Representing Expected Moral Value: Outcome Magnitude, Probability, and Expected Value in the Context of Moral Judgment

Amitai Shenhav, Joshua D. Greene

Dept of Psychology, Harvard University

682.12

Background

- Neuroeconomic studies have sought to isolate brain regions responsible for representing risk, reward magnitude, and overall expected value in order to inform affective judgments. Almost all of these studies have utilized stimuli that involve the opportunity for (tangible) personal gain or loss.
- "Trolley"-type moral dilemmas present individuals with choices in which promoting the "greater good" (maximizing the number of lives saved) requires causing the death of another (e.g., pulling a switch that causes a trolley to run over one worker rather than the five it would otherwise hit). Determining the acceptability of an action in such cases must then rely to some degree on an evaluation of expected gains and/or losses of lives that would result.

Purpose

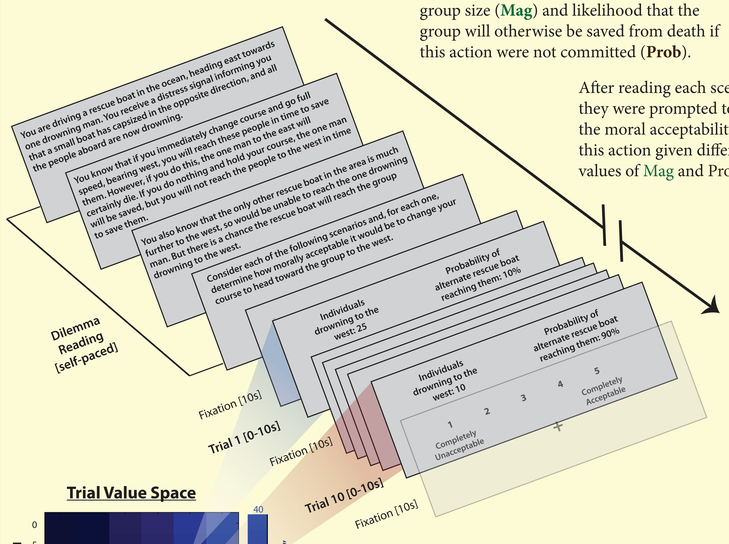
To determine which brain regions are involved in representing **magnitude**, **probability**, and **expected value** in the context of moral dilemmas in which the decision-maker is an unaffected third party with no material stake in the outcome.

Behavioral Task

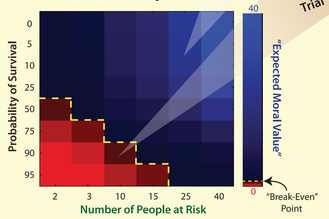
Subjects were presented with 5 different scenarios in which they evaluated the moral acceptability of a given action within the context of that scenario. In all scenarios the action would result in the certain death of one individual in the service of preventing the uncertain deaths of a group of other individuals. (Example run shown below)

The scenario descriptions left unstated the group size (**Mag**) and likelihood that the group will otherwise be saved from death if this action were not committed (**Prob**).

After reading each scenario, they were prompted to rate the moral acceptability of this action given different values of Mag and Prob.



Trial Value Space



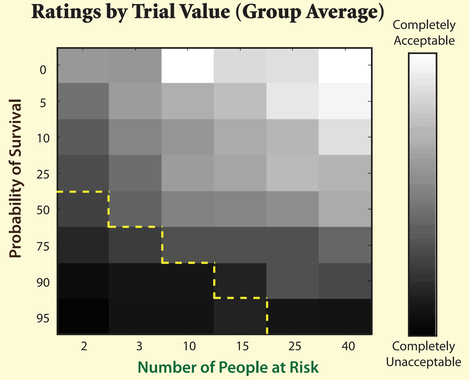
Mag and Prob values for each trial were chosen randomly from 50 unique pairs derived from a full factorial model of 6 Mag X 8 Prob values (plus two additional catch trials of Prob=100%).

"Expected moral value" (EV) of a given trial can be calculated as the interaction between the Mag and Prob parameters (as shown on left). The trials for which the EV of action exceeds the EV of inaction are shown in blue.

Behavioral Results

- When entered into a mixed-effects multiple regression, subjects' ratings of moral acceptability were shown to be highly sensitive to Mag ($t(34,1) = 9.71$), Prob ($t(34,1) = 15.3$), and EV ($t(34,1) = 4.29$) [all p-values < 0.0001)].
  - This is especially noteworthy given that these values were randomly distributed between very different dilemma contexts.

Ratings by Trial Value (Group Average)



- Subjects' trial RT's, however, were shown not to be significantly influenced by Mag and Prob ($t(34,1) = -1.57$, $p=0.13$; $t(34,1) = 1.26$, $p=0.22$), but were faster as increased EV ($t(34,1) = -3.98$, $p<0.0005$)

Neuroimaging Methods

- **35 Subjects** (16 male, ages 18-42) each performed five fMRI runs
- Whole-brain fMRI was performed on a 3T Siemens Trio magnet with the following parameters:
  - TR = 2.5s, TE = 28ms, FOV = 256x256mm, Matrix = 96x96
  - 36 slices, thickness = 3mm, gap = 0.5mm
- Data were concatenated across runs within each subject and trials were analyzed using a variable-duration canonical (SPM) HRF within a multiple regression model that included regressors for reaction time (to control for differences in time-on-task) as well as:
  - a) **Mag** (natural-log transformed based on best fit to behavioral data)
  - b) **Prob** (coded as probability of group NOT surviving if prescribed action isn't taken, shown as positive direction in y-axis on left)
  - c) **EV** (multiplicative interaction of Mag and Prob as shown in color map on left)

