LEBS
Vol.4 No.2 (2013) 9-12
Letters on Evolutionary Behavioral Science

Refusal of Killing a Stranger to Save Five Brothers: How Are Others' Judgments Anticipated and Favored in a Moral Dilemma Situation?

Ryo Oda*

Academic affiliation: Nagoya Institute of Technology, Nagoya 466-8555, Japan

*Author for correspondence (oda.ryo@nitech.ac.jp)

One evolutionary theory of morality, examined here, is based on theories of kin selection while another has proposed that moral judgment is based on a Kantian rule-based system. Using the Trolley Problem, Kurzban et al. (2012) asked subjects to decide whether they would kill one person to save five others, varying the relationship of the subject with the others involved. They revealed that nearly half of the subjects reported that they would be unwilling to push one stranger to his/her death to save five brothers in a footbridge version of the Trolley Problem. That is, nearly half of the subjects thought that moral rules should not be broken even if they sacrificed Hamiltonian inclusive fitness. In the present study, I tried to replicate this somewhat surprising result in Japanese participants, to investigate the robustness of the finding. I also examined how participants anticipated and favored the moral judgment of other people. If a Kantian decision was made according to the coordination system suggested by Kurzban et al. (2012), a Kantian decision, rather than a Hamiltonian decision, would be anticipated and favored as the decision of people generally. The results seem to support the discussion of Kurzban et al. (2012), that the computational system that delivers Kantian moral judgment functions to coordinate condemnation decisions.

Keywords
morality, moral dilemma, altruism, kin selection, Trolley Problem

Introduction

Kin selection theory (Hamilton, 1964) has the potential to explain aspects of human social behavior. For example, people are less likely to kill kin than non-kin (Daly & Wilson 1988). One evolutionary theory of morality is based on this theory and predicts that moral judgment and behavior are designed to maximize inclusive fitness. Other theories have proposed that moral judgment is rule-based, such as Immanuel Kant's idea of *nonconsequentialism*, and that moral rules should not be broken, regardless of the consequences. Using the Trolley Problem (Foot, 2002), Kurzban, DeScioli, and Fein (2012) asked subjects to decide whether they would kill one person to save five others, varying the relationship of the subject with the others involved. They proposed that the Kantian rule-based structure of moral cognition is not explained by kin selection, reciprocity, or other altruism theories.

A surprising result of their study was that nearly half of the subjects reported that they would be unwilling to push one stranger to save five brothers in a footbridge version of the Trolley Problem. They asked 94 participants to decide whether they would push a stranger to his/her death and 43.6% of them answered that they would not. They suggested that a kin-selection system and a moral system have distinct functions. DeScioli and Kurzban (2009) argued that condemnation mechanisms causally precede conscience, and that conscience functions, at least in part, as a defense system, designed to avoid attacks from third-party condemners. However, the third-party condemners have the adaptive problem that they should coordinate their condemnation decisions with others because condemning a wrongdoer alone involves a greater risk of retaliation. Kurzban et al. (2012) suggested that the computational system that delivers Kantian moral judgment functions to coordinate condemnation decisions. They stated that, "In such a case, agents should use whatever structural features of the moral situation that others are using, *independent of the welfare consequences of those structural features*" (p. 333). They did not, however, investigate in detail what their participants thought about the morality of others. If a Kantian decision was made according to a coordination system, a Kantian decision, rather than a Hamiltonian decision, would be anticipated and favored as the decision of people in general. The tendency would be stronger in participants who would not push the stranger to his/her death than in those who would push the stranger.

There were two main objectives of the present study. First, I tried to replicate the results of Kurzban et al. (2012) using Japanese participants because moral judgment is affected by cultural backgrounds (e.g., Haidt, Koller, & Dias, 1993). While Kurzban et al. (2012) recruited their participants from Amazon's "Mechanical Turk" and did not describe their cultural backgrounds, I

doi: 10.5178/lebs.2013.26.
Received 1 August 2013.
Accepted 7 August 2013.
Published online 2 September 2013.
© 2013 by Human Behavior and Evolutionary Society of Japan

believe that relatively few Japanese participants were included. Second, I investigated how my participants anticipated and favored the moral judgment of other people. I asked participants not only to judge by themselves but also to anticipate how people in general would answer the footbridge version of the Trolley Problem. I also asked them to evaluate impressions of two imaginary persons: a person who answered that s/he would push the stranger and a person who answered s/he would not.

