Journal of Experimental Psychology: General
2016, Vol. 145, No. 11, 1448–1459

© 2016 American Psychological Association
0096-3445/16/$12.00 http://dx.doi.org/10.1037/xge0000190

Reflexive Intergroup Bias in Third-Party Punishment

Daniel A. Yudkin
New York University

Tobias Rothmund and Mathias Twardawski
University of Koblenz-Landau

Natasha Thalla
Lehigh University

Jay J. Van Bavel
New York University

Humans show a rare tendency to punish norm-violators who have not harmed them directly—a behavior known as *third-party punishment*. Research has found that third-party punishment is subject to *intergroup bias*, whereby people punish members of the out-group more severely than the in-group. Alhough the prevalence of this behavior is well-documented, the psychological processes underlying it remain largely unexplored. Some work suggests that it stems from people's inherent predisposition to form alliances with in-group members and aggress against out-group members. This implies that people will show reflexive intergroup bias in third-party punishment, favoring in-group over out-group members especially when their capacity for deliberation is impaired. Here we test this hypothesis directly, examining whether intergroup bias in third-party punishment emerges from reflexive, as opposed to deliberative, components of moral cognition. In 3 experiments, utilizing a simulated economic game, we varied participants' group relationship to a transgressor, measured or manipulated the extent to which they relied on reflexive or deliberative judgment, and observed people's punishment decisions. Across group-membership manipulations (American football teams, nationalities, and baseball teams) and 2 assessments of reflexive judgment (response time and cognitive load), reflexive judgment heightened intergroup bias, suggesting that such bias in punishment is inherent to human moral cognition. We discuss the implications of these studies for theories of punishment, cooperation, social behavior, and legal practice.

Keywords: third-party punishment, intergroup bias, cooperation, altruism, fairness

Human behavior is characterized not just by extraordinary altruism, but also by extraordinary punishment. Many organisms retaliate against those who have harmed them directly—a behavior known as "second-party punishment" (Fehr & Gächter, 2000; Jensen, Call, & Tomasello, 2007). Humans, however, also demonstrate a willingness to punish those who have harmed someone else—known as "third-party punishment" (Buckholtz et al., 2008; Carlsmith, Darley, & Robinson, 2002; Fehr & Fischbacher, 2004). Because third-party punishment can deter members of a cooperative community from acting selfishly, it may have played a role in promoting the kinds of cultural and technological advancements that characterize the human species (Boyd & Richerson, 1992, 2009; Fowler, 2005; Gardner & West 2004). Third-party punishment is observed in children; Hamlin, Wynn, Bloom, & Mahajan, 2011; McAuliffe, Jordan, & Warneken, 2015) and across a range of human societies (Henrich et al., 2006), though not in chimpanzees (Reidl, Jensen, Call, & Tomasello, 2012), suggesting it is a uniquely human trait (but see Raihani, Grutter, & Bshary, 2010).

Considerable research has demonstrated that people judge and punish out-group members more harshly than in-group—a phenomenon known as "intergroup bias" (Baumgartner, Götte, Gügler, & Fehr, 2012; Efferson, Lalive, & Fehr, 2008; Hewstone, Rubin, & Willis, 2002; Mussweiler & Ockenfels, 2013; Tajfel & Turner, 1979). One study, for instance, found that Swiss Army officers punish transgressions committed by a member of a different platoon more harshly than those committed by one of their own—especially when they are in a competitive environment against other platoons (Goette, Huffman, Meier, & Sutter, 2010). Similarly, undergraduates who were asked to imagine a scenario involving the theft of a significant sum of money assigned higher fines to foreign offenders than to family or classmates (Lieberman & Linke, 2007). This pattern of intergroup bias in third-party punishment has also been observed in people's soccer clubs and political parties (Schiller, Baumgartner, & Knoch, 2014). Similar behavior has been observed among tribesmen in Papua New

This article was published Online First September 15, 2016.

Daniel A. Yudkin, Department of Psychology, New York University; Tobias Rothmund, Department of Psychology, University of Koblenz-Landau, Landau, Germany; Mathias Twardawski, Graduate School "Learning and Teaching Processes," University of Koblenz-Landau, Landau, Germany; Natasha Thalla, Department of Psychology, Lehigh University; Jay J. Van Bavel, Department of Psychology, New York University.