Neuroimaging Results

Below are the statistical maps (thresholded for cluster size of 20 voxels) for positively increasing BOLD activity for each of the 3 regressors* of interest:

p-values (uncorrected)

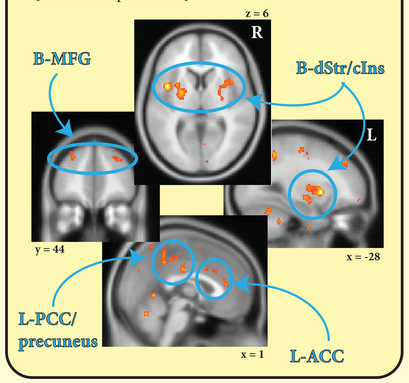
$p < 20^{-3}$ — $p < 10^{-5}$

**none of these regressors produced significant decreases in activation along the same parameter*

Magnitude

(numbers of people who could be saved/killed)

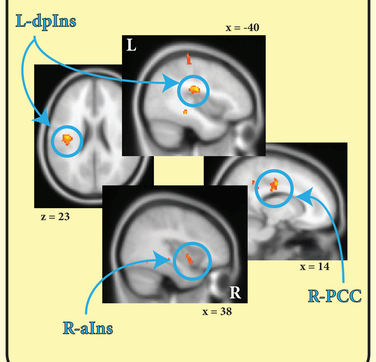
- **Bilateral dorsal striatum** [dStr] (caudate nucleus and putamen) and **central insula** [cIns]
- **Bilateral middle frontal gyrus** [MFG]
- **Anterior and posterior cingulate cortices** [ACC; PCC/precuneus]



Probability

(the probability that the group will die if action is not taken)

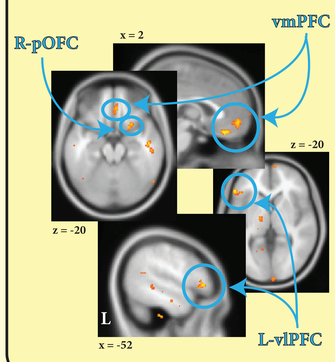
- **Left dorsal posterior insula** [dpIns]
- **Right anterior/ central insula** [aIns]
- **R-PCC**



Expected Value

(overall value of allowing one person to die to save the group)

- **Ventromedial prefrontal cortex** [vmPFC]
- **Left ventrolateral prefrontal cortex** [vlPFC]
- **Right posterior orbitofrontal cortex** [pOFC]



Relevance to Decision-Making Literature

- The dorsal striatum is implicated in operant conditioning, encoding stimulus-action associations based on reward/loss value.
  Balleine, Delgado & Hikosaka, 2007
- Coactivation of insula/putamen has been correlated and causally related to perception/experience of disgust.
  Calder et al., 2000; Thielscher & Pessoa, 2007
- Neuroeconomic studies have shown regions of putamen, ACC, and MFG to parametrically increase as the magnitude of expected monetary gains and losses.
  Knutson et al., 2005; Tom et al., 2007
- The anterior insula (particularly R-aIns) has been implicated in perceptions of risk (and uncertainty) and personality constructs that center around avoidance thereof.
  Paulus et al., 2003; Huettel et al., 2006
- The L-dpIns also activates in certain contexts involving risk, but more generally shows activation (jointly with contralateral aIns) correlated with parametrically increasing aversiveness of somatosensory stimuli.
  Craig, 2002; Paulus et al., 2003
- Studies in both humans and non-human primates have shown the mPFC and OFC to play a role in calculating expected value of both primary and secondary reinforcers and integrating the determined value into present and future behavior.
  Knutson et al., 2005; Wallis, 2007

Conclusions

1) Subjects were highly sensitive to "expected moral value" and its components (the number of lives saved and probability of their death) when making judgments of the moral acceptability of sacrificing the life of another.

2) Brain regions whose activations were differentially correlated with increases in each of these parameters shared a number of commonalities with those implicated in decisions involving real reward or loss for the subject. **This is despite the fact that our task involved judgments that were a) hypothetical and b) concerned only with effects on individuals other than the subject.**

References

Balleine BW, Delgado MR, Hikosaka O (2007). *The Role of the Dorsal Striatum in Reward and Decision-Making.* J Neurosci 27 (31): 8161-65.

Calder AJ, Keane J, Manes F, Antoun N, Young AW (2000). *Impaired recognition and experience of disgust following brain injury.* Nat Neuro 3 (11): 1077-78.

Craig AD (2002). *How do you feel? Interoception: the sense of the physiological condition of the body.* Nat Rev Neuro 3 (8): 655-66.

Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML (2006). *Neural Signatures of Economic Preferences for Risk and Ambiguity.* Neuron 49: 765–75

Knutson B, Taylor J, Kaufman M, Peterson R, Glover G (2005). *Distributed Neural Representation of Expected Value.* J Neurosci 25 (19):4806-12

Paulus MP, Rogalsky C, Simmons A, Feinstein JS, Stein MB (2003). *Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism.* Neuroimage 19 (4): 1439-48

Thielscher A, Pessoa L (2007). *Neural correlates of perceptual choice and decision making during fear-disgust discrimination.* J Neurosci 27 (11): 2908-17.

Tom SM, Fox CR, Trepel C, Poldrack RA (2007). *The Neural Basis of Loss Aversion in Decision-Making Under Risk.* Science 315: 515-18.

Wallis JD (2007). *Orbitofrontal Cortex and Its Contribution to Decision-Making.* Annu Rev Neuro 30: 31-56.