \varphi)$$

$$\Delta\,total = C_{group}\,N_{group} - C_{one}$$
$$\Delta\,single = C_{one} - C_{group}$$

Here ΔU denotes the subjective utility of choosing the Rawlsian option over Utilitarian option which was determined by two parameters, alpha and phi, capturing different aspects of Rawlsian moral preferences.The weighting parameter that freely varied between 0 and 1 was able to capture how much participants preferred Rawlsian strategy over Utilitarian strategy, indicating most of participants preferred mixture between two. Additionally, the best model was with agreeability parameter with a constant term, suggesting that the agreeable amount of differential harm between individuals is represented either as a decremental or incremental amount rather than ratio.

After the utility comparison between two options were computed based on those two free parameters, the softmax function was applied to calculate the choice probabilities. For the softmax function, we considered an inverse temperature parameter, τ (tau), to take into account stochasticity in value-based choices.

$$P\ (Choose\ one) = \frac{1}{1 + e^{-\tau \cdot \Delta U}}$$

The winning model was able to accurately reproduce the observed proportion of Rawlsian decisions (*R* = 0.99, *p* <.001). We also carried out the parameter recovery test to check the identifiability of model parameters (Ahn et al., 2017). This was done by comparing the estimated parameters values on simulated data with “true” parameters used to generate the data with our winning model. Based on the winning model and its parameter estimation, the mean of the chosen and unchosen options’ subjective utility was calculated in a trial-by-trial manner and used as regressors of parametric modulation in GLM2.

fMRI acquisition and preprocessing

fMRI scanning was performed on a 3T Siemens Magnetom Trio MRI scanner at Seoul National University, South Korea. Functional images were obtained using a multiband T2*-weighted echo-planar imaging (EPI) sequence using the following parameters : repetition time = 1.5, echo time (TE) = 30 ms, flip angle = 85°, 64 slices, field of view = 256mm, voxel size = 2.3 × 2.3 × 2.3mm, slice thickness = 2.3mm. We acquired 295 volumes in each of three sessions. T1-weighted anatomical images were acquired with magnetization prepared rapid gradient echo (MPRAGE) sequence with following parameters: repetition time = 2.3, echo time = 2.36ms, field of view = 256mm, voxel size = 1 × 1 × 1mm, slice thickness = 1mm).

Preprocessing of fMRI data was performed using fMRIPrep 1.4.0, which is based on *Nipype* 1.2.0 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502). Preprocessed images were spatially smoothed with a Gaussian kernel of FWHM = 8mm and analyzed using SPM12.

GLM1: model of maximin computation

The first GLM was built to determine brain regions showing parametric responses at decision onset, independently from participants' choices, to the objective amount of harm that would be assigned to an individual at maximum from choosing "the one" relative to "the group", as indicated by Rawlsian maximin strategy. fMRI time-series were regressed onto a GLM with the onset of the decision phase as the main event regressor. The event was modeled with a duration corresponding to the participant's RT on that trial. Two parametric modulators were added for each regressor of onset: the objective amount of harm (the amount of time) involved in the one option and in the group option for each individual. The GLM contained five additional event regressors of no interest, describing the onsets of: (i) wait phase, (ii) implement phase 1 (selected option being implemented), (iii) implement phase 2 (selected option being highlighted), (iv) fixation cross, (v) button presses. The event of each phase in the task was modeled with their corresponding duration, while the event of button press was modeled as a stick function (duration = 0s). We included a series of nuisance regressors of no interest to control for motion effects. Specifically, we added six motion regressors obtained during realignment and spike regressors determined on the basis of FD and DVARS thresholds within the fMRIPrep pipeline. The number of spike regressors varied by individual depending on the degree of their head motion during the scanning.

GLM2: model of relative chosen value

We constructed the second GLM to determine brain regions encoding the utility of chosen and unchosen options at decision onset. The utility of options was identified by our winning model. The main event regressor was the onset of the decision phase, which was modeled with a duration corresponding to the participant's RT on that trial. Two parametric modulators were entered for each regressor of onset: the subjective value of the chosen and unchosen options, derived from each participant's choice model. We added the same set of event regressors and nuisance of regressors of no interest as in GLM1.

IS-RSA analysis

To probe whether participants who share similar idea of agreeability also show similar brain responses in encoding harm during moral decision-making, we applied the IS-RSA approach and examined the association between similarity in agreeability parameter and the similar neural responses. The analysis was performed with dyadic regression models implemented on R code adapted from van Baar et al. (2021). The main aim of running the dyadic regression model was to identify brain regions where similarity in agreeability parameter predicted similar neural engagement (van Baar et al., 2020; 2021). We examined the similar neural engagement in parametric response to harm assigned to individual in the worst-off position. We controlled for the similarity in the maximin parameter to eliminate the confounding effect caused by similarity in choice behavior.

