Subverting Impartiality:
Remorse as a Basis for Exploiting Moral Wiggle Room

Daniel Fein

2009–2010 Penn Humanities Forum
Undergraduate Mellon Research Fellowship

Abstract

The extent to which third parties punish perpetrators can be influenced by several variables, including the severity of the crime, the way that the crime is committed, and the social categories of the perpetrator and victim. Our study examines the extent to which the effect of social category, or relative impartiality, is modulated by the severity of the crime and the perceived remorse of the perpetrator. Subjects read a series of scenarios where a moral crime was committed by a family member, friend, or stranger, and answered questions related to wrongness, attributed remorse, and punishment. We found no significant relationship between wrongness and impartiality or remorse and impartiality, because we found no significant differences in impartiality between scenarios. This suggests that humans treat in-group members differently in moral situations than in non-moral situations.

Subverting Impartiality:

Remorse as a Basis for Exploiting Moral Wiggle Room

Moralistic impartiality is a puzzling ideal for humans to hold. As DeScioli and Kurzban (2007) describe, for third parties the apparent opposite of impartiality is *loyalty*. The authors describe loyalty, or partiality, as more than a biological default, but as a seemingly complex evolved trait. They describe certain features of our cognitive architecture that appear to be designed specifically for fostering loyalty: "kin recognition devices, [and] a suite of mechanisms that recognize individuals, record interaction histories, tally delivered and received benefits, and execute reciprocal responses." The existence of these systems points to a powerful drive towards partiality, and makes the apparent existence of impartiality in some instances remarkable.

Third Party Punishment

While it seems clear that punishing family and friends carry large costs, third party punishment tends to be costly even when directed towards strangers. Not only does effective punishment require resources to carry out, but also risks alienating the punisher from some segment of the population. Therefore, natural selection is expected to have built systems designed for accurately assessing various types of costs and benefits for each opportunity to punish (Lieberman & Linke, 2007).

Some of the factors that can influence the decision of a third party judge are related to the way the crime was committed. Judges have been found to punish acts more harshly than omissions, even with all other variables held constant (Ritov & Baron, 1990; Haidt & Baron 1996, Bruening 2008). Other punishment-influencing factors are related to the type of crime, with different domains eliciting different levels of punishment, even if the harm to the victim is

removed or held constant (Asao, DeScioli, & Kurzban 2008). The social categories of the victim and perpetrator have also been found to significantly influence third party judgment. Third parties are more likely to seek retribution against a perpetrator who commits a moral crime against a group member than against a non-group member (Bernhard, Firschbacher, & Fehr, 2006), and are more likely to punish a non-group member than a group member (Lieberman & Linke, 2007).

Another dimension that can affect the relative cost of third-party punishment is the level of opposition to the perpetrator. Third-party judges incur fewer costs when punishing a perpetrator to whom there is a greater amount of opposition. In these cases, judges can "jump on the bandwagon," and receive both the support and protection of the public. Conversely, a judge that fails to properly punish a perpetrator who faces substantial opposition may incur costs related to his decision not to punish harshly enough. These observations suggest that there are numerous complex mechanisms that have evolved for the function of assessing wrongness and distributing punishment in ways that are adaptive for the third party judge.

Impartiality

Previous studies of impartiality with regard to moralistic punishment have been limited to single scenarios. Lieberman and Linke (2007) found that when third party judges were presented with equivalent thefts committed by a family member, a schoolmate, or a foreigner, the family member was punished the least and the foreigner was punished the most . Bernhard, Fischbacher, and Fehr (2006) found that in a dictator game, a norm-violating dictator expects less punishment when the third party judge is an in-group member.

While both of these studies have contributed to the body of knowledge around impartiality, they focus more on instances of partiality than on the relative levels of impartiality. For the reasons described above, it is hard to imagine an evolved system that treats all social categories in moral situations exactly the same. Therefore, we would never expect to see impartiality reflected as a total blindness to social category. Instead, we would expect to see impartiality consistently coming into conflict with loyalty, and for relative levels of impartiality to be influenced by many factors in the environment and features of the crime.

This leads to an important question regarding moral impartiality that has yet to be explored empirically: In which situations are third-party judges most likely to act impartially, and in which are judges most likely to be biased? One way to test this is to look at the relationships between punishments of different social categories in multiple situations, and to see if they are consistent.