We thank members of the New York University *Social Perception and Evaluation Lab* (@vanbavellab) for helpful comments on this article. We acknowledge the New York University Dean's Undergraduate Research Fellowship and Forschungsinitiative Rheinland-Pfalz, Germany, for financial support. Responsibilities: DAY, TR, NT, and JVB designed the experiments; MT and NT collected data; DAY, MT and NT analyzed the data with input from JVB; and DAY wrote the manuscript with critical edits from TB, MT, NT, and JVB.

Correspondence concerning this article should be addressed to Jay J. Van Bavel, Department of Psychology, New York University, 6 Washington Pl, 4th Floor, New York, NY 10003. E-mail: jay.vanbavel@nyu.edu

Guinea, who punished out-group offenders more harshly than in-group members—especially when the victim is a member of their own tribe (Bernhard, Fehr, & Fischbacher, 2006). And this pattern of intergroup bias emerges early in development: Both 6- and 8-year-olds punish peer transgressors more severely when they are members of a different team (Jordan, McAuliffe, & Warneken, 2014).

The fact that intergroup bias in third-party punishment has been observed in such a wide variety of experimental methods, across cultures, and in very young children raises the possibility that this behavior stems from reflexive processes in the human mind. Consistent with this idea, anthropological research suggests that, because humans spent much of their evolutionary history in small tribes in competition for scarce resources, they have a natural tendency to view out-group members with distrust and hostility (Balliet, Wu, & De Dreu, 2014; Halevy, Weisel, & Bornstein, 2011; King & Wheelock, 2007; Richerson & Boyd, 2001; Wrangham & Peterson, 1997). Other research has shown that people view group identities as a boundary for cooperative behavior (Makimura & Yamagishi, 2003; Yamagishi, Jin, & Kiyonari, 1999; Yamagishi & Kiyonari, 2000) and may use them to forge alliances in order protect themselves against possible retaliation in punitive contexts (DeScioli & Kurzban, 2013).

Although this research gives hints as to the cognitive underpinnings of intergroup bias in third-party punishment, the question of whether intergroup bias in punishment stems from inherent features of the moral mind has never been examined experimentally. Here, we attempt to redress this gap by building off a growing body of evidence suggesting that evaluations may be driven by different types of mental processing, ranging from reflexive to deliberative (Chaiken & Trope, 1999; Cushman, 2013; Greene, 2013). Reflexive judgments are effortless and automatic, stemming from evolutionarily ancient subcortical regions of the brain such as the limbic system and the amygdala (Greene, 2007; Greene & Haidt, 2002); deliberative judgments, by contrast, are more effortful and controlled, arising from cortical areas associated with executive functioning and working memory (Coolidge & Wynn, 2001; D'Esposito, Postle, & Rypma, 2000). Although the components of the human mind and brain are widely distributed and highly interactive (see Cunningham, Zelazo, Packer, & Van Bavel, 2007), this dichotomy has heuristic value for understanding and predicting human behavior (Chaiken & Trope, 1999).

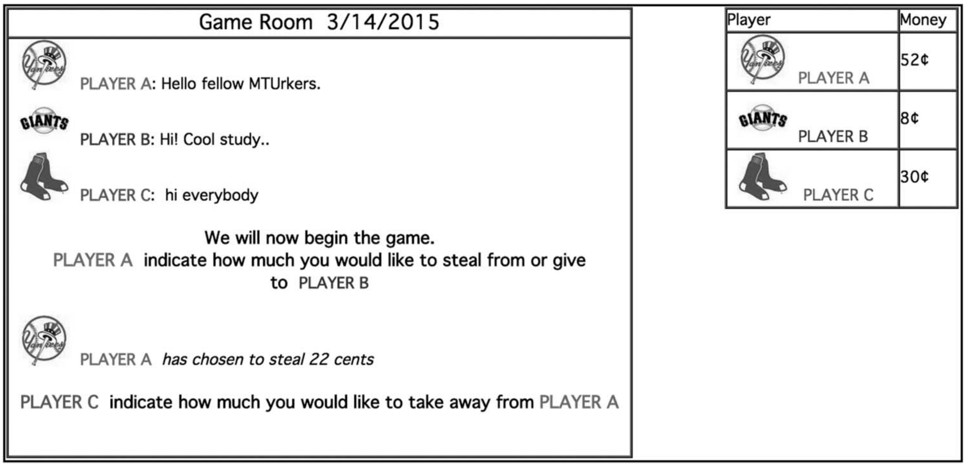
Recent work has shown that different types of processing can give rise to distinct moral judgments. For instance, whereas reflexive decision-making results in deontological judgments (those based on actors' rights or duties), providing people the opportunity for greater deliberation shifts their judgments in favor of consequentialist considerations (i.e., those based on outcome; Greene, 2007; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008). Other work suggests that humans are generally more cooperative under time pressure (Rand, Greene, & Nowak, 2012). Inducing reflexive judgment via a cognitive load task caused people to cooperate more with in-group members than out-group members (De Dreu, Dussel, & ten Velden, 2015). Thus, reflexive and deliberative processes may elicit different moral judgments and decisions (although see Van Bavel, FeldmanHall, & Mende-Siedlecki, 2015, for the limitations of this approach).