To this end, first, we created a geometric representation of individual variabilities in variables of our interest: each dimension of Rawlsian preference (separately for agreeability dimension and maximin dimension), neural similarity in parametric BOLD response to worst-off harm amount for each voxel we obtained from aforementioned GLM. The geometric representations for those variables were obtained by computing similarity between all pairs of participants calculated as the euclidean distance from the observation of two participants converted to a similarity score range between 0 and 1 using the following formula:

$$\text{similarity score} = \frac{1}{1+d(s_1,s_2)}$$

where $d(s_1, s_2)$ is the euclidean distance between two subjects s_1 and s_2

This procedure allowed us to be all our variables of interest expressed on the same range between 0 and 1.

The dyadic regression models mapped the geometric representation of behavioral data to brain data, by regressing similarity in beta estimates onto similarity in agreeability parameter with covariates of similarity in maximin parameter. This allowed us to identify similar multivariate neural patterns that were specifically associated with similarity in agreeability parameter, while controlling for similarity in maximin strategy. Random intercepts of participants were included in our dyadic regression model to correct for statistical dependencies which arise from the repeated occurrence of individual data in pair-wise observation. The dyadic regression model was estimated for each individual voxel, where the observations corresponded to all possible pairs of participants without repetition. The beta estimates of agreeability dimension from dyadic regression models were mapped onto 3D brain space to generate the resulting beta map. We tested for statistical significance of the beta map using voxel-wise thresholding at *p* (FDR) < .05 and reported surviving clusters with the size of 5 or more contiguous voxels, following the same threshold used in a recent work employed the same approach (van Baar et al., 2021).

Supplementary Tables and Figures

Supplementary Table 1

Model	Parameters	LOOIC
Model 1 (Best Model)	$U(one) = -(1-\alpha) \cdot C_{one} - \alpha\ (C_{one} + \ \phi\ )$ $U(group) = -(1-\alpha) \cdot C_{group} \cdot N_{group} - \alpha \cdot C_{group}$ $\Delta U = (1\ -\ \alpha) \cdot \Delta total\ -\ \alpha \cdot (\Delta single +\ \varphi)$ $\Delta total = C_{group} \cdot N_{group}\ - C_{one}$ $\Delta single = C_{one}\ - C_{group}$ $P\ (Choose\ one) = \dfrac{1}{1 + e^{-\tau\ \cdot \Delta U}}$	6112.0

Model 2 (alpha constrained to 0)	$U(one) = -C_{one}$ $U(group) = -C_{group} \cdot N_{group}$ $\Delta U = \Delta total$ $\Delta total = C_{group} \cdot N_{group} - C_{one}$ $P\ (Choose\ one) = \frac{1}{1 + e^{-\tau \cdot \Delta U}}$	12989.0
Model 3 (alpha constrained to 1)	$U(one) = -C_{one}$ $U(group) = -C_{group}$ $\Delta U = -\Delta single$ $\Delta single = C_{one} - C_{group}$ $P\ (Choose\ one) = \frac{1}{1 + e^{-\tau \cdot \Delta U}}$	10400.4
Model 4 (no agreeability parameter)	$U(one) = -(1 - \alpha) \cdot C_{one} - \alpha \cdot C_{one}$ $U(group) = -(1 - \alpha) \cdot C_{group} \cdot N_{group} - \alpha \cdot C_{group}$ $\Delta U = (1 - \alpha) \cdot \Delta total - \alpha \cdot \Delta single$ $\Delta total = C_{group} \cdot N_{group} - C_{one}$ $\Delta single = C_{one} - C_{group}$ $P\ (Choose\ one) = \frac{1}{1 + e^{-\tau \cdot \Delta U}}$	6653.5
Model 5 (agreeability parameter as multiplicative term)	$U(one) = -(1 - \alpha) \cdot C_{one} - \alpha \cdot \varphi \cdot C_{one}$ $U(group) = -(1 - \alpha) \cdot C_{group} \cdot N_{group} - \alpha \cdot C_{group}$ $\Delta U = (1 - \alpha) \cdot \Delta total - \alpha \cdot \Delta single$ $\Delta total = C_{group} \cdot N_{group} - C_{one}$ $\Delta single = C_{one} \cdot \varphi - C_{group}$ $P\ (Choose\ one) = \frac{1}{1 + e^{-\tau \cdot \Delta U}}$	6463.4

