A reanalysis of the personal/impersonal distinction in moral psychology research

Jonathan McGuire\textsuperscript{a,}\textsuperscript{*}, Robyn Langdon\textsuperscript{a}, Max Coltheart\textsuperscript{a}, Catriona Mackenzie\textsuperscript{b}

\textsuperscript{a} Macquarie Centre for Cognitive Science, Macquarie University, Sydney, NSW 2109, Australia
\textsuperscript{b} Department of Philosophy, Macquarie University, Sydney, NSW 2109, Australia

Abstract

Greene and colleagues [Greene, J., Sommerville, B. R., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293, 2105–2108.] have revealed an apparent distinction in folk psychology between ‘up close and personal’ and ‘impersonal’ moral dilemmas. Reasoning about these types of dilemmas is purportedly supported by partially dissociable neural systems. However, further investigation of the data supporting this hypothesis indicated that only a small number of stimuli used by Greene et al. are driving the effect originally found. Implications of the apparent distinction initially reported and of other research in the domain of moral psychology are discussed.

The work of Greene and colleagues in moral psychology (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), particularly their proposed distinction between ‘personal’ and ‘impersonal’ moral actions, has been influential in the realm of moral psychology (e.g. Killgore, Killgore, Day, Li, Kamimori, & Balkin, 2007; Koenigs et al., 2007; Valdesolo & DeSteno, 2006), and has also been held to be relevant in fields as varied as philosophy (Greene, 2003; Levy, 2006; Singer, 2005), medical ethics (Tassy, Le Coz, & Wicker, 2008), advertising (Plassmann, Ambler, Braeutigam, & Kenning, 2007), and the psychology of genocide (Midlarsky, 2005).

‘Personal’ moral actions are emotionally charged and are defined as those that “could reasonably be expected to lead to serious bodily harm... to a particular person or a member or members of a particular group of people... where this harm is not the result of deflecting an existing threat onto a different party” (Greene et al., 2001, p. 2107). Moral actions that do not meet the full set of criteria are considered ‘impersonal’. An example of a personal moral dilemma used by Greene et al. (2001) is the ‘footbridge case’, a variant of philosophical ‘trolley problems’ (Foot, 1967; Thomson, 1976), in which one must decide if it is appropriate to push a stranger to his death in front of a runaway trolley in order to save the lives of five workmen. In this example, a personal moral violation (pushing the man to his death) is required in order to maximise utilitarian outcomes (saving the five workmen). While Greene et al. (2001) do acknowledge these criteria to be preliminary, to our knowledge there are no data demonstrating the necessity of all three criteria in eliciting stronger emotional responses, nor do they correspond to well-established philosophical distinctions.

Greene et al. (2001) propose a dual-process model of moral judgement involving a fast, unconscious and effortless ‘affective’ system and a slow, conscious, and effortful ‘cognitive’ system. These systems are purported to be subserved by partially dissociable neural systems (Greene, 2003; Greene & Haidt, 2002; Greene et al., 2004). The affective (‘hot’) system is presumed by Greene and colleagues to favour rules and to be preferentially activated by ‘personal’ moral considerations, while the cognitive (‘cold’) system favours consequences. Cognitive control mechanisms are hypothesised to be recruited in order to resolve rule/consequence conflicts (Greene & Haidt, 2002; Greene et al., 2001).

In support of this model, Greene et al. (2001) found that participants took longer to accept as appropriate personal actions that would maximise utilitarian outcomes than to reject such actions as inappropriate. This pattern was not found for impersonal actions that maximised utilitarian outcomes (see Fig. 1). Greene et al. (2001) argued that contemplating personal dilemmas which pit rules against outcomes recruits both the hot system, which favours rejecting the actions, and the cold system, which favours endorsing them, and the increased reaction time shown was indicative of the cognitive control required to overcome the initial hot response.