Overall, then, this research implies that reflexive versus deliberative processing may exacerbate intergroup bias in third-party punishment. We thus set out to test this question by experimentally observing people's punishment of in-group and out-group members under conditions of both reflexive and deliberative judgment. This research is important because they have the potential to shed light on the functional role such punishments are designed to serve. Several different outcomes were possible in this experiment, each with their own implications about the role of punishment in groups. First, people may demonstrate more intergroup bias when punishing reflexively than deliberatively—the *reflexive intergroup bias hypothesis*. This would be consistent with past work suggesting that reflexive judgment exacerbates intergroup bias (e.g., De Dreu et al., 2015; Greenwald, McGhee, & Schwartz, 1998). Second, reflexive judgment may eliminate intergroup bias in third-party punishment—the *reflexive egalitarian hypothesis*. Such a finding would be in line with the notion that humans are reflexively egalitarian in their punishment decisions (Valdesolo & DeSteno, 2008). Third, reflexive judgment may heighten people's punishment of *in-group* members—the *reflexive group regulation hypothesis*. This would be in line with theories of group regulation and with the black sheep effect, which suggest that third-party punishment may have emerged as a way for people to police the behavior of members of their own group (e.g., Fearon & Laitin, 1996; Fehr & Gächter, 2002; Marques, Yzerbyt, & Leyens, 1988; Mendoza, Lane, & Amodio, 2014). Overall, then, we sought to examine the effects of reflexive judgment on third-party punishment, hoping to better understand the cognitive processes that underlie this important human behavior.

Overview

The current research was designed to help arbitrate this theoretical debate about the relative contribution to judgment of reflexive versus deliberative decision-making in third-party punishment. We conducted three experiments employing two methods and three distinct intergroup contexts. In Experiment 1, we measured participants' response time in the punishment decision. Faster response times have been shown to be associated with more reflexive judgment (e.g., Bargh & Chartrand, 1999; Rand et al., 2012). We thus planned to examine whether responding quickly to the transgression had different effects depending on whether people were punishing in-group or out-group members. Next, we tested the causality of this relationship in Experiments 2 and 3 by manipulating cognitive load. This manipulation interferes with executive functioning and increases the reliance on reflexive judgment (e.g., Gilbert & Osborne, 1989; Manoach et al., 1997).

Across experiments, participants entered a simulated online environment in which they interacted with two other "players" (see Figure 1). In all conditions, participants observed one of the players "steal" the majority of another player's money. Participants were then given the opportunity to punish the transgressor by deducting some, all, or none of his or her money and removing it from the game. No one received the deducted money. The primary dependent measure was the percentage of the transgressor's money deducted as punishment. To test for the effects of group membership on punishment decisions, we randomly assigned subjects to punish either an in-group or an out-group transgressor. We then either measured (Experiment 1) or manipulated (Experiments 2 and 3) the extent to which people relied on reflexive judgment. In an effort to extend the generalizability of our findings, we em-



Game Room 3/14/2015

PLAYER A: Hello fellow MTUrkers.

PLAYER B: Hi! Cool study..

PLAYER C: hi everybody

We will now begin the game.

PLAYER A **indicate how much you would like to steal from or give to** PLAYER B

PLAYER A *has chosen to steal 22 cents*

PLAYER C **indicate how much you would like to take away from** PLAYER A

Player	Money
PLAYER A	52¢
PLAYER B	8¢
PLAYER C	30¢

Figure 1. Screenshot from the punishment phase of Experiment 3. Participants (Player C) are logged into a “Game Room,” where they see the icons and greetings of two other players. Next they observe the transgressor (Player A) “steal” a significant portion of Player B’s resources and are asked to indicate the amount of money they want to deduct as punishment from Player A and remove from the game (shown here). In this example, Player A has elected to steal $0.22, so Player A now has $0.52, Player B has $.08, and Player C has $0.30. Finally, participants are informed of the totals to be paid to all the players, thanked for their participation, and logged off the Game Room.

ployed a different group membership manipulation in each experiment, including American football teams, baseball teams, and nationality. The victim of the transgression was always portrayed as a member of an unrelated out-group. We collected several other variables, including in-group identification, mental effort expended, and perceived fairness of the self and others’ actions.