Supplementary Table 2 : GLM1 : Cone > Cgroup

Region Name	Extent	t-value	x	y	z
Supra Marginal Gyrus / rTPJ	678	5.50	66	-49	30
Mid Frontal Gyrus	154	4.52	-34	37	41
		3.53	-27	49	43
Inf Frontal Gyrus	252	4.30	-31	40	-12
		3.57	-24	49	-10
		3.44	-29	68	-7
Mid Temporal Gyrus	173	4.14	62	-18	-12
Superior Frontal Gyrus	136	3.93	20	28	64
		3.73	22	19	69
dmPFC	392	3.84	8	47	27
		3.82	11	49	4
		3.57	15	40	32
vlPFC	172	3.76	48	37	-16
		3.55	52	47	-14

Superior Frontal Gyrus/ dlPFC	53	3.64	31	40	48

Supplementary Table 3 : GLM1, multiple regression : Cone > Cgroup positive correlation with alpha

Region Name	Extent	t-value	x	y	z
Inferior Frontal Gyrus / dlPFC	2720	7.88	57	12	27
		6.42	45	2	27
		5.79	41	37	20
Superior Parietal Lobe/ Precuneus	2577	7.10	31	-60	48
		6.86	50	-39	57
		6.11	29	-49	41
dmPFC / Superior Frontal Gyrus	1380	6.29	1	19	55
		5.85	11	26	34
		5.58	-8	12	53
lTPJ / Inferior Parietal Lobule	1638	6.25	-45	-42	48
		5.66	-27	-72	55
		5.55	-22	-60	48
Insula	556	5.84	34	19	9
		5.77	34	23	-5

		5.30	45	16	7
Mid Temporal Gyrus	247	5.44	55	-39	-16
		5.02	50	-53	-12
Insula	1212	5.13	-31	28	7
		5.00	-52	35	25
		4.90	-48	7	32
Fusiform Gyrus	386	4.75	-38	-74	-7
		4.31	-34	-65	-28
Mid Frontal Gyrus	348	4.37	-22	-5	57
		4.11	-27	-14	50
		3.90	-38	0	55

Supplementary Table 4 : GLM1, multiple regression: Cone > Cgroup ; negative correlation with alpha

Region Name	Extent	t-value	x	y	z
vmPFC	288	5.59	-1	61	-12
		4.47	11	51	-14
		3.61	6	37	-12
Inferior Frontal Gyrus / vlPFC	204	4.90	41	9	-21
		4.79	52	-7	-21

Supplementary Table 5 : GLM2, Chosen utility > Unchosen utility

Region Name	Extent	t-value	x	y	z

Posterior Cingulate Cortex	3744	7.27	20	-46	13
		6.66	-10	-56	16
		6.61	25	-42	20
Caudate	1552	7.07	6	16	-7
		6.52	4	26	-7
		6.15	6	37	-5
Mid Temporal Gyrus	226	6.23	-45	-77	30
		3.61	-38	-63	27
Mid Temporal Gyrus	316	5.62	-59	-5	-14

Supplementary Table 6: IS-RSA analysis result showing similar multivariate patterns of brain activity in agreement

Region Name	Extent	t-value	x	y	z
Cingulate Gyrus	144	5.76	-20	-40	30
Caudate	72	5.41	-22	-42	10
Caudate	176	5.38	-20	-20	32
Cerebellar Tonsil	64	5.09	-10	-44	-60

Supplementary Table 7 : PPI Choose Rawlsian > Choose Utilitarian ; positive correlation with alpha

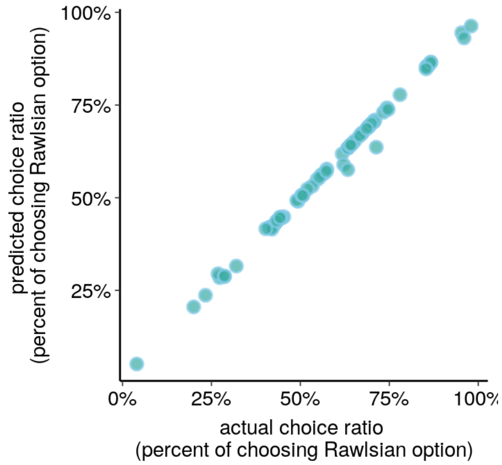
Region Name	Extent	t-value	x	y	z
dmPFC	292	5.27	4	54	16
		5.08	4	54	34

