The influence of prior record on moral judgment

Dorit Kliemann a, b, *, Liane Young b, Jonathan Scholz b, Rebecca Saxe b

a Department of Neuropsychology and Behavioral Neurobiology, Center for Cognitive Sciences, Hochschulring 18, Cognium Building, D-28359 Bremen, Germany
b Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract

Repeat offenders are commonly given more severe sentences than first-time offenders for the same violations. This practice escalates penalties with offense history is widespread in both criminal and civil law, in many countries and over many centuries (Durham, 1987). Moreover, the practice fits with common sense: intuitively, it seems “right” that persistent offenders receive more severe punishments. Nevertheless, both justice and economic models of the law advocate against escalating penalties. According to the justice model, punishment is justified only if the amount of punishment is proportional to the harm caused by the violation. Escalating penalties violate this rule, punishing repeat offenses disproportionately (Ashworth, 2005; Durham, 1987). According to the economic model, an optimal punishment regime is one in which the expected punishment for a violation equals the social cost of the violation. The model was correlated with individual subjects’ behavioural responses to unfair play in the game. These results thus provide insight for both legal theory and moral psychology.

1. Introduction

Repeat offenders commonly receive more severe sentences than first-time offenders for the same violations. This principle of escalating penalties with offense history is widespread in both criminal and civil law, in many countries and over many centuries (Durham, 1987). Moreover, the practice fits with common sense: intuitively, it seems “right” that persistent offenders receive more severe punishments. Nevertheless, both justice and economic models of the law advocate against escalating penalties. According to the justice model, punishment is justified only if the amount of punishment is proportional to the harm caused by the violation. Escalating penalties violate this rule, punishing repeat offenses disproportionately (Ashworth, 2005; Durham, 1987). According to the economic model, an optimal punishment regime is one in which the expected punishment for a violation equals the social cost of the violation. The model was correlated with individual subjects’ behavioural responses to unfair play in the game. These results thus provide insight for both legal theory and moral psychology.

Repeat offenders are commonly given more severe sentences than first-time offenders for the same violations. This practice escalates penalties with offense history is widespread in both criminal and civil law, in many countries and over many centuries (Durham, 1987). Moreover, the practice fits with common sense: intuitively, it seems “right” that persistent offenders receive more severe punishments. Nevertheless, both justice and economic models of the law advocate against escalating penalties. According to the justice model, punishment is justified only if the amount of punishment is proportional to the harm caused by the violation. Escalating penalties violate this rule, punishing repeat offenses disproportionately (Ashworth, 2005; Durham, 1987). According to the economic model, an optimal punishment regime is one in which the expected punishment for a violation equals the social cost of the violation. The model was correlated with individual subjects’ behavioural responses to unfair play in the game. These results thus provide insight for both legal theory and moral psychology.

Repeat offenders are commonly given more severe sentences than first-time offenders for the same violations. This practice escalates penalties with offense history is widespread in both criminal and civil law, in many countries and over many centuries (Durham, 1987). Moreover, the practice fits with common sense: intuitively, it seems “right” that persistent offenders receive more severe punishments. Nevertheless, both justice and economic models of the law advocate against escalating penalties. According to the justice model, punishment is justified only if the amount of punishment is proportional to the harm caused by the violation. Escalating penalties violate this rule, punishing repeat offenses disproportionately (Ashworth, 2005; Durham, 1987). According to the economic model, an optimal punishment regime is one in which the expected punishment for a violation equals the social cost of the violation. The model was correlated with individual subjects’ behavioural responses to unfair play in the game. These results thus provide insight for both legal theory and moral psychology.

In spite of these considerations, legal practice in the US over the past 30 years has tended towards increasing, rather than decreasing, reliance on prior record during sentencing, as in the “Three-Strikes” policy in California (Austin, Clark, Hardyman, & Henry, 2000). Many efforts have been made to account for this phenomenon (Ashworth, 2005; Dana, 2001). One theory, for example, treats escalating penalties as deterrent or preventative incapacitation: if the offender is incarcerated, he or she will be less able to commit another offense (Ashworth, 2005).

An alternative is that escalating penalties express society’s moral condemnation of persistent wrongful action (Dana, 2001; Sunstein, 2005), regardless of utilitarian calculations. The current study investigates this alternative: do laypersons indeed judge first-time offenders as less blameworthy, and repeat offenders as more blameworthy, for the same harm caused? How are the effects of prior record related to other aspects of folk morality, such as attribution of intent to moral agents (Cushman, personal communication; Pizarro, Laney, Morris, & Loftus, 2006; Woolfolk, Doris, & Darley, 2006)? Specifically, does negative prior record lead subjects to attribute more intentionality to agents for causing negative outcomes; if so, is this effect a cause or a consequence of a change in moral judgment, i.e. increase in blame?

Consider the following example scenario: Ashley works at the computer help desk and often friends bring their computers. Once,
Chris found someone else’s clothes lying wet in the washing machine in the basement of his building. He put all of the clothes into the dryer and turned it on the regular cycle, shrinking his neighbor’s new sweater four sizes.

In each scenario, one of the ten competitors from the Game performed an action that either led to a positive or negative outcome. Scenarios did not explicitly state the agent’s intentions or the action’s moral status. Corresponding to each story, a target sentence was presented to ask for subject’s rating of the intentional (Behavioural Experiment) or moral status (fMRI Experiment) of the action.

<table>
<thead>
<tr>
<th>Negative outcome story</th>
<th>Target sentence</th>
<th>Positive outcome story</th>
<th>Target sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jessica once went on a camping trip with her ex-boyfriend. On the second day, there was a thunderstorm, and a big branch fell onto the tent, hitting him on the ankle. She wrapped an ace bandage tightly around the swelling ankle, which made the swelling get worse, and the pain even more intense</td>
<td>Jessica made her ex-boyfriend’s swelling ankle worse when she wrapped it</td>
<td>Jessica once went on a camping trip with her ex-boyfriend. On the second day, there was a thunderstorm, and a big branch fell onto the tent, hitting him on the ankle. She wrapped the swelling ankle in a sheet that was soaked from the cold rainwater. The cold water numbed the pain and helped him recover</td>
<td>Jessica wrapped her ex-boyfriend’s swelling ankle and it helped him to recover</td>
</tr>
<tr>
<td>Chris shrank his neighbor’s new sweater</td>
<td></td>
<td>Chris was doing his laundry very late at night, in the basement of his building. Mixed in with his own dry clothes were someone else’s clothes. He kept folding until all the clothes were done: his own, and the stranger’s</td>
<td></td>
</tr>
<tr>
<td>Chris folded some of the stranger’s dry clothes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1

Story Task vignettes

<table>
<thead>
<tr>
<th>Negative outcome story</th>
<th>Target sentence</th>
<th>Positive outcome story</th>
<th>Target sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Her ex-boyfriend brought his computer, which had crashed. Ashley restarted the computer, the hard-drive was re-formatted and all of Chris’ files were lost. For actions resulting in a negative outcome (e.g., lost files on the computer) caused by an agent with a negative prior record (e.g., a negative prior personal experience with Ashley), we hypothesize that participants judge the agent as (1) more blameworthy and (2) having acted more intentionally, compared to agents with no prior record. If so, we further ask whether the increase in blame precedes or follows the increased attribution of intentionality. The current study investigated these questions, using behavioural and neuroimaging (functional magnetic resonance imaging, fMRI) methods.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In each scenario, one of the ten competitors from the Game performed an action that either lead to a positive or negative outcome. Scenarios did not explicitly state the agent’s intentions or the action’s moral status. Corresponding to each story, a target sentence was presented to ask for subject’s rating of the intentional (Behavioural Experiment) or moral status (fMRI Experiment) of the action.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**2. Methods**

Subjects (fMRI Experiment: nine male, seventeen female, aged 19–33 years; Behavioural Experiment: three male, four female, aged 18–48 years) were naive to experimental hypotheses, right-handed and recruited by email at the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology (MIT). All subjects had normal or corrected-to-normal vision, were native English speakers, participated for payment and gave written informed consent in accordance with the requirements of MIT’s Committee on the Use of Humans as Experimental Subjects. Each subject participated two sessions: the Game, and the Story Task.

**2.1. Game**

Subjects were told that they were recruited in groups of 10–12 people, who would not meet face to face, but would play the Game against each other over a computer network. When they arrived, subjects were met by an experimenter and taken to a hallway containing 12 experimental rooms, each labeled with the experiment name, and a subject number. Subjects were taken into one room containing a single computer and a sheet of paper, their photograph was taken, and they were given written instructions for the Game (described below). Subjects were informed that in the second experimental session, members of the group would read stories about one another. They were asked to provide two or three short stories that would be rewritten by the experimenters and later presented to the players. They were provided hints of possible story types (e.g., something nice you did for a stranger, something that turned out worse than expected).

After a few minutes, the experimenter returned, collected the stories, and showed the subject a page containing photographs of the 10 “other players”. Subjects were asked to mark on the page whether they knew any of the people in the photographs. The photographs were taken from the FRI CVL database of face images (http://www.lrv.fri.uni-lj.si/facedb.html). Solina, Peer, Batagelj, Jovan, & Kovac, 2003) and showed six male and female white college-aged faces. All subjects marked that they did not know any of the players. Then subjects were instructed to wait for a cue on the screen, that everyone else was ready, and that the Game was about to begin. The experimenter then left the room, and within 2 min triggered the Game remotely.

Subjects then played 100 trials of repeated sequential economic investment game. For each trial, an Investor and a Trustee were chosen: the subject was ran-
domly assigned to be either the Investor or the Trustee, and one of the other ten competitors was chosen to play the opposite role. Role assignments were initially displayed on the screen, along with both players' names and photographs (Fig. 1). On average, subjects played ten games against each competitor, five as Investor and five as Trustee. The Investor was assigned four Money Units (MUs), and chose to invest between one and four with the Trustee (Investment \( I \)). The invested amount of money was tripled \((I \times 3)\) and given to the Trustee, who then decided what fraction of the tripled money to repay the Investor (Repayment \( R \), rounded to the nearest integer), which was displayed on the screen. Then, the final distribution across players was displayed on the screen (e.g., Liz: 8, Chris: 4). Subjects were told that earning during the game would partly influence their final pay for the experiment, and that playing cooperatively with other players would maximize their earnings. (In fact, subjects earned almost exactly the same amount in the Game, and all subjects were paid the same.) Half of the competitors played unfairly as Trustee, resulting in final distributions skewed toward the subject (“unfair”, \( R \) randomly selected between 0 and 1/4). The other half played fairly, resulting in equitable final distributions (“fair”, \( R \) between 1/3 and 2/3). The assignment of specific face photographs to fair versus unfair play was counterbalanced across subjects.

2.2. Story Task

In the second session, subjects first read short vignettes, each of which described an action of one of the players from the Game. The vignettes described an action that either lead to a positive or to a negative outcome (for example see Table 1). Scenarios were designed to leave the moral status of the actions partly ambiguous: no information about the intentions or beliefs that led to the action was provided. For every player, two stories with a positive and two with a negative outcome were presented, resulting in 40 stories in total. Each specific story was attributed to a previously fair player for half of the subjects, and to a previously unfair player for the other half. The photograph and the name of the player accompanied each story to help subjects identify the player, and to serve as implicit reminders of their behaviour in the Game. Stories were presented across runs in a counterbalanced order across conditions (fair versus unfair player, positive versus negative outcome).

After reading all of these stories, subjects were told that they would have to make judgments about the actions described in these stories: either about the intentions of the person (Behavioural Experiment) or about the blame- or praiseworthiness of the action (fMRI Experiment). A short sentence summarizing the story’s outcome was presented, accompanied by the player’s photograph and name (Fig. 2). Sentences were presented in randomized order. Each sentence was presented for 8 s followed by 6, 8 or 10 s rest period. In the Behavioural Experiment, subjects were asked to judge “How intentional was the action?” on a scale from 1 (not intentional) to 4 (definitely intentional). In the fMRI Experiment, subjects were asked to judge “How blame-/praiseworthy is the action?” (“blameworthy” for negative outcomes, “praiseworthy” for positive outcomes) on a scale from 1 (least blame-/praiseworthy) to 4 (most blame-/praiseworthy). Subjects made their response on a keyboard (Behavioural Experiment) or button box (fMRI Experiment).

2.3. Behavioural Experiment

The Story Task was conducted in the same room as the Game, within 24 h of the Game session. Stimuli were presented via Matlab 7.3 (Mathworks) and Psychtoolbox extensions (http://www.psychtoolbox.org) running on an iMac in white font on black background.

2.4. fMRI Experiment

The Story Task was conducted in the scanner, within 48 h of the Game session. Subjects were scanned using a Siemens Magnetom Tim Trio 3T system (Siemens Medical Solutions, Erlangen, Germany) in the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, using 30 4-mm-thick near-
axial slices with whole brain coverage (TR = 2 s, TE = 30 ms flip angle = 90°). Stimuli were presented in the scanner using a Hitachi (CP-X1200 series) projector displayed on a rear projection screen (Da-Lite) via Matlab 5.0/7.3 (Mathworks) and Psychtoolbox extensions (http://www.psychtoolbox.org) running on an Apple G4 laptop in white font on black background. Reaction time and response data were obtained during both MRI-experiments with a fiber-optic MR-safe button response box.

In addition to the Story Task, these subjects participated in a localizer experiment in the same scan session, contrasting reasoning about false non-moral beliefs (belief stories) with reasoning about non-social control situations (photograph stories), following the methods reported in Saxe and Kanwisher (2003, Experiment 2). Each story was presented for 10 s, followed by a short fill-in-the-blank question about the story (4 s). The stories were presented in counterbalanced order across runs and across subjects. Following the scan session, subjects took a short recognition memory test in order to determine if they remember the behaviour of each player during the Game. Subjects had three response options (fair–neutral–unfair) to differentiate between fair and unfair players. The memory task was conducted on a MacBook laptop immediately after subjects came out of the scanner. Finally, there was a short debriefing session about the story (4 s). The stories were presented in counterbalanced order across runs and across subjects.

Based on the localizer whole brain contrast (false belief versus false photograph), Theory of Mind regions of interest (ROIs) in each participant were defined as clusters of contiguous voxels with a higher BOLD response during ‘false belief’ than ‘false photograph’ stories (P < 0.0001, uncorrected), within 5 mm of the peak voxel in anatomical areas implicated in Theory of Mind by previous studies: PC, middle MPFC (mMPFC), dorsal MPFC (dMPFC), ventral MPFC (vMPFC) and bilateral TPJ (Ciaramidaro et al., 2007; Fletcher et al., 1995; Gallagher et al., 2000; Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007; Ruby & Decety, 2003; Saxe, Carey et al., 2004; Saxe & Kanwisher, 2003; Saxe, Moran, Scholz, & Gabrieli, 2006; Saxe & Powell, 2006; Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004; Vogeley et al., 2001; Young, Cushman, Hauser, & Saxe, 2007).

Within these ROIs, the average percent signal change (PSC) relative to rest baseline (PSC = 100 raw BOLD magnitude for (condition × fixation) raw BOLD magnitude for fixation) was calculated for each condition at each time point (averaging across all voxels in the ROIs and all blocks of the same condition). Adjusted for hemodynamic lag, PSC during stimuli presentation in each of the ROIs was compared across experimental conditions. Because the data defining the ROIs were independent from the data used in the fMRI Experiment, Type I errors were drastically reduced. All peak voxels are reported in Montreal Neurological Institute Coordinates.

Statistical analysis (Behavioural and fMRI Experiment) utilized post hoc paired-samples t-tests and repeated-measures ANOVAs, both conducted with an alpha level of 0.05.
to the second half [first half: 2.56; second half: 2.19; \( t_{(12)} = 4.24; P < 0.001 \)]. Average investment was significantly higher with fair than with unfair players in both halves [first half [fair: 3.23; unfair: 2.56; \( t_{(12)} = 9.01; P < 0.001 \]]; second half [fair: 3.32; unfair: 2.19; \( t_{(12)} = 11.67; P < 0.001 \)].

### 3.2. Behavioural Experiment

Subjects were asked to evaluate the outcomes of players’ actions on a scale from ‘not intentional’ (1) to ‘definitely intentional’ (4). To examine whether the outcome of the stories and the player’s former fairness influenced subject’s judgments, we conducted a 2 \( \times \) 2 (fairness [fair versus unfair] by outcome [positive versus negative]) repeated-measures ANOVA. The analysis yielded a significant main effect of outcome \([F_{(1,6)} = 32.29; P = 0.001; \text{partial } \eta^2 = 0.84] \) and a significant interaction of fairness and outcome \([F_{(1,6)} = 9.35; P = 0.02; \text{partial } \eta^2 = 0.61] \). In general, positive outcomes (mean: 3.3 of 4) were judged to be more intentional than negative outcomes (mean: 2.2 of 4). Subjects judged that both previously fair and previously unfair players intended the positive outcomes to the same degree (fair/positive: 3.35; unfair/positive: 3.28; \( t_{(6)} = 0.59; P = 0.58 \)) (see Fig. 3b). By contrast, subjects judged that previously unfair players intended the negative outcomes to a significantly greater degree than previously fair players (fair/negative: 1.9; unfair/negative: 2.4; \( t_{(6)} = -3.852; P = 0.008 \)). All seven subjects showed the same effect. There were no effects of condition on reaction times.

The magnitude of the fairness effect on judgments of negative stories’ intentional status (the difference in judging the intentional status of negative actions for unfair versus fair) was furthermore positively correlated with the fairness effect on subjects’ investments during the Game (the difference between investments with fair versus unfair players) \([\text{Pearson’s } r = 0.81; P = 0.03 \text{ (two-tailed)}] \). The more subjects differentiated between fair and unfair players in their investments, the more they judged that unfair players intended the negative outcomes more than fair players did.

### 3.3. fMRI Experiment, behavioural data

Subjects were asked to evaluate competitor’s actions in the short stories on a scale from ‘least blame-/praiseworthy’ (1) to ‘most blame-/praiseworthy’ (4). To examine whether the outcome of the stories and the player’s former fairness influenced subject’s judgments, we conducted a 2 \( \times \) 2 (fairness [fair versus unfair] by outcome [positive versus negative]) repeated-measures ANOVA. The analysis yielded a significant main effect of outcome \([F_{(1,25)} = 22.88; P < 0.001; \text{partial } \eta^2 = 0.49] \) and a significant interaction of the two factors \([F_{(1,25)} = 12.97; P = 0.001; \text{partial } \eta^2 = 0.34] \). In general, praise-judgments (praise for positive outcomes, mean: 3.19 of 4) were higher than blame-judgments (blame for negative outcomes, mean: 2.81 of 4) (see Fig. 3c). When judging a story with a positive outcome, subjects judged previously fair players as deserving more praise than unfair players [fair/positive: 3.29; unfair/positive: 3.08; \( t_{(25)} = 2.74; P = 0.01 \)]. By contrast, subjects judged previously unfair players as deserving more blame than fair players when their actions led to negative outcomes [fair/negative: 2.64; unfair/negative: 2.99; \( t_{(25)} = 3.08; P = 0.005 \)]. Twenty-two of 26 subjects showed this effect of fairness on moral judgment of negative outcomes [Binomial Test; \( P = 0.001 \)].

Across individual subjects, the magnitude of the fairness effect on moral judgments of negative actions (the difference in assigned blame to negative actions of unfair versus fair players) was not correlated with the effect of fairness on investments \([\text{Pearson’s } r = 0.021; P = 0.9 \text{ (two-tailed)}] \). One factor that may have contributed to the absence of this effect was that, according to their responses in the debriefing, some of the subjects deliberately tried to avoid letting their judgments be biased by their memory of the competitor’s fairness during the Game. If so, then responses during the experiment may underestimate the influence of prior fairness on moral intuitions for some subjects.

We also analyzed subjects’ reaction times by condition. The analysis revealed a significant main effect of outcome \([F_{(1,25)} = 6.8; P = 0.015; \text{partial } \eta^2 = 0.21] \) and an interaction between fairness and outcome \([F_{(1,25)} = 5.12; P = 0.03; \text{partial } \eta^2 = 0.17] \). Subjects were significantly slower when judging previously fair players whose actions led to negative outcomes than to any other condition \([\text{all } P < 0.05] \). There were no effects of subject’s gender on moral judgment.

After the scanner session, subjects rated the fairness of each subject on a scale of 1 (‘unfair’), 2 (‘neutral’) to 3 (‘fair’). The recognition memory task suggested that subjects could explicitly distinguish between fair and unfair players [fair: 2.52; unfair: 1.5; \( t_{(24)} = 13.76; P < 0.001 \)]. Correlation analyses revealed that individual differences in explicit memory for fairness were not correlated with any behavioural or neural measure of moral judgment.

At the end of the experiment, a debriefing explored whether subjects believed that the players were ‘real people’. Nine subjects did not believe that the competitors were “real”, seven were “not sure” and ten subjects believed that they were playing with and reading stories about real people. Post hoc analyses revealed no significant effects of this variable on subjects’ moral judgments, however, so all subjects contributed to the analyses below.

### 3.3.1. fMRI Experiment, imaging results

Our primary goal was to investigate the effect of prior record on the neural representation of the agent’s thoughts and intentions, during moral judgments. To do so, we analyzed the fMRI results in individually defined regions of interest, based on the localizer experiment. As predicted, a whole brain random effects analysis of the localizer experiment replicated former studies’ results (see Fig. 4a and b) [Gallagher et al., 2000; Gobbini et al., 2007; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Saxe & Powell, 2006; Young et al., 2007]: increased BOLD response during false belief, compared to false photograph stories was observed in bilateral TPJ, dMPFC, mMPFC, vMPFC and PC. We defined individual ROIs as follows: RTPJ (26/26 subjects), LTPJ (19/26), PC (24/26), mMPFC (15/26), dMPFC (20/26) and vMPFC (12/26) (Table 2). Inspection of the time-series revealed a late effect in raw PSC time courses of some regions (predominantly the RTPJ); the PSC in each ROI was therefore calculated for two time intervals: during the moral judgment (up to average RT, 5 s after sentence onset), and immediately after judgment. Allowing for the hemodynamic lag, these intervals were: during (Time1: 4–10 s) and after (Time2: 12–18 s) moral judgment. Each ROI’s response was then analyzed using a 2 \( \times \) 2 \( \times \) 2 (fairness [fair versus unfair] by outcome [positive versus negative]) repeated-measures ANOVA.
by time [Time1 versus Time2]) repeated-measures ANOVA of the average PSC.
The most robust effects of fairness were observed in the RTPJ. The response in this region showed a significant interaction of the three factors \(F(1,25) = 8.19; P = 0.008; \text{partial } \eta^2 = 0.25\). We therefore analyzed the response separately at Time1 and Time2. At Time1 there were no significant effects. Only after moral judgment (Time2), RTPJ showed a significant interaction between fairness and outcome \(F(1,25) = 9.31; P = 0.005; \text{partial } \eta^2 = 0.27\). In stories with negative outcomes, activation was significantly increased for previously unfair, as compared to fair players [unfair/negative: 0.09; fair/negative: −0.06; \(t_{25} = –2.92; P = 0.007\)] (see Fig. 5). Twenty of 26 subjects showed this pattern of response [Binomial Test; \(P = 0.01\)]. There was no effect of fairness on the response to positive outcomes.

Since there was some variation across individuals in the effect of fairness on the RTPJ response, we next investigated whether the magnitude of this effect was predicted by subject’s behavioural performance during the Game and/or the fMRI Experiment. To test this hypothesis, we computed three difference scores for each subject:

1. **RTPJ difference**: the difference in RTPJ response to negative outcomes at Time2, for unfair versus fair players.
2. **Investment difference**: the difference between investments with fair versus unfair players during the Game (a measure of subject’s original sensitivity to the fairness manipulation).
3. **Moral judgment difference**: the difference between assigned blame to unfair versus fair players for negative actions in the Story Task.

We found that the RTPJ difference was positively correlated with the original investment difference [Pearson’s \(r = 0.45; P = 0.02\) (two-tailed)] (see Fig. 6). The more a subject differentiated between fair and unfair players during the initial Game, the more that subject’s RTPJ would later differentiate between fair and unfair players, after moral judgments of negative outcomes. There was no correlation between the RTPJ difference and differences in moral judgments in the scanner [Pearson’s \(r = 0.66; P = 0.75\) (two-tailed)].

Other Theory of Mind brain regions showed similar, but less robust, response profiles. In the PC, the \(2 \times 2 \times 2\) ANOVA revealed a marginally significant interaction of time (Time1 versus Time2) and outcome (positive versus negative) \(F(1,23) = 4.06; P = 0.056; \text{partial } \eta^2 = 0.150\). A separate analysis of each time suggested that the PC showed the same profile as the RTPJ, but less reliably. There were no effects of fairness or outcome at Time1. At Time2, the response was higher when unfair players’ actions led to negative outcomes, as compared to positive outcomes [unfair/positive: −0.03; unfair/negative: 0.09; fair/negative: –2.14; \(P = 0.043\)].

Both the LTPJ \(F(1,18) = 7.02E–02; P = 0.049; \text{partial } \eta^2 = 0.19\) and the dMPFC \(F(1,19) = 4.51; P = 0.047; \text{partial } \eta^2 = 0.19\) showed a significant three-way interaction of time (Time1 versus Time2), fairness (fair versus unfair) and outcome (positive versus negative) in the \(2 \times 2 \times 2\) ANOVA. However, separate analysis of the two time intervals in both regions showed no significant effects at either time. Nevertheless the response in the LTPJ resembled that of the RTPJ and PC; the highest response was when unfair subject’s action led to negative outcomes [fair/positive mean: 0.01; fair/negative mean: −0.01; unfair/positive mean: 0.04; unfair/negative mean: 0.15].
The vMPFC and the mMPFC showed no significant main effects or interactions by condition, either in the $2 \times 2 \times 2$ ANOVA, or in a collapsed analysis of both time intervals (4–18 s).

Finally, we conducted whole brain analyses of the Story Task directly. To investigate neural correlates of representing positive versus negative outcomes, we conducted a whole brain random effects analysis of the Story Task with two contrasts: (1) positive > negative and (2) negative > positive. The first contrast (positive > negative) produced no reliable regions of activation. The second contrast (negative > positive) revealed a significant cluster of activation in the dMPFC at $P<0.001$ (global peak at $x=4, y=42, z=46$) (see Fig. 7). Next, we used two contrasts to look for overall effects of prior record on processing of action outcomes: (3) fair > unfair, and (4) unfair > fair. These contrasts produced no reliable activations.

4. Discussion

The current study provides neural and behavioural clues concerning the role of prior record in both moral judgment and attributions of intention. As predicted, prior record influenced both of these aspects of social cognition: previously unfair competitors were judged to be more blameworthy (broadly mirroring legal practice) and to have acted more intentionally when causing negative outcomes, as compared to previously fair competitors. Neural activation in regions associated with mental state reasoning was also affected by prior record (e.g., fairness), as we discuss below.

Consistent with previous studies (Berg et al., 1995; de Quervain et al., 2004; Haselhuhn & Mellers, 2005; Singer, Kiebel, Winston, Dolan, & Frith, 2004; Singer et al., 2006), the economic game provided subjects with negative (unfair) and positive (fair) personal experiences with the competitors. During the Game, subjects quickly and selectively decreased their investments with unfair players. The resulting personal impressions were robust and enduring: previous experience with the other players subsequently biased moral judgments made up to 2 days later. Unfair players' harmful actions were judged to be both more intentional and more blameworthy, compared to judgments of the same actions when performed by fair players.

The bias in the moral judgment was accompanied by a distinctive neural response. Theory of Mind brain regions, especially the RTPJ, showed significantly higher BOLD response to harmful outcomes caused by unfair as opposed to fair players. There was no effect of prior record on response to positive outcomes, mirroring the pattern observed in judgments of intentionality. In addition, the differential response in the RTPJ was correlated with individual subjects' earlier discrimination of fair and unfair players during the economic game, emphasizing a link between the subjects' prior experience and mental state reasoning in the later moral judgment task.

We found no evidence for increased BOLD response, in the RTPJ or other Theory of Mind brain regions, to actions that violated prior expectations in general (e.g., positive outcomes produced by unfair players, negative outcomes produced by fair players). These results speak against the recent proposal that the role of the RTPJ in social tasks is in comparing predictions with incongruent outcomes and directing attention towards salient or unexpected events (Decety & Lamm, 2007). Decety and Lamm (2007)'s view was informed by the existence of a nearby region of RTPJ that is implicated in exogenous attention (the right inferior parietal component of the ventral attention network, Corbetta and Shulman (2002)). By contrast, we suggest that distinct regions within the RTPJ may be involved in attentional reorienting and Theory of Mind. Both Decety and Lamm (2007), and Scholz and colleagues (personal communication) observed approximately 10 mm of separation between peaks for the exogenous attention and Theory of Mind tasks. Although Decety and Lamm (2007) concluded that this difference was small in the context of their meta-analysis, Scholz and colleagues (personal communication) found that the same difference was reliable within individual subjects.

Given this anatomical separation, we believe that the functional localizer approach used in the current study allowed us to identify and investigate the specific sub-region of the RTPJ implicated in Theory of Mind. In this region, we observed no effect of "violation of expectation" on the neural response. Instead, we observed selective enhancement of the response to negative outcome produced by previously unfair players, which we interpret below in terms of the interaction between Theory of Mind and moral judgment.

Recent behavioural research in moral psychology emphasized the importance of Theory of Mind in moral judgments (Cushman, Young, & Hauser, 2006; Koenigs et al., 2007; Mikhal, 2007; Pizarro, Uhlmann, & Salovey, 2003; Woolfolk et al., 2006). Cushman (personal communication) found that the agent's beliefs and desires are the first and second most important factors, respectively (followed by outcomes) in determining observers' moral judgments. Consistent with those results, we have previously reported that brain regions involved in Theory of Mind are systematically recruited during moral judgment. When beliefs were presented explicitly, the RTPJ, PC and LTPJ showed an initial response at the time the belief was presented that did not depend on the valence of the belief (negative versus neutral), and an additional response at the time that the outcome was presented, that did depend on the valence of the outcome (Young et al., 2007; Young & Saxe, 2008). We concluded that Theory of Mind brain regions are involved in both the initial encoding of belief information, and the subsequent integration of beliefs with outcomes, in order to support mature moral judgment.

Unlike our previous experimental stimuli, though, in real life people's mental states, including their beliefs, are often not explicitly available, but must be inferred from other information (Young & Saxe, in press). We hypothesized that one key source of such information would be the observer's impressions of the actor's prior record, based on personal experience and/or knowledge of the actor's offense history (Nadelhoffer, 2004a, 2004b; Phelan & Sarkissian, 2008; Pizarro et al., 2006; Woolfolk et al., 2006). The behavioural results of the current experiment support this hypothesis. Subjects read vignettes that described actions with positive or negative actions, but did not explicitly state the beliefs or desires of the protagonists (Young & Saxe, in press). Nevertheless, subjects made systematically differential intentional attributions across conditions. Given a negative personal interaction, observers judged that the protagonist was more likely to have intended the negative outcomes, compared to judgments of the same outcomes follow-
ing observers' positive personal experiences with the protagonist. Across individuals, judgments of the intentional status of actions were correlated with the effect of fairness on subjects' investments during the Game. One unpredicted but robust finding was that the effect of fairness on the RTPJ response occurred late in the time-series, after moral judgments were made. We propose that after subjects judged the action to be blameworthy, they continued to consider the possible mental states of the protagonist. One possibility is that subjects first feel an impulse to blame previously unfair people for causing negative outcomes, and then subsequently seek to justify this impulse by attributing to them negative intentions (Cushman et al., 2006; Haidt, 2007).

The current results may thus be a specific instance of a general phenomenon, known as the Side-Effect Effect (S-E-E) (Knobe, 2003, 2004, 2006; Knobe & Burra, 2006; Leslie, Knobe, & Cohen, 2006), a recent puzzle in moral psychology. The Side-Effect Effect is the observation that when a protagonist's action causes a side-effect that is foreseen but not directly intended (i.e., the protagonist says "I don't care about [the side-effect]"), observers judge that the side-effect was “intentional” if the side-effect is negative, but not if the side-effect is positive. The S-E-E can be induced by changing a single word in the description of the side-effect (e.g., “harm” versus “help”). An elegant recent series of experiments shows that this effect is robust across many variations of the task format (Pettit & Knobe, 2008). Even 4-year-old children show an adult-like pattern of this effect (Leslie et al., 2006).

The S-E-E poses a challenge to a traditional model of folk morality, according to which Theory of Mind serves only as an input to moral judgment: observers try to establish whether the outcome was intended or not based on evidence about beliefs and desires, and then rely exclusively on this information to generate a moral judgment. Instead, it seems that moral judgment (whether the action is blameworthy or praiseworthy) also influences Theory of Mind judgment (whether the side-effect is perceived as having been brought about intentionally or not) (Knobe, 2005).

One interpretation of the S-E-E is that subjects first feel an impulse to blame the protagonist who knowingly caused the negative side-effect; in order to justify this impulse, subjects then attribute to the protagonist a clearer negative intention. This interpretation fits with two aspects of the current results: (1) the timing of the effect of fairness in the RTPJ, which emerged only after moral judgment and (2) the correlation between the magnitude of the effect in the RTPJ and the earlier investment during the Game. Both the S-E-E and the current fMRI results may therefore suggest a common psychological mechanism for post hoc blame justification (Nadelhoffer, 2004a, 2004b; Phelan & Sarkissian, 2008).

The current results may provide insight into the effects of prior record on intuitive moral judgment. Legal practice suggests a range of other factors that may also be relevant for folk morality. In the law, prior record has distinct consequences for severity of punishment (i.e. sentencing) versus for judgments of blameworthiness (i.e. conviction). The role of offense history is further modulated by the similarity of means and consequences between prior actions and the current accusation (i.e. habit or routine practice). These determinants of legal practice may have interesting psychological and neural implications. More generally, future research using cognitive neuroscience methods will help to characterize the common ground between folk morality and legal practice, as well as the specific contexts in which they diverge.

Acknowledgments

The authors are grateful to Josh Tenenbaum, John Mikhail, Christina Triantafyllou, Laura Schulz, Mike Frank, and Manfred Herrmann for their intellectual and technical support, and to Allen Mintz for help collecting the behavioural data. The fMRI data collection was supported by the NCRR (grant P41RR14075), the MIND Institute, and the Athinoulia A. Martinos Center for Biomedical Imaging.

References