Experiment 1: NFL Teams

Participants

Because there was no previous research on this specific set of hypotheses, we assumed a medium effect size (Cohen’s $d = .5$, power $(1 - \beta) = 80\%$) and recruited 100 American participants through Amazon’s Mechanical Turk (MTurk) to participate in a short study titled “Football Fortunes” in exchange for $0.30. MTurk has been shown to provide a reliable and diverse subject pool that behaves in ways consistent with known psychological phenomena (Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013). Of the original 100, 9 failed to complete the study and 4 were dropped for failing a basic attention check, leaving a total of 87 (33 female; $M_{\text{age}} = 29.9$). We report how we determined our sample size, all data exclusions (if any), all manipulations, and all primary measures in this and all subsequent studies. All additional analyses can be found at Open Science Framework at osf.io/pmvtj/.

Materials and Procedure

Participants were informed that the purpose of the experiment was to see how people make decisions; they were told that they would soon be entering the Game Room, an online forum that allowed players to interact in real-time and make decisions involving real money. After completing basic demographic questions, participants went on to receive a series of formal instructions about the protocol of the interaction sequence that would take place in the Game Room. Participants learned that they would be randomly assigned to one of three roles: Player A, B, or C. Each player would start with the same amount of money ($0.30), which was displayed in a table on the screen. The interaction in the Game Room would consist of two “turns.” In Turn 1, Player A would decide how much money he or she would like to give or steal from Player B. The range of options included everything from giving all of his or her own money away to stealing all of Player B’s money. The word “steal” was employed in the instructions in order to ensure that participants understood the act as a transgression. In Turn 2, Player C would make a decision about how much money he or she wanted to deduct from or give to Player A. Any money taken from Player A would be removed from the game and received by no one.

After receiving these instructions and passing a short comprehension quiz, all participants were “randomly” assigned to the role of Player C, and then asked to indicate their favorite NFL football team to be used as an avatar in the game. Next they completed a series of items assessing their level of fandom with the team (“I am PROUD to be a _______ fan”; “I value being a fan of ______”; “Being a fan of _______ is an important part of my identity”) on a 6-point Likert scale ranging from *strongly agree* to *strongly disagree* ($\alpha = .93$) (Van Bavel & Cunningham, 2012). They were then logged into the Game Room and joined, after a brief interval, by the two simulated Players A and B.

The group identity of the players was randomly manipulated such that Player A (the “thief”) either was a fan of the same football team as the participant or was a fan of a different team (Buffalo Bills). Player B never shared a team membership with Player A or Player C (either Baltimore Ravens or the Arizona Cardinals). Once in the Game Room each player was given a chance to enter a greeting. Player A wrote, "hi mturkers. cool study. i love football season." Player B wrote, "whats up everyone/ Go Cardinals!!" Player C was subsequently given an opportunity to respond, and was then told by the computer moderator that the game would begin.

In each condition, Player A took 14 cents away from Player B, leaving B with 16 cents and Player A with 44. Player C (the participant) was then given the opportunity to indicate how much of Player A's money he or she wanted to deduct as punishment and remove from the game. The game then ended and participants were directed to answer some follow-up questions on a sliding scale, which included, "How fairly/unfairly did Player A act?" (*very unfairly* to *very fairly*), "How much do you like/dislike Player A?" (*dislike very much* to *like very much*), "How fairly/unfairly did you act?" (*very unfairly* to *very fairly*), and "How similar/dissimilar do you feel to Player A?" (*very dissimilar* to *very similar*). (All follow-up questions can be found at the Open Science Framework at osf.io/pmvtj/).