Methods

Participants

In total, 115 Japanese undergraduates (50 males, 65 females, mean age: 19.4±0.9) at two universities participated. They responded to a paper and pencil questionnaire in classrooms and received no monetary reward for their involvement.

Questionnaire

I used the "footbridge dilemma" version of vignettes Kurzban et al. (2012) had used in which participants faced the Trolley Problem, whereby they could push and kill one stranger to save five brothers. The vignette was translated by the author into Japanese and presented as a booklet with a series of questions. Following Kurzban et al. (2012), I first asked subjects to report whether they would push and kill the person. Next, I asked participants to indicate whether pushing the individual onto the tracks was morally wrong and also whether *not* pushing the individual onto the tracks was morally wrong. Then, I asked participants to compare pushing and not pushing and asked which of the two was *more* morally wrong, and asked participants to evaluate the moral wrongness of each act on a 1-7 scale. Finally, I asked whether the participants would want someone else to push the stranger instead of the participants themselves if someone else were on the footbridge. Following these questions, I asked participants to answer as they anticipated people in general would answer the same questions.

After answering the decision of themselves and people in general, participants were presented with two vignettes that described two persons. One described a person who answered that s/he would push a stranger (referred to as 'Hamiltonian' hereafter), and the other described a person who answered s/he would not push the stranger (referred to as 'Kantian' hereafter). The genders of the two persons were made ambiguous. Participants were requested to evaluate favorable impressions of 'Hamiltonian' and 'Kantian' by a nine-grade evaluation as well as to anticipate impressions evaluated by people in general.

Results

Like the results of Kurzban et al. (2012), nearly half of my participants reported that they would be unwilling to push one stranger to save five brothers (Table 1). While smaller numbers of my participants answered that pushing was wrong (85.1% vs. 74.8%), there was no significant difference between the results of Kurzban et al. (2012) and my study. On the other hand, significantly fewer participants in my study answered that not pushing was wrong (66.0% vs. 20.9%; Fisher's $p < .001$). Although Kurzban et al. (2012) reported that women were less likely than men to say that it is wrong not to push, 14 of 65 women (21.5 %) and 10 of 50 men (20.0%) in my study answered that not pushing was wrong, which revealed no gender difference. While 63.8% of the subjects in Kurzban et al. (2012) answered that pushing was worse than not pushing, 79.1% of my participants reported that pushing was worse (Fisher's $p = .019$). Although more participants wanted someone else to push in Kurzban et al. (2012) than the subjects in my study, there was no significant difference (73.4% vs. 65.2%; Fisher's $p = .297$).

Table 1. Push a stranger to save brothers: self

N	115
Would you push?	56.5%
Is it wrong to push?	74.8%
Is it wrong not to push?	20.9%
How wrong is pushing?	5.3 (1.3)
How wrong is not pushing?	3.4 (1.6)
Is pushing worse?	79.1%
Would you want someone else to push?	65.2%

Table 2. Push a stranger to save brothers: people in general

Push	Yes	No	All
N	65	50	115
Would people in general push?	80.0%	36.0%	60.9%
Would people in general think it wrong to push?	69.2%	68.0%	68.7%
Would people in general think it wrong not to push?	38.5%	14.0%	27.8%
How people in general think pushing wrong?	5.1 (1.4)	5.0 (1.3)	5.0 (1.3)
How people in general think not pushing wrong?	3.9 (1.5)	3.5 (1.4)	3.7 (1.5)
Would people in general think pushing worse?	66.2%	76.0%	70.4%
Would people in general want someone else to push?	92.3%	68.0%	81.7%