The current study

Our model predicts that third-party impartiality should fluctuate with the level of opposition to the perpetrator of the crime. If the level of opposition to the perpetrator varies with the severity of the crime committed, then we would expect judges to be more impartial for more severe crimes. On the other hand, since there exists a positive relationship between punishments in general and severity of crimes, third party judges might be more biased towards in-group members as the crimes become worse. If this were the case, we would still predict a positive relationship between punishment and crime severity, but would also predict larger biases between social categories as the overall wrongness and punishment increase.

A second way to explore this question is to hold constant the details of the crime, but to vary the level of remorse the third-party judge perceives on the part of the perpetrator. One can imagine several features of a crime that might influence the level of support for or opposition to the perpetrator, and thus influence the amount of "wiggle room" afforded to the judge in assigning a punishment. Some examples include strength of evidence, the intentions of the perpetrator, the causality of the act committed by the perpetrator, and the remorse felt by the perpetrator after committing the act. All of these variables might contribute to the punishments assigned by third-party judges, and to the level of impartiality they are likely to exhibit.

For this study we chose to focus on remorse because of its intrinsic link to the identity and character of the perpetrator, as opposed to other examples that deal more with the details of the crime itself. In addition, prior studies have included remorse as a dependent variable, and have asked subjects to gauge the remorse of perpetrators of crimes (Lieberman and Linke, 2007).

We would predict that when the perpetrator does not explicitly state a level of remorse, the judge would take the opportunity to exploit the "wiggle room" provided and to punish kin and friends less harshly. Lieberman and Linke (2007) describe a systematic tendency to attribute more remorse to kin and friends than to strangers who commit exactly the same time. We expect these results to replicate across the different crimes that we present to third party judges.

We also predict that when expressions of remorse are made explicit in vignettes and are held constant between kin and strangers, judges will no longer benefit from the "wiggle room" that ambiguous levels of remorse afford them. We predict that in these cases – where either high or low levels of remorse are explicitly stated by both kin and strangers – judges will be more impartial than in cases where there are no explicit expressions of remorse.

Here we report two studies to explore the nature of impartiality in third-party judgments of moral crimes. First, we report a study in which subjects were presented with a series of vignettes depicting crimes of various severities and asked to rate how wrong they were and to assign punishments to the perpetrators. In our second study, we presented subjects with a vignette of a single crime, and varied whether remorse was explicitly stated by the perpetrator and whether the statement indicated high or low levels of remorse. In each study, we varied the social category of the perpetrator between subjects. This design allowed us to examine differences in judgment for crimes committed by perpetrators of different social categories, and to explore which features of a crime seem to influence relative levels of impartiality.

Study 1

Method

Participants.

Participants were recruited from an online community (N = 300). Sixty subjects were excluded from analysis because of participation in more than one condition. The remaining subjects (N = 240) chose to participate in one of three conditions (n = 79 to 83 per condition). Participants were paid over the internet for completing the study.

Procedure.

Participants were asked to read a series of six scenarios featuring either a family member, friend, or stranger committing different types of crimes. Each subject read every scenario in the same order. The categories of family member, friend, and stranger were held constant within subjects, so that each subject only saw scenarios with one type of perpetrator. The scenarios read as follows:

1. *Your family member [your friend, a stranger] is in a national park. The family member [friend, stranger] tosses a piece of paper on the ground and continues walking.*

2. *Your family member [your friend, a stranger] orders $50 of food from an expensive restaurant. The family member [friend, stranger] enjoys the meal, and walks out without paying the bill.*

3. *Your family member [your friend, a stranger] breaks into a home while the homeowners are asleep. The family member [friend, stranger] steals some expensive property including electronics and jewelry that is estimated to be worth $1,500.*

4. *A man is on a ladder, fixing the roof of his house. Your family member [your friend, a stranger] walks by and pushes the ladder down, causing the man to fall and break his leg.*

5. *Your family member [your friend, a stranger] is cut off by another car while sitting in traffic. The two cars pull over and the family member [friend, stranger] gets into an altercation with the other driver. The family member [friend, stranger] becomes extremely angry and strikes the other driver. The other driver dies from his injury.*

6. *Your family member [your friend, a stranger] finds out that a business partner has cheated him out of a significant amount of money. The family member [friend, stranger] devises a plan to kill the business partner. When the two are alone, the family member [friend, stranger] kills the business partner.*

After each of the scenarios, subjects were prompted to answer a series of questions about the perpetrator and the act. These included questions about the moral wrongness of the act (on a scale of 0 to 100, with 0 being not wrong at all and 100 being the absolute most wrong), the perceived remorse felt by the perpetrator after committing the act (on a scale of 0 to 100, with 0 being no remorse at all and 100 being the absolute most remorse), and degrees of punishment.