Results and Discussion

To test the question of whether reflexive judgment leads to an increase in intergroup bias in third-party punishment, we log-transformed punishment response time to normalize it (see Whelan, 2008). We assigned dummy variables 0 and 1 to people who observed in-group and out-group transgressors, respectively, then entered transgressor group membership, mean-centered response time, and their interaction into a linear regression (Aiken, West, & Reno, 1991). We calculated the percent of the transgressor's money participants deducted as punishment and entered this as the dependent variable. The results indicated a main effect of transgressor identification, showing that participants deducted a higher percentage of out-group transgressors' money ($M = 52.5\%$, $SD = 31.8$) than in-group transgressors' money ($M = 35.0\%$, $SD = 31.6$), $B = .714$, $SE = .261$, $p = .008$, **d** = .59. Importantly, the results also revealed a significant interaction, $B = -.211$, $SE = .10$, $p = .038$, **d** = .46. Whereas participants with a slower response time (-1 *SD*) deducted similar proportions from in-group (41.4%) and out-group (44.8%) members, $B = .03$, $SE = .10$, $p = .73$, **d** = .08, those with faster response times ($+1$ *SD*) punished out-group members (62.5%) significantly more than in-group members (29.1%), $B = .33$, $SE = .10$, $p = .001$, **d** = .72. Put another way, although those who punished quickly deducted a higher proportion of out-group members' money (62.5%) than those who punished slowly (44.8%), $B = -.125$, $SE = .063$, $p = .054$, **d** = .62, punishment of in-group members demonstrated the opposite trend (although this difference did not reach statistical significance), with fast punishers confiscating directionally less of in-group members' money (29.1%) than slow punishers (41.4%), $B = .087$, $SE = .078$, $p = .274$, **d** = .34. Consistent with the *reflexive intergroup bias hypothesis*, out-group members received harsher punishments than in-group members when participants responded swiftly.

Experiment 2: National Identity

Experiment 1 provided preliminary justification for the notion that reflexive judgment is associated with intergroup bias in third-party punishment. However, recent research calls into question the extent to which faster RTs can be taken as unequivocal evidence of reflexive judgment. In contrast to work linking decision speed to processing style (see Chaiken & Trope, 1999; Kahneman, 2011), other factors can influence the speed with which a decision is made. For instance, decision conflict drives slower RTs (Evans, Dillon, & Rand, 2015). Additionally, slower or faster decisions can be elicited as an artifact of experimental design in cases when a series of decisions is skewed in the relative attractiveness of different options (Krajbich, Bartling, Hare, & Fehr, 2015). Although the one-shot nature of our experimental procedure renders implausible the view that our effects could be the result of strength-of-preference artifacts, it is possible that participants were less conflicted when punishing out-group members than in-group members—which could produce a pattern of results similar to the ones we observed. It is therefore impossible to infer a causal relationship between reflexive judgment and intergroup bias in punishment solely on the basis of Experiment 1.

To determine if reflexive judgment directly *causes* intergroup bias in punishment, we randomly subjected participants to a cognitive load manipulation in two follow-up experiments. Prior research has shown that cognitive load induces a reliance on reflexive judgment (Gilbert & Osborne, 1989; Manoach et al., 1997). As with Experiment 1, participants were randomly assigned to observe either an in-group or an out-group transgressor steal a third-party's resources. They then were given the opportunity to punish the transgressor by deducting some or all of his or her money from the game. The primary dependent variable was the amount of money deducted as punishment. In line with Experiment 1, we expected that reflexive judgment, as operationalized by cognitive load, would exacerbate intergroup bias in punishment, supporting the *reflexive intergroup bias hypothesis*. Furthermore, to ensure that the results would generalize to other intergroup contexts, we varied the manner in which group membership was operationalized: In Experiment 2, we recruited an Indian sample and used national identity as the manipulation of group membership.

Participants

On the basis of the results from Experiment 1, we anticipated a small-to-medium effect size (Cohen's $f = .18$, power $[1 - \beta] = 80\%$) and recruited 268 Indian participants through MTurk to participate in a short study titled "Social Fun and Games" in exchange for $0.30. Of our initial sample, 54 participants either failed to complete the study, failed an attention check, or did not grant consent to use their data, leaving a final sample of 213 (86 female, $M_{\text{age}} = 29.5$).

Materials and Procedure

The design of this study was similar to that of Experiment 1. Group identity was manipulated by nationality: Participants were asked to select an icon corresponding to the flag of their home country. In-group transgressors were ostensibly from the same country as participants (India); out-group transgressors selected a flag from the United States. The victim indicated an Estonian nationality. To manipulate cognitive load, participants in the "high load" condition were asked to memorize the letter string "7T4$RF%" and told that they would be tested on it at a later phase of the experiment, whereas participants in the "low load" condition were given no such instruction.