		4.91	6	63	20

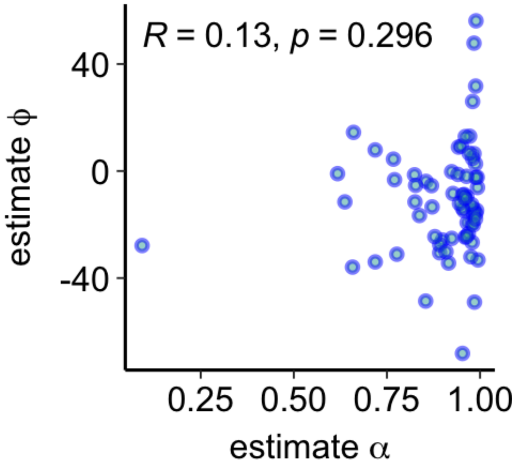
Supplementary Table 8: PPI result for contrast of Choose Rawlsian > Choose Utilitarian ; negative correlation with alpha

Region Name	Extent	t-value	x	y	z
Fusiform Gyrus	27962	7.54	48	-44	-14
		7.35	-15	12	30
		7.27	13	-53	0
Cingulate Gyrus	254	5.41	-13	-23	43
		5.20	11	-37	39
		5.18	-8	-32	39
Anterior Cingulate Cortex	154	4.62	-13	42	0
		4.26	-3	54	-3

Supplementary Figure 1.
The correlation between actual choice ratio and model predicted choice

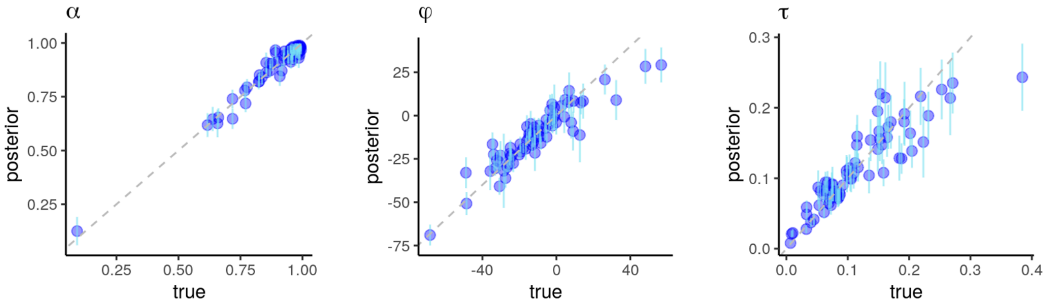


Supplementary Figure 2.
The correlation between two model parameters reflecting different dimensions of Rawlsian moral preference