The degrees of punishment were the independent variables, with judgments of moral wrongness being the dependent variables.

Participants were asked to assign an appropriate punishment to the perpetrator in terms of both fines and jail time, independently from each other. The "fine" question was phrased: "Let's say [your family member's, your friend's, the stranger's] punishment is to pay a fine. What fine should they have to pay?" Participants were given choices of fines to assign between $0 and $100,000 or more ($0, $50, $100, $500, $1000, $1500, $2500, $5000, $10000, $25000, $50,000, $100,000 or more). The "jail time" question was phrased: "Let's say [your family member's, your friend's, the stranger's] punishment is to serve some time in jail. How much jail time should they have to serve?" Participants could assign any period of jail time between 0 years and 50 years, in increments of months, and a visual guide was provided with examples of government recommended jail times for various crimes.

Results

As predicted, we found highly significant differences in judgments of moral wrongness between the different scenarios for every treatment: family member ($F_{5,461} = 37.11$, $P < .001$), friend ($F_{5,468} = 45.44$, $P < .001$), and stranger ($F_{5,492} = 13.80$, $P < .001$).

However, between treatments and within individual scenarios the differences in perceived wrongness were not always significant. Differences were not significant for scenario 1 ($F_{2,237} = .15$, $P = .864$) and scenario 5 ($F_{2,237} = .21$, $P = .807$), were significant for scenario 2 ($F_{2,237} = 4.13$, $P = .017$), scenario 3 ($F_{2,237} = 3.29$, $P = .039$), and scenario 6 ($F_{2,237} = 3.90$, $P = .022$), and were highly significant for scenario 4 ($F_{2,237} = 7.86$, $P < .001$).

Surprisingly, we did not find any meaningful relationship between perceived moral wrongness and either type of punishment. In fact, even between treatments within any individual

scenario, a single factor ANOVA showed that neither punishment type exhibited any meaningful relationships (fines: *F*s < 2.82, *Ps* > .062; jail time: *F*'s < 1.94, *P*'s > .145).

We found strong results indicating that perceived remorse of the perpetrator is affected by social category. Subjects tended to attribute the most remorse to family members, less to friends, and the least to strangers, within each of the six scenarios (*F*'s > 5.72, *P*'s < .004).

Study 2

Method

Participants.

Participants were recruited from an online community ($N = 400$). Thirty-eight subjects were excluded from analysis because of participation in more than one condition. The remaining subjects ($N = 362$) chose to participate in one of six conditions (n = 44 to 89 per condition). Participants were paid over the internet for completing the study.

Procedure.

Participants were asked to read a vignette in which either a family member or stranger steals $1,500 from a restaurant. In each scenario, the family member or stranger was observed outside the restaurant after committing the crime. Subjects witnessed either a phone conversation where the perpetrator expressed a high level of remorse, a phone conversation where the perpetrator expressed a low level or remorse, or no phone conversation and no expression of remorse. The "high remorse" scenario reads as follows:

One evening you go to dinner at an expensive local restaurant with a group of people that you do not know well [a group of people, one of whom is your brother]. A person at your table [your brother] watches as a large party of about 20 people leaves cash on their table for the check and then exits the restaurant. Before the server goes to the table to collect the money,

you see the person [your brother] walk past the table, take the $1500 left for the bill, and leave through the front door. The person [your brother] is now $1500 richer. As you leave the restaurant, you notice that the person [your brother] appears distraught, and you overhear them [him] talking on the phone. They say [he says], "I've done something very wrong, and I feel horrible about it." Just then, the restaurant manager comes out and looks right at the person [your brother], and says, "I saw you grab the money. You can't steal from me and get away with it."

The "low remorse" scenario reads as follows:

One evening you go to dinner at an expensive local restaurant with a group of people that you do not know well [a group of people, one of whom is your brother]. A person at your table [your brother] watches as a large party of about 20 people leaves cash on their table for the check and then exits the restaurant. Before the server goes to the table to collect the money, you see the person [your brother] walk past the table, take the $1500 left for the bill, and leave through the front door. The person [your brother] is now $1500 richer. As you leave the restaurant, you notice that the person [your brother] appears distraught, and you overhear them [him] talking on the phone. They say [he says], "I've just acquired a bunch of money, and I feel great about it.." Just then, the restaurant manager comes out and looks right at the person [your brother], and says, "I saw you grab the money. You can't steal from me and get away with it."