Results and Discussion

Manipulation checks. To test whether the manipulation of cognitive load was successful, we compared the means in log-transformed first-click punishment reaction times (RTs) between groups. In line with literature suggesting that cognitive load occupies deliberative systems and working memory (see Barrouillet et al., 2007; Greene et al., 2008), we expected that those in the "high load" condition would take longer to punish than those in the "low load" condition. Surprisingly, no significant differences emerged between conditions, $p > .5$. Another method of treatment, which entails trimming outliers that lie above the 95th percentile (32.9 seconds; Ratcliff, 1993), showed effects trending in the predicted direction (low load: $M = 9.2$, $SD = 5.7$; high low: $M = 10.7$, $SD = 7.4$), $t(201) = -1.56$, $p = .118$, $\mathbf{d} = .22$; this provides limited support that the manipulation of cognitive load was successful. To ensure that these outliers do not alter the findings reported below, we analyzed the data both with and without them; all significance values are the same under both approaches. The results reported include all outliers; analyses excluding outliers can be found at Open Science Framework at osf.io/pmvtj/.

Effect of group and load on punishment. To test the hypothesis that intergroup bias in third-party punishment is exacerbated under load, we analyzed the data with a 2 (*transgressor identity*: in-group vs. out-group) × 2 (*cognitive load*: high vs. low) analysis of variance (ANOVA), with money deducted as the dependent variable. The results revealed a significant interaction between load condition and transgressor identity, $F(1, 209) = 6.61$, $p = .011$, $\mathbf{f} = .18$ (see Figure 2B). To better understand the effects of reflexive judgment on punishment of in-group and out-group members, we tested the simple effects of transgressor group membership under each load condition. Results indicated that inducing cognitive load caused significant differences in the punishment of in-group and out-group members. Those under low load showed no such difference in punishment according to transgressor identification (out-group: $M = 47.4\%$, $SD = 27.1$; in-group: $M = 51.0\%$, $SD = 27.1$), $F(1, 209) = 0.49$, $p = .484$, $\mathbf{d} = .09$. By contrast, those in the high cognitive load condition punished out-group members ($M = 60.6\%$, $SD = 26.5$) significantly more than in-group members ($M = 44.6\%$, $SD = 30.6$), $F(1, 209) = 8.19$, $p = .005$, $\mathbf{d} = .40$. Consistent with the *reflexive intergroup bias hypothesis*, out-group members received harsher punishments than in-group members when participants were under cognitive load.

Because participants were performing this study at their own computers, it is possible that they could undermine the efficacy of the cognitive load manipulation by writing the code down. This possibility only makes our results more conservative, because it would only serve to weaken the effects. Nevertheless, we wanted to verify that the manipulation was successful only among those who had not written down the code. When we reanalyzed the data with only the 88.3% of participants who explicitly stated they had not written down the code, the critical interaction term remained statistically significant, $F(1, 174) = 4.03$, $p = .046$, $f = .14$, and none of the significance values of the other reported statistics were altered. These data conceptually replicate the results of Experiment 1—further supporting the notion that reflexive judgment elicits intergroup bias in third-party punishment and the *reflexive intergroup bias hypothesis*.

Experiment 3: Baseball Teams

Experiment 2 bolstered the claims of Experiment 1 by showing how reflexive judgment is causally implicated in increased intergroup bias in punishment. However, the RT data provided only qualified evidence that the cognitive load manipulation was successful, and we did not test participants' degree of mental effort directly. We conducted Experiment 3 in order to address these issues and replicate the results of Experiment 2. Additionally, we