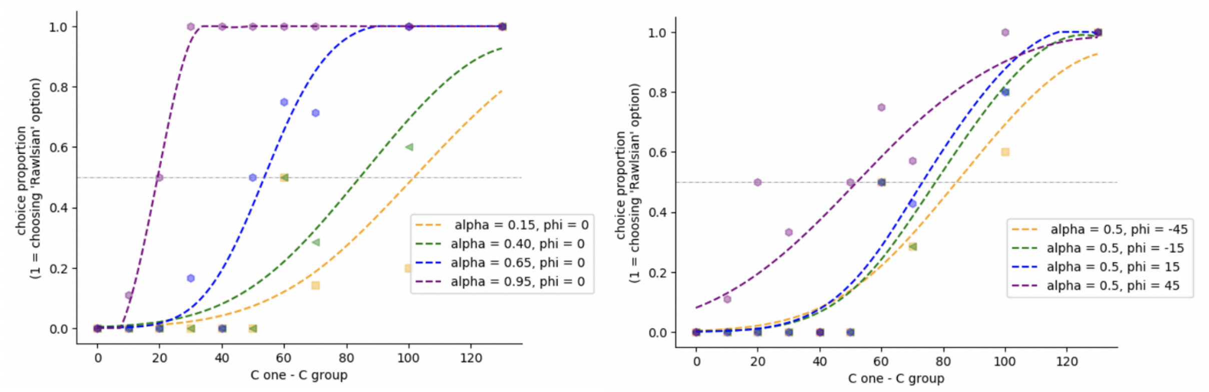


c

Supplementary Figure 3.
Parameter Recovery for three model parameters

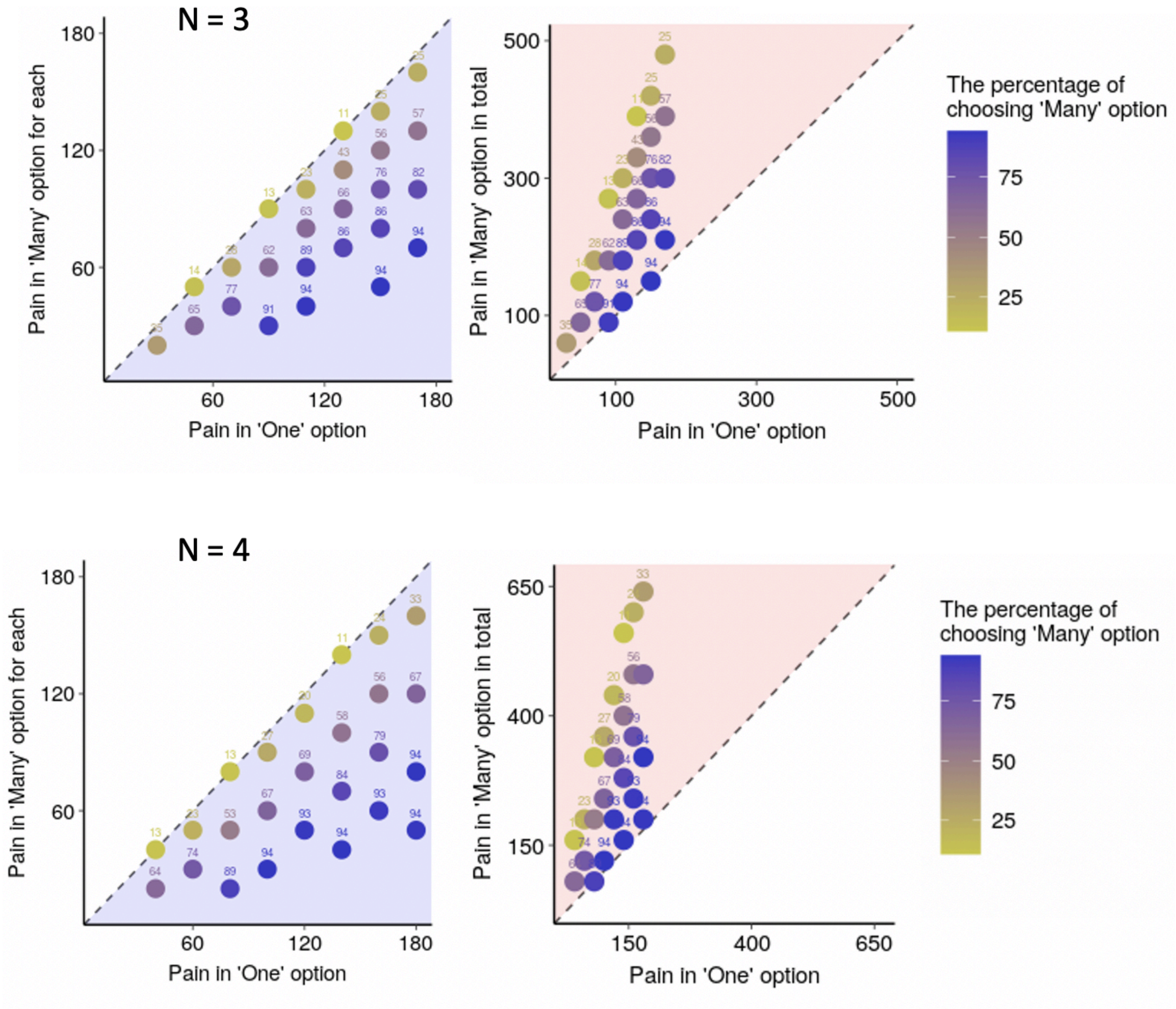


Supplementary Figure 4. The relationship between model parameters and choice in the task using simulation