After reading the scenario, subjects were again asked to rate the wrongness of the crime, to assign punishments to the perpetrator both in terms of jail times and fines, and to rate the level of the remorse they thought the perpetrator felt after committing the crime. Subjects were given the same guidance and answer choices as in Study 1.

For vignettes where the remorse of the perpetrator was stated explicitly, we asked subjects to rate their perception of the perpetrators' remorse as a manipulation check. We expected subjects to find perpetrators in the "low remorse" condition less remorseful than those in the "high remorse" condition, and for the control group to fall somewhere in between.

Results

Our primary prediction – that there would be an interaction effect between the level of remorse expressed and the social category of the perpetrator – was not supported by the data. A univariate ANOVA showed that the interaction was insignificant for measures of perceived wrongness ($F_{2, 358} = .235$, $P = .791$), jail time assessed ($F_{2, 360} = .881$, $P = .415$), or fines assessed ($F_{2, 360} = .839$, $P = .433$).

Oneway ANOVAs also did not show any significant main effects between social category and wrongness ($F_{1, 358} = .013$, $P = .909$), jail time assessed ($F_{1, 360} = 1.792$, $P = .182$), or fines assessed ($F_{1, 359} = 2.183$, $P = .140$), or between expressed remorse and jail time assessed ($F_{2, 360} = 2.191$, $P = .113$) or fines assessed ($F_{2, 359} = 2.614$, $P = .075$). However, we did find a significant main effect between expressed remorse and wrongness ($F_{2, 358} = 3.462$, $P = .032$).

Discussion

Our results indicate that relative levels of impartiality may not be mediated by the type of crime that the third party is judging, but seem to be influenced by perceptions of the remorse of the perpetrator. In Study 1, we found no relationship between the perceived wrongness of the crime and the level of impartiality reflected in the punishments recommended for family members, friends, and strangers. However, we found significant differences in the attributions of remorse between family members, friends, and strangers in each condition. In Study 2, we found that when remorse is explicitly stated by the perpetrator – both high remorse and low remorse –

subjects' average punishments of kin and strangers were less divergent than in the control group, where no expression of remorse was stated by the perpetrator. However, these results were not statistically significant, due to large amounts of noise in the data.

The result of Study 1 is inconsistent with previous studies testing the effect that social category has on moralistic punishment. There are several reasons that this could be the case. One possibility is that previous studies did not accurately assess levels of moralistic impartiality. It could be that when presented with multiple situations, humans become more impartial than if presented with one. It could also be that that the choices given to participants for levels of punishment were too broad. Since our scenarios included such a wide range of crimes and we did not want to bias the wrongness judgments of the subjects, we provided a huge range of potential punishments. We found that some subjects remained at either floor or ceiling for all of the situations, and this greatly affected our results. Lieberman and Linke (2007) found the same problem, even within their one scenario and smaller range of punishments. For their analysis of jail time recommendations, they used only a subset of responses in order to mitigate the effect. In the future, it might make more sense to restrict the punishment choices to smaller ranges and to vary the ranges from situation to situation. While this would bias the measure for the perceived moral wrongness of the scenarios, it may be the only way to obtain accurate punishment data.

A significant finding from Study 1 was that subjects within treatments were able to agree on the relative levels of wrongness between situations. Lieberman and Linke (2007) experienced ceiling effects for this measurement, even within their one, relatively mild, crime. The fact that we found meaningful variation indicates that our scenario design was successful – subjects recognized that some scenarios were more morally wrong than others. This was crucial to Study 1, since moral wrongness was our dependent variable. However, we did find that a large number

of subjects assigned the same level of moral wrongness to each scenario. These participants tended to judge each crime as extremely morally wrong; some even rated all six scenarios as "100," or the "absolute most wrong." The fact that some participants made no distinction between the wrongness of littering and premeditated murder reflects a critical aspect of morality in general. With moral judgments, as opposed to other judgments, it seems that humans use a different system for assessing wrongness and punishment that does not always seem entirely logical. For some of our participants, if a moral crime is committed it is automatically considered the absolute most wrong type of crime; there seems to be little nuance between levels of severity. Even within our overall finding that perceived moral wrongness increased with the scenario number, the difference between each situation was surprisingly small.