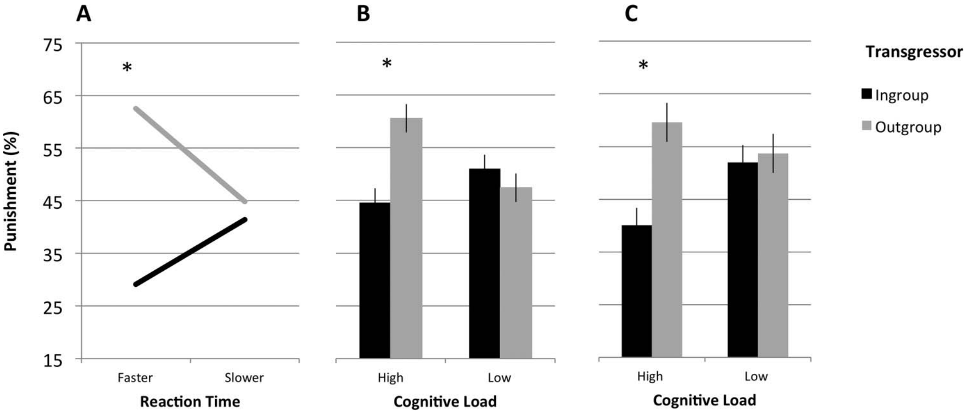


Figure 2. Faster response time (A) and high cognitive load (B and C) predict increased intergroup bias in third-party punishment. Panel A ($n = 87$) depicts predicted punishment percentages of in-group and out-group National Football League fans at 1 standard deviation above and below the RT mean. Although participants with faster-than-average decision times deducted a significantly greater amount of money from out-group members than in-group members, those with slower-than-average decision times demonstrated no such discrepancy. Panels B ($n = 386$) and C ($n = 213$) depict a similar pattern across high and low cognitive load in Indian nationals and American baseball fans, respectively. Error bars reflect ±1 standard error of the mean; $^* p < .01$.

used baseball teams as our group manipulation in order to examine if our results generalize across group contexts.

Participants

On the basis of the results of Experiment 2, we anticipated a small effect size (Cohen's f = .14, power [1 − β] = 80%) and recruited 400 American participants through MTurk to participate in a short study titled "You Win Some, You Lose Some" in exchange for $0.30. Of these, 14 failed an attention check or to complete the study, leaving a total of n = 386 (133 females, M_{age} = 33.9).

Materials and Procedure

The design of this experiment was largely the same as that of Experiment 2. Participants were asked to memorize either the sequence "7T4$RF%" (high load) or "74" (low load) and were told they would be prompted to report this sequence later in the experiment. To ensure the cognitive load manipulation was successful, and consistent with past research in this domain (Paas, Tuovinen, Tabbers, & Van Gerven, 2003), at the conclusion of the study we asked participants to indicate on a 9-point scale how much mental effort they had invested in remembering the code (*very, very low mental effort* to *very, very high mental effort*), and asked whether they had written down the code.

The amount of money participants saw "stolen" from Player B was $0.22 out of $.30. Group identity was varied via favorite baseball team; The game was titled "Baseball Heroes." To ensure random assignment in both in-group and out-group conditions, we divided the 30 American major league teams into three bins of 10 teams each. Depending on the bin containing the team the participant had chosen, we assigned out-group members to consist of a team randomly selected from the alternate bin. We were thus able to ensure that team relationships were fully varied among all potential team combinations.

Results and Discussion

Manipulation check. As expected, participants in the high load condition (M = 5.45, SD = 2.06) reported investing significantly more mental effort than those in the low load condition (M = 2.85, SD = 2.11), t(383) = −12.19, p < .001, **d** = 1.25. Furthermore, as predicted, analysis of log-transformed RTs suggested that those in the high load condition took significantly more time to respond to the punishment measure than those in the low condition, t(371) = −2.05, p = .041, **d** = .21.[1] Together, these results support the notion that the cognitive load manipulation was effective in influencing deliberative processing.

Effect of group and load on punishment. To test the effects of reflexive judgment on third-party punishment decisions, we conducted a 2 (*transgressor identity*: in-group vs. out-group) × 2 (*cognitive load*: high vs. low) ANOVA, with money deducted as the dependent variable. Results showed a significant interaction between load condition and transgressor identity, F(1, 381) = 5.71, p = .017, **f** = .12 (see Figure 2C). To better understand this interaction, we tested the simple effects of transgressor group membership in each load condition. Those in the low load condition demonstrated no difference in punishing out-group transgressors (M = 53.7%, SD = 33.9) compared to in-group transgressors (M = 52.0%, SD = 37.2), F(1, 381) = .116, p = .73. In contrast, those in the high load condition demonstrated significant differences in punishment according to transgressor identification, with out-group members (M = 59.6%, SD = 33.7) punished more harshly than in-group members (M = 40.1%, SD = 36.2), F(1, 381) = 14.05, p < .001, **f** = .19. Consistent with the *reflexive intergroup bias hypothesis*, participants demonstrated more intergroup bias when they were under cognitive load.