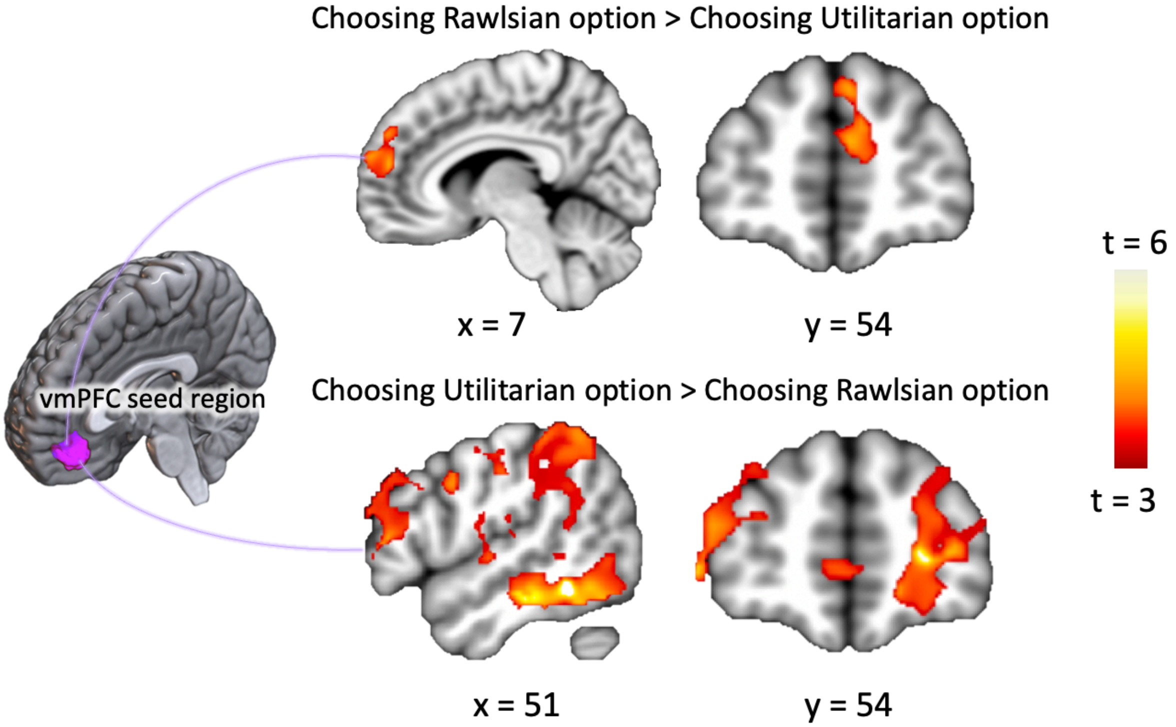
We generated the simulated choice data with the winning model by changing parameter values one at a time while fixing another to examine the respective roles of alpha and phi parameters on choices across our task parameter space. We fitted the simulated data to the cumulative normal distribution as a function of relative difference in the amount of harm for an individual when choosing the one compared to for choosing the group as in maximin computation ($C_{one} > C_{group}$). This revealed that varying alpha in our model changed the slope of the cumulative normal distribution, modifying the sensitivity to the relative difference in harm given to the individual. On the other hand, Phi had an effect on adjusting the point of subjective equivalence, changing the point in which people regard fair to assign the harm more or less to a single individual.

Supplementary Figure 5. Task parameter and choice ratio



Preference for Rawlsian option as a function of harm for one vs. group separated by the number of people involved in ‘many’ option.

Supplementary analysis : PPI analysis


Choosing Rawlsian option > Choosing Utilitarian option
x = 7
y = 54
t = 6
t = 3
vmPFC seed region
Choosing Utilitarian option > Choosing Rawlsian option
x = 51
y = 54

Heightened functional connectivity between vMPFC and a set of brain regions in mentalizing regions as a function of choice type was modulated by individual's Rawlsian moral preference. People with higher Rawlsian preference showed the greater increase in connectivity between dmPFC and vmPFC when making Rawlsian choice compared to Utilitarian choice. Rawlsian preference was also associated with heightened connectivity between vmPFC and brain regions in mentalizing regions when making utilitarian choice compared to Rawlsian choice.

Functional connectivity responsive to moral decision

Our results from parametric brain responses to maximin computation provided evidence that concern for the worst-off individual in harm allocation decision engages brain regions implicated in mentalizing. Moreover, individual differences in Rawlsian preference for prioritizing the reduction of the worst off individual's suffering over aggregate utility was associated with differential neural sensitivity to the maximin computation, with neural activity in mentalizing and valuation network scaling with such preferences respectively in positive and negative directions.

These findings motivated us to hypothesize that, given the role of vmPFC in encoding all-things-considered decision value (Shenhav & Greene, 2014; Hutcherson et al., 2015), individual differences in Rawlsian preference may be explained by the degree to which one incorporates concerns for the worst-off individual into integrative values of moral decision through mentalizing. This possibility is also consistent with previous works showing that social information represented in brain areas specialized for encoding social functions modulates computation of decision value in domain-general valuation regions (Izuma, Saito & Sadato, 2008; Lin, Adolphs & Rangel, 2011; Ruff & Fehr, 2014; Zoh, Chang, & Crockett, 2021). Therefore, we predicted that mentalizing region will express differential functional connectivity with vmPFC, during Rawlsian choice relative to uUtilitarian choice as a function of Ralwisan preference. To test this, we implemented psychophysiological interaction (PPI) analyses with vmPFC as a seed region and added individual differences in moral preferences as second (group) level regressor onto the contrast between the trials where Rawlsian option was chosen over Utilitarian one.