The finding that our participants recognized and agreed on differences in the wrongness of different scenarios in Study 1 makes the lack of any relationships within our punishment data even more interesting. The levels of punishment between treatments exhibit no meaningful relationships, and at times appear to be relatively haphazard. This could have been a result of subjects not understanding the scenarios the way that we intended them to, but the wrongness results suggest otherwise. It seems that the participants did recognize differences in the severity of the different scenarios, but that these did not translate into any changes in the way that they punished different social categories. In addition, participants' attributions of remorse were significantly "partial." Within each scenario, family members were judged to feel significantly more remorse than friends, and friends significantly more than strangers.

The fact that this result was so robustly replicated while others did not seems to indicate that it is a more consistent feature of our cognitive architecture. Our results suggest that humans systematically perceive more remorse on the part of kin and friends than of strangers, even for

committing the same crime. We hypothesize that this consistently differential attribution of remorse might represent a way for judges to avoid punishing in-group members while simultaneously maintaining their impartiality in the eyes of the public.

In Study 2, we built on this finding by attempting to hold these perceptions of remorse constant while varying the social category of the perpetrator of a crime. We predicted that if subjects were unable to plausibly perceive differences in the remorse of perpetrators of different categories, they also would be unable to justify punishing kin less harshly than strangers.

One unsurprising result from Study 2 was that our manipulations of remorse were successful: we observed that subjects consistently rated both kin and strangers to be more remorseful in the high remorse condition than in the control, and more remorseful in the control than in the low remorse condition.

Our results, while in the direction we predicted, were not statistically significant for any of our dependent measures, except for the main effect of expressed remorse on perceived wrongness. While there were greater differences between the mean punishments assessed for kin and strangers in the control condition than in either of the remorse conditions, we found no interaction between our two independent measures (social category and expressed remorse). As with prior studies on moralistic punishment that we have conducted, the data we gathered contain much statistical noise. There seem to be large individual differences in both level of punishment and assessment of wrongness that make relationships between variables insignificant.

However, our data do show general trends that will be interesting to examine further in experiments designed to generate less noise. For example, our data suggest that explicitly stating low levels of remorse may be less harmful to perpetrators than stating high levels of remorse is

helpful. This seems to support the idea that people possess an underlying assumption that non-kin perpetrators of crimes are unremorseful.

Most generally, these studies confirm that the function and characteristics of moralistic punishment are complex and difficult to examine experimentally. We look forward to conducting additional experiments designed to decrease the noise in our data. What is clear is that subjects consistently make decisions about punishment based on a large variety of factors. Teasing apart the effects that different factors have on decisions to punish will be an important step towards understanding moralistic punishment as a whole.

References

Asao, K., DeScioli, P., and Kurzban, R. (2008). *The puzzle of morality: An investigation of moral cognition across domains*. [PowerPoint slides].

Baron, J., Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*, 74-85.

Bernhard, H., Fischbacher, U. and Fehr, E. (2006). Parochial altruism in humans. *Nature, 442*, 912-915

Bruening, R.A. (2008). Effect of Evidence of Wrongdoing on Third Party Moral Cognition. Unpublished Manuscript.

Dana, J., Weber, R., and Kuang, J. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*, 67-80.

DeScioli, P., and Kurzban, R. (2007). Mysteries of morality. Unpublished Manuscript.

Haidt, J., and Baron, J. (1996). Social roles and the moral judgment of acts and omissions. *European Journal of Social Psychology, 26*, 201-218

Hamilton, W.D. (1963) The evolution of altruistic behavior. *American Naturalist, 97*, 354-356

Kurzban, R., and Leary, M.R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. *Psychological Bulletin, 127*, 187-208.

Lieberman, D., and Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology, 5*, 289-305

O'Neill, P., and Petrinovich, L. (1998) A preliminary cross-cultural study of moral intuitions. *Evolution and Human Behavior*, *19*, 349-367.

Petrinovich, L., O'Neill, P., and Jorgensen, M. An empirical study of moral intuitions: toward an evolutionary ethics. *Journal of Personality and Social Psychology, 64*, 467-478

Ritov, I., Baron J. (1990). Reluctance to vaccinate: Omission bias and ambiguity.

Journal of Behavioral Decision Making, 3, 263-277