As in Experiment 2, there remained the question as to the efficacy of the cognitive load manipulation, because participants could have written down the code. Once again, we reanalyzed the data using just those participants who said they had not written down the code (89.2%). The results held, with participants under load punishing transgressors more or less depending on their group membership, F(1, 336) = 6.59, p = .011, f = .13. Furthermore, none of the other reported statistics changed from significant to nonsignificant (or vice versa). The data are therefore consistent with reflexive intergroup bias in punishment.

General Discussion

The purpose of this research was to examine the effects of reflexive judgment on third-party punishment of in-group and out-group members. Past research suggests that intergroup bias in punishment may stem from reflexive features of the moral mind. Here we tested this *reflexive intergroup bias hypothesis* directly by examining the effects of reflexive judgment on people's punishment of in-group and out-group transgressors. We operationalized reflexive judgment first by measuring RT and then by manipulating cognitive load. Across three experiments, with three different manipulations of group identity, reflexive judgment reliably elicited increased intergroup bias in third-party punishment. However, we found no evidence that people become egalitarian (the *reflexive egalitarian hypothesis*) or more punitive toward in-group members (the *reflexive group regulation hypothesis*) under conditions of reflexive judgment. As such, our results are consistent with the reflexive intergroup bias hypothesis.

This finding is in line with a long history in social psychology of intergroup discrimination and bias (Hewstone, Rubin, & Willis, 2002; Tajfel & Turner, 1979). People are known to form groups based on minimal information, and to use such group boundaries as a basis for hostility and aggression. Such reactions are often exacerbated when people are forced to rely on implicit attitudes and judgments (Duckitt, Wagner, Du Plessis, & Birum, 2002; Greenwald et al., 2002). For instance, when in the presence of a racial out-group member, stereotypes are automatically activated and people tend to construe and evaluate ambiguous behavior in line with existing stereotype (e.g., Devine, 1989). Such differences in reflexive evaluations can lead to behavioral demonstrations of out-group antagonism, such as in the shooter task, in which participants make quick judgments about whether to shoot black or white criminals and civilians (Correll, Park, Judd, & Wittenbrink, 2002). Furthermore, cognitive load increases people's discrepancies in cooperation with in-group versus out-group members (De Dreu et al., 2015; see also Liu, He, & Dou, 2015; Wang et al.,

[1] These results remain significant when using the same 95th-percentile method of eliminating outliers used in Experiment 2.

2011). Our research extends these findings in the domain of third-party punishment, showing that even when punishers are unrelated to the victim of the transgression, they are more willing to punish out-group members more harshly than in-group members when punishing reflexively.

The research is further consistent with anthropological evidence suggesting that humans evolved in small tribal communities in competition for scarce resources (Lancaster, Kaplan, Hill, & Hurtado, 2000). Because intergroup relations were often reduced to hostile encounters in which out-group loss was directly related to in-group gain, aggression was a frequent feature of intergroup interaction (Choi & Bowles, 2007). Because punishment can be used to bolster the in-group's relative standing over rival out-groups (Jackson, 1993), it may have served as a natural extension of these well-established biases between groups (Baumgartner et al., 2012; Hewstone, Rubin, & Willis, 2002; Tajfel & Turner, 1979). Furthermore, recent theoretical advances suggest that third-party punishment originally arose in situations in which witnesses to transgressions were forced to take sides in the dispute and sought to form alliances with others in order to mitigate the risk of retaliation (Descioli & Kurzban, 2009, 2013). Our research underscores the notion that group membership serves as a powerful boundary along which such coalitions may be established (Baumgartner et al., 2012; Tajfel & Turner, 1979; Van Bavel, Packer, & Cunningham, 2008; Yamagishi & Kiyonari, 2000). Reflexive intergroup bias in third-party punishment, in other words, may stem from people's natural tendency to see moral issues through the lens of group membership (Xiao, Coppin, & Van Bavel, 2016).

The fact that intergroup biases are exacerbated under reflexive judgment raises the question about the evolutionary and adaptive origins of such behavior. Reflexive and deliberative judgments are known to arise from discrete areas of the brain (Cunningham, Zelazo, Packer, & Van Bavel, 2007; Kuo, Sjöström, Chen, Wang, & Huang, 2009; Pochon et al., 2002), with the former driven by evolutionarily ancient subcortical systems such as the amygdala (Van Dillen, Heslenfeld, & Koole, 2009) and the latter by relatively recent-emerging areas of the frontal cortex (Coolidge & Wynn, 2001; D'Esposito, Postle, & Rypma, 2000). The fact that