When choosing Rawlsian option over Utilitarian one, participants with stronger Rawlsian preferences showed greater increase in connectivity between vmPFC and dmPFC that has been implicated in mentalizing. This supported our prediction that Rawlsian preference is explained by the degree to which one engages in mentalizing during moral decision-making, through which they may incorporate concerns for the worst-off individual in their decisions. Interestingly, we also found that, even in the case of making Utilitarian option over Ralwisan option, Rawlsian preference was associated with greater increase in connectivity between vmPFC and TPJ and IPL, areas of which also encompass mentalizing network (Saxe, 2006; Koster-Hale & Saxe, 2013; Jamali et al., 2021; Schurz et al., 2014; Schurz et al., 2021). Together, this suggests that Rawlsian preference is associated with greater engagement of mentalizing region in the brain, regardless of choices participants made.

PPI model: functional connectivity with vmPFC

We applied the method of generalized psychophysiological interactions (gPPI, McLaren et al., 2012) to determine brain regions with which differential functional connectivity with vmPFC as a function of choice type (choosing Ralwisan option >

Utilitarian option) is modulated by individual differences in Rawlsian preference. Nine participants were excluded from PPI analysis, since they did not provide sufficient variation in their choices to allow for contrast estimation. For the gPPI analysis, we first constructed vmPFC region-of-interest (ROI) by masking a contrast from GLM2, which showed parametric effects of relative chosen value, with a *priori* meta-analysis map of vmPFC carrying a SV signal (5-way conjunction analysis map) in Bartra et al., (2013). Next, we built a gPPI model which contained PPI regressors for the event of choosing Rawlsian option and choosing Utilitarian option. The model also contained five additional event regressors of no interest for each phase of the task. The contrast of choosing Rawlsian option compared to Utilitarian option was calculated using the model. We regressed parameter alpha onto the contrast to look for brain regions with which they showed increased functional coupling during the decision onset of the trial where Rawlsian option was chosen relative to the Utilitarian option. All the results with univariate fmri analysis reported in the text survived whole-brain correction for multiple comparisons ($P < 0.05$, FWE-corrected at the cluster level after voxel-wise thresholding at $P < 0.001$)

REFERENCES

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences*, *17*(1), 1-10.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389-400.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363-366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273-292.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford*, *5*, 5-15.

Jarvis Thomson, J. (1985). The trolley problem.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in cognitive sciences*, *11*(8), 322-323.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144-1154.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908-911.

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., ... & Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of personality and social psychology*, *120*(2), 443.

Baumard, N., & Sheskin, M. (2015). Partner choice and the evolution of a contractualist morality. *The moral brain: a multidisciplinary perspective*, *20*, 35-48.

Baumard, N. (2016). *The origins of fairness: How evolution explains our moral nature*. Oxford University Press.

Levine, S., Chater, N., Tenenbaum, J. B., & Cushman, F. (2023). Resource-rational contractualism: A triple theory of moral cognition. *Behavioral and Brain Sciences*, 1-38.

Rawls, J. (1971). A Theory of Justice. 2. *The Law of Peoples*, *67*.

Baron, J., & Jurney, J. (1993). Norms against voting for coerced reform. *Journal of Personality and Social Psychology*, *64*(3), 347.

Baron, J. (1995). Blind justice: Fairness to groups and the do‐no‐harm principle. *Journal of behavioral decision making*, *8*(2), 71-83.

Konow, J. (2001). Fair and square: the four sides of distributive justice. Journal of Economic Behavior & Organization, 46(2), 137-164.

Sandel, M. (2009). Justice: What's the right thing to do?.

Baumard, N., & Sheskin, M. (2015). Partner choice and the evolution of a contractualist morality. *The moral brain: a multidisciplinary perspective*, *20*, 35-48.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772.

Kogut, T., & Ritov, I. (2005). The singularity effect of identified victims in separate and joint evaluations. *Organizational behavior and human decision processes*, *97*(2), 106-116.

Lee, S., & Feeley, T. H. (2016). The identifiable victim effect: A meta-analytic review. *Social Influence*, *11*(3), 199-215.

Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes, 102*(2), 143–153.

Schelling, T. C. (1968). The life you save may be your own. *Problems in public expenditure*, 127-162.

Västfjäll, D., Slovic, P., Mayorga, M., & Peters, E. (2014). Compassion fade: Affect and charity are greatest for a single child in need. *PloS one*, *9*(6), e100115.