Lending further support to their dual-process model, Greene et al. (2001) found a broad pattern of increased activation in brain areas involved in processing emotional information and decreased activation of areas associated with working memory during consideration of personal dilemmas. They also found that the...
dorsolateral prefrontal cortex and anterior cingulate cortex (ACC), involved in conflict resolution, were recruited when participants endorsed personal moral violations.

The personal/impersonal distinction has however been criticised for being overly crude and unable to explain the variability found in responses to the trolley problems (Mikhail, 2007). Greene et al. (2001, 2004) have also been criticised for using more emotive language and references to family members or friends in their personal dilemmas, for failing to control for cognitive processing requirements across conditions, and for ambiguously asking whether actions were ‘appropriate’ (Schaich Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006).

Furthermore, scrutiny of the stimuli used by Greene et al. (2001) reveal a number of potential problems. It is not clear that all stimuli used are truly ‘dilemmas’, particularly the non-moral dilemmas. Hence we undertook two reanalyses of Greene et al.’s behavioural data, as provided to us personally by Joshua Greene. Firstly, we repeated Greene et al.’s subject analysis. We then performed an item analysis, in which items rather than subjects are treated as cases for analysis. Whenever two sets of N1 and N2 items are administered to a set of M subjects and a statistical analysis by subjects yields a significant difference, it is statistically possible that this significant difference arises from responses to just one of the items in the first item set and just one of the items in the second set, in which case the conclusion being drawn (that subjects are in general behaving differently to items of the first type compared to items of the second type) is unwarranted. This is a problem first recognised (re psycholinguistic research) by Clark (1973). His solution was that the data must be analysed across items as well as across subjects; if the significant subject effect is due to just one or two items, the effect will not be significant in the item analysis. To carry out item analyses as well as subject analyses has thus been standard practice in psycholinguistics since then (Forster, 2008). Exactly the same issue arises with questionnaire data such as those of Green et al. (2001). So we performed an item analysis of those data, which allowed us to determine whether any effects that are statistically significant in the subject analysis are (as one would hope) resulting from some general properties of the items, rather than arising from just one or two aberrant items.

Analysis 1

Method

Analysis 1 recreated Greene et al. (2001) subject analysis, except that non-moral items were excluded.

Non-moral dilemmas were excluded from all analyses for several reasons. At least ten of these could not reasonably be described as ‘dilemmas’, as they elicited very high levels of agreement (between 95% and 100%) across subjects. An example of such a ‘non-dilemma’ is the appropriateness of harvesting 20 bushels of turnips rather than 10. Of the remaining non-moral dilemmas, none involves a social component, and all can be better identified as pragmatic decisions. For these reasons we argue that these dilemmas do not provide a useful comparison to the moral (personal and impersonal) dilemmas, and so have analysed only the data for the personal and impersonal moral dilemmas.

The dependent variable for each subject was the mean reaction time (RT) across items in each condition. IVs were dilemma type (personal vs. impersonal) and response (appropriate vs. inappropriate action). The model also assessed the interaction of dilemma type by response.

Results and discussion

Response ($F(1121.869) = 10.597, \ p = 0.001$), dilemma type ($F(1121.171) = 14.467, \ p < 0.001$) and interaction ($F(1121.171) = 31.605, \ p < 0.001$) were all significant, as found by Greene et al. (2001). The same results were obtained when the log of RT was used as DV to control for the effect of RT skew. The interaction occurred due to higher mean RTs for responses of ‘appropriate’ to personal dilemmas. See Table 1 for the means of all analyses.

When non-moral items were excluded from analysis, the interaction between dilemma type and response first reported by Greene et al. (2001) remained significant, so the results of analysis 1 supported the purported distinction between personal and impersonal moral dilemmas.