I compared answers of participants who would be willing to push with those of participants who would not (Table 2). Among the 65 participants who answered that they would push the stranger, 52 (80.0%) thought that people generally would also push, while among the 50 participants who reported that they would not push, only 18 (36.0%) imagined that people in general would push the stranger. This difference was significant statistically (Fisher's $p <$ .001). Although there was no significant difference in the proportion of participants who anticipated that people generally would think pushing was wrong, participants who answered that they would not push tended to think that people in general would think pushing was not wrong (Fisher's $p =$ .006). There was no significant difference in the proportion of participants who anticipated that people in general would think pushing was worse than not pushing. On the other hand, significantly more participants who answered they would push thought that people in general would want someone else to push versus the participants who answered they would not push (Fisher's $p = .001$).

Figure 1 shows the favorable impression scores of the 'Hamiltonian' and 'Kantian' choices, rated by participants who would push the stranger and those who would not push. The impression scores were analyzed with an ANOVA, with subject persons ('Hamiltonian' or 'Kantian') and evaluators (themselves or people in general) as within-subject variables, and decision of participants (push or not push) as a between-subjects variable (Table 3). The main effects of subject persons and interaction between subject persons and decisions of participants were significant. Moreover, the second-order interaction was significant. Thus, we analyzed the interaction between subject persons and evaluators in each decision taken by the participants. 'Kantian' was favored more than 'Hamiltonian' when participants who would not push answered ($F(1,49) = 19.557$, $p = .001$, $\eta^2_G = .159$). There was neither a main effect of evaluators ($F(1,49) = 0.654$, $p = .423$, $\eta^2_G = .001$) nor an interaction between evaluators and subject persons

Table 3. Result of three-way ANOVA on the favorable impression score

Factor	$F_{(1,113)}$	p	η^2_G
Main effects			
Subject person	13.75	< .001	.050
Evaluator	0.25	.618	.001
Decision of participant	0.38	.537	.001
Interactions			
Subject × Evaluator	0.25	.616	.001
Subject × Decision	10.47	.002	.038
Evaluator × Decision	0.35	.554	.000
Subject × Decision × Evaluator	5.19	.025	.010

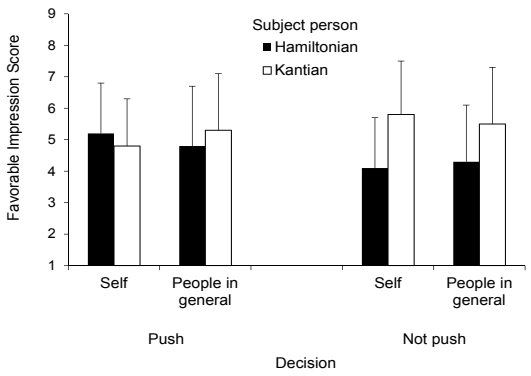

Figure 1. Mean and SE of favorable impression score to 'Hamiltonian' (black bar) and 'Kantian' (white bar) divided by each evaluator and decision of participants.

($F(1,49) = 1.342$, $p = .252$, $\eta^2_G = .006$). However, when participants who would push answered, there was, similarly, neither a main effect of subject persons ($F(1,64) = 0.138$, $p = .712$, $\eta^2_G = .001$) nor evaluators ($F(1,64) = 0.004$, $p = .948$, $\eta^2_G < .001$). The interaction between evaluator and subject person was significant ($F(1,64) = 4.579$, $p = .036$, $\eta^2_G = .015$). The simple effect of the interaction indicated no significant difference between 'Hamiltonian' and 'Kantian' when participants who would push evaluated themselves ($F(1,64) = 2.477$, $p = .121$, $\eta^2_G = .014$) whereas 'Kantian' was favored more than 'Hamiltonian' when the participants considered the impression of people in general ($F(1,64) = 5.218$, $p = .025$, $\eta^2_G = .016$).