Slovic, P., Västfjäll, D., Erlandsson, A., & Gregory, R. (2017). Iconic photographs and the ebb and flow of empathic response to humanitarian disasters. *Proceedings of the National Academy of Sciences*, *114*(4), 640-644.

Fetherstonhaugh, D., Slovic, P., Johnson, S., & Friedrich, J. (1997). Insensitivity to the value of human life: A study of psychophysical numbing. *Journal of Risk and uncertainty*, *14*(3), 283-300.

Slovic, P. (2013). The more who die, the less we care. In *The feeling of risk* (pp. 69-77). Routledge.

Wiss, J., Andersson, D., Slovic, P., Västfjäl, D., & Tinghög, G. (2015). The influence of identifiability and singularity in moral decision making. *Judgment and Decision Making*, *10*(5), 492-502.

Ye, Z., Heldmann, M., Slovic, P., & Münte, T. F. (2020). Brain imaging evidence for why we are numbed by numbers. *Scientific Reports*, *10*(1), 9270.

Kameda, T., Inukai, K., Higuchi, S., Ogawa, A., Kim, H., Matsuda, T., & Sakagami, M. (2016). Rawlsian maximin rule operates as a common cognitive anchor in distributive justice and risky decisions. *Proceedings of the National Academy of Sciences*, *113*(42), 11817-11822.

Parfit, D. (2011). *On what matters* (Vol. 1). Oxford University Press.

Scanlon, T. M. (2000). *What we owe to each other*. Belknap Press.

Habermas, J. (1990). *Moral consciousness and communicative action*. MIT press.

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in cognitive sciences*, *18*(10), 512-519.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, *117*(42), 26158-26169.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. 2. Chapman & Hall/CRC; 2004.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. Journal of Mathematical Psychology, 55, 1–7.

Lockwood, P. L., Apps, M. A., & Chang, S. W. (2020). Is there a 'social'brain? Implementations and algorithms. *Trends in cognitive sciences*, *24*(10), 802-813.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Saxe, R. (2006). Uniquely human social cognition. *Current opinion in neurobiology*, *16*(2), 235-239.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, *79*(5), 836-848.

Jamali, M., Grannan, B. L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., & Williams, Z. M. (2021). Single-neuronal predictions of others' beliefs in humans. *Nature*, *591*(7851), 610-614.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9-34.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., ... & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological bulletin*, *147*(3), 293.

Frohlich, N., & Oppenheimer, J. A. (1992). *Choosing justice: An experimental approach to ethical theory* (Vol. 22). Univ of California Press.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, *94*(4), 857-869.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The quarterly journal of economics*, *117*(3), 817-869.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive science*, *36*(1), 163-177.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108.

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the national academy of sciences*, *116*(48), 23989-23995.

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, *111*(48), 17320-17325.

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: distributive justice and neural encoding of equity and efficiency. *science*, *320*(5879), 1092-1095.

Kappes, A., Nussberger, A. M., Siegel, J. Z., Rutledge, R. B., & Crockett, M. J. (2019). Social uncertainty is heterogeneous and sometimes valuable. *Nature Human Behaviour*, *3*(8), 764-764.

Davis, A. L., Miller, J. H., & Bhatia, S. (2018). Are preferences for allocating harm rational?. *Decision*, *5*(4), 287.

van Baar, J. M., Halpern, D. J., & FeldmanHall, O. (2021). Intolerance of uncertainty modulates brain-to-brain synchrony during politically polarized perception. *Proceedings of the National Academy of Sciences*, *118*(20), e2022491118.

van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature communications*, *10*(1), 1483.

Ahn, W. Y., Haines, N., & Zhang, L. (2017). Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry (Cambridge, Mass.)*, *1*, 24.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*(5), 1413-1432.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature neuroscience*, *16*(1), 105-110.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review*, *117*(3), 864.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience*, *34*(13), 4741-4749.

Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J., & Rangel, A. (2015). Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in ventromedial prefrontal cortex. *Journal of Neuroscience*, *35*(36), 12593-12605.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, *58*(2), 284-294.

Lin, A., Adolphs, R., & Rangel, A. (2012). Social and monetary reward learning engage overlapping neural substrates. *Social cognitive and affective neuroscience*, *7*(3), 274-281.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549-562.

Zoh, Y., Chang, S. W., & Crockett, M. J. (2022). The prefrontal cortex and (uniquely) human cooperation: a comparative perspective. *Neuropsychopharmacology*, *47*(1), 119-133.

McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *Neuroimage*, *61*(4), 1277-1286.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, *76*, 412-427.