Table 1

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Response</th>
<th>Dilemma type</th>
<th>Mean</th>
<th>Std. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis 1</td>
<td>Not appropriate</td>
<td>Impersonal</td>
<td>4888.8</td>
<td>256.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Personal</td>
<td>4580.6</td>
<td>256.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Impersonal</td>
<td>4490.0</td>
<td>254.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Personal</td>
<td>6087.9</td>
<td>254.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Impersonal</td>
<td>6088.3</td>
<td>502.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Personal</td>
<td>5970.4</td>
<td>571.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Impersonal</td>
<td>5731.8</td>
<td>502.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Personal</td>
<td>5436.0</td>
<td>571.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Impersonal</td>
<td>6008.3</td>
<td>502.2</td>
</tr>
<tr>
<td>Analysis 2</td>
<td>Not appropriate</td>
<td>Impersonal</td>
<td>4826.6</td>
<td>271.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Personal</td>
<td>6245.4</td>
<td>295.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Impersonal</td>
<td>4508.3</td>
<td>259.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Personal</td>
<td>5984.9</td>
<td>261.3</td>
</tr>
<tr>
<td>Analysis 3</td>
<td>Not appropriate</td>
<td>Impersonal</td>
<td>6126.6</td>
<td>605.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Personal</td>
<td>6809.9</td>
<td>605.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Impersonal</td>
<td>5369.7</td>
<td>586.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Personal</td>
<td>6763.2</td>
<td>605.9</td>
</tr>
</tbody>
</table>
Analysis 2

Method

Item analysis was conducted next. Here the dependent variable was computed for each individual dilemma and was the mean RT across subjects in each of the 2 (dilemma type) × 2 (response) conditions. This analysis was performed to assess the generalisability of the above results to different populations of personal and impersonal dilemmas.

Results and discussion

No effects were significant: Response (F(1,74) = .057, p = 0.811); dilemma type (F(1,74) = .096, p = 0.757); and interaction (F(1,74) = .568, p = 0.453). The same pattern of results was obtained using the log of RT as DV.

This implies that the effects found in the subject analysis are not generalisable to other populations of moral dilemmas, instead being driven by just a few particular dilemmas within this set of stimuli.

Analysis 3

Method

Inspection of the data made it clear that items with a very low percentage of subjects judging the action to be appropriate had strikingly fast RTs, indicating subjects rejected the proposed action very quickly. We removed these items from the subject analysis to assess whether the results previously found in subject analysis remained significant, or were an artefact of this unusual subgroup of dilemmas.

Items with less than 5% of the subjects judging the actions to be ‘appropriate’ were considered ‘poorly endorsed’ items, and were removed from analysis. Nine of the 40 dilemmas were considered ‘poorly endorsed’ (one impersonal and eight personal). Subject analysis was then performed with the same factors as in our first analysis.

Results and discussion

Upon subject analysis with poorly endorsed items removed, only dilemma type was revealed to be significant: response (F(1,74) = 2.640, p = 0.104); dilemma type (F(1,74) = 68.855, p < 0.001); interaction (F(1,74) = 0.027, p = 0.870) (see Fig. 2). Importantly, the interaction was no longer significant. The same pattern of results was obtained using the log of RT as DV. The only significant result was that the personal dilemmas elicited longer RTs overall, and that the underlying mechanism responsible for the dilemma type by response interaction reported by Greene et al. (2001) was due entirely to these poorly endorsed personal dilemmas (or ‘personal non-dilemmas’). Thus the apparent interaction in the subject analysis of Greene et al. (2001) was due entirely to these poorly endorsed personal dilemmas.

General discussion

The purpose of our reanalyses of this data was to assess the generalisability of the previous findings. Item analysis suggested that the results found by Greene et al. (2001) were driven by idiosyncratic item characteristics, and were not generalisable to other putative populations of moral dilemmas. Furthermore, when poorly endorsed items were excluded from analysis, the important dilemma type by response interaction became non-significant, indicating that this interaction was driven solely by extremely fast RTs to the poorly endorsed items, as shown by comparison of Figs. 1 and 2.

Once these poorly endorsed items were removed, it became clear that ‘personal’ dilemmas elicit longer RTs overall, and that the apparent interaction in the subject analysis of Greene et al. (2001) was due entirely to these poorly endorsed personal dilemmas (or ‘personal non-dilemmas’). Thus the underlying mechanism responsible for the dilemma type by response interaction reported by Greene et al. (2001) is the opposite of that which Greene et al suggest. Rather than longer RTs for responses of ‘appropriate’ to ‘personal’ moral dilemmas, it is in fact the extremely fast responses of ‘inappropriate’ for a small subset of dilemmas that generate this interaction.

Finally, the results of both item analyses were non-significant, indicating that even with removal of poorly endorsed items, these results are not generalisable to the purported wider populations of ‘personal’ and ‘impersonal’ moral dilemmas. We hope that the current analysis serves to put the hypothesised ‘personal/impersonal’ distinction to rest.

These results also suggest that Greene et al.‘s imaging results should be revisited. With no other reason to support a dual-process interpretation of the data, the most parsimonious interpretation may be that the patterns of activation may be reflecting the hedonic values of the stimuli. We also note the difficulty in interpreting
imaging data, particularly when assumptions of modularity cannot be supported (Van Orden & Paap, 1996), or when regions of interest are poorly selective in function (Poldrack, 2006). Both of concerns apply to the regions of interest here; for example, as noted by Greene et al. (2001), the function of the ACC is not certain.

Our findings have a number of implications for research investigating a dual-process model of moral judgement. Most obviously, they indicate that there is no reason to assume that emotionally salient moral decisions are processed in a qualitatively different way to those dilemmas that are not emotionally salient. Furthermore, there is no evidence here to support the theory that there are two competing moral systems at work.

Our concerns regarding stimulus and research design are also significant in regards to other studies reporting results that purportedly support a dual-process model of moral judgement. Mendez, Anderson, and Shapira (2005) claimed to have shown that patients with frontotemporal dementia and is less influenced by the emotional aspects of moral dilemmas than healthy controls. In a similar vein, Ciaramelli, Muccioli, Ladavas, and de Pellegrino (2007) and Koenigs et al. (2007) claim to show that patients with ventromedial prefrontal lesions are more likely to endorse ‘personal’ moral violations than healthy controls. Finally, Valdesolo and DeSteno (2006) used emotion induction to investigate the influence of emotion on moral decision making.

However, Mendez et al. (2005) and Valdesolo and DeSteno (2006) only utilised the trolley and footbridge dilemmas, which vary not only in emotional salience, but also in regards to the level of physical contact involved. Furthermore, these dilemmas differ on whether harm is intended as a means to an end or is merely a foreseen side effect of that end. These considerations have been shown to influence moral judgments (Cushman, Young, & Hauser, 2006). Ciaramelli et al. (2007) used stimuli selected randomly from the battery used by Greene et al. (2001), and as such may be subject to the same criticisms as outlined in the current paper, although Koenigs et al. (2007) maintained their results upon removal of low-conflict scenarios.

Interpretation of these sorts of results is also difficult. As has been noted (Moll & de Oliveira-Souza, 2007), findings such as those in frontotemporal dementia and VMPFC patients do not show conflict between ‘hot’ and ‘cold’ processes and so cannot be used to distinguish between a dual-process theory and one in which emotions act as an input to the reasoning process or as a guide to the salience of situational information.

The results obtained by Greene et al. (2001) have been shown to be artefacts of stimulus design. That is, Greene et al. (2001) did not use matched dilemmas in their different conditions, rather relying on completely different scenarios between conditions. For example, the footbridge and trolley cases, which are the most similar between conditions, differ on whether the action involves physical contact and harm as a means to an end, both of which have been indicated as factors in moral judgement (Cushman et al., 2006). We suggest that future studies in this field use stimuli differing only the variable of interest. Item analyses should also be performed, particularly when novel stimuli are being used, to ensure that results are generalisable across the supposed populations of moral dilemmas under investigation. More research needs to be done at a behavioural level in order to fine-tune the questions being asked before work identifying the neural correlates of moral decision making can be useful.

Acknowledgments

We thank Dr. Alan Taylor for invaluable assistance with data analysis and Dr. Joshua Greene for providing us with the data used for this paper.

References