Discussion

In this study, 43.5% of participants refused to push a stranger to save five brothers. This frequency, surprisingly, agreed closely with that of Kurzban et al. (2012), suggesting that this supposed inconsistency with kin selection theory is, in fact, robust despite differences in cultural backgrounds. However, a possible reason of this replication could be that the case I used in this study was that leading the most extreme Hamiltonian decision by contrasting a stranger with brothers and that this might weaken potential cultural differences.

The moral judgment regarding pushing did differ from the previous study. While there was no significant difference in the number of participants who thought pushing was wrong, significantly fewer participants in my study answered that not pushing was wrong than did participants in the study of Kurzban et al. (2012). Moreover, more of my participants reported that pushing was worse than not pushing. These results suggest that Japanese undergraduates felt less guilty with regard to not overcoming the situation positively, and leaving things to chance. However, there was no difference between the two studies in evaluating the moral wrongness of each act on a seven-point scale. This might have been due to the difference between a forced-choice method and an independent

continuous evaluation method. For example, unlike a forced-choice method there was a neutral choice on a seven-point scale. Further detailed examination is needed on this issue.

In several items asking participants to anticipate decisions of people in general, there were significant differences between participants who would push and those participants who would not. Most of the participants who would push thought that people in general would also push, while most of the participants who would not push anticipated that people in general, similarly, would not. Moreover, fewer participants who would not push than participants who would push anticipated that people in general would think it wrong not to push. On the other hand, there was no difference in the frequency of participants who anticipated that people in general thought it would be wrong to push. Although it had been expected that fewer participants who would push would anticipate that people in general would think it wrong to push, no such tendency was found. That is, almost 70% of participants anticipated that people in general would think it wrong to push, regardless of their own decision. Participants who would push (Hamiltonians) tended to anticipate that people in general would choose the same decision as them, but they did not expect that the decision would be affirmed morally. However, there was no difference in evaluation using the seven-point scale. As discussed above, the difference might have been caused by the difference between a forced-choice method and an independent continuous evaluation method. In the vignette study 'Kantian' was favored more than 'Hamiltonian'. However, when participants who would push evaluated by themselves, the favorable impressions did not differ. Although participants who would not push (Kantians) favored persons who made the same judgment as they had done, and anticipated that people in general also would favor the 'Kantian' view, participants who would push (Hamiltonians) did not prefer 'Hamiltonian' more than 'Kantian', even when they evaluated themselves.

These results indicate that the participants who refused to push a stranger to save five brothers anticipated that people in general would also refuse to push and would have a good impression of a person who would not push. Moreover, participants who answered that they would push anticipated that people in general would think it wrong to push and would favor the person who would not push. These findings seem to support the discussion of Kurzban et al. (2012) that the computational system that delivers Kantian moral judgment functions to coordinate condemnation decisions. Referring to the dual-process theories in cognitive psychology, Stanovich (2004) proposed that rationality for genes does not necessarily agree with rationality for their vehicle when the vehicles are "long-leashed" species, such as humans. This idea might be useful when we examine the evolution of human morality. Kantian moral judgment might be an example of rationality for the vehicle. Further studies are needed to investigate whether the participants who refused to push a stranger could actually benefit by their decision.

Acknowledgments

I would like to thank the reviewers for their helpful comments.

References

Daly, M., & Wilson, M. (1988). *Homicide*. New York: Aldine de Gruyter.

DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition*, 112, 281-299. (doi: 10.1016/j.cognition.2009.05.008)

Foot, P. (2002). The problem of abortion and the doctrine of double effect. In *Virtues and Vice: And other essays in moral philosophy* (pp.19-33). Oxford: Oxford University Press. (doi:10.1093/0199252866.003.0002)

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65, 613-628. (doi: 10.1037/0022-3514.65.4.613)

Hamilton, W. D. (1964). The genetical evolution of social behaviour I and II. *Journal of Theoretical Biology*, 7, 1-16 & 17-52. (doi:10.1016/0022-5193(64)90038-4&90039-6)

Kurzban, R., DeScioli, P., & Fein, D. (2012). Hamilton vs. Kant: Pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior*, 33, 323-333. (doi:10.1016/j.evolhumbehav.2011.11.002)

Stanovich, K. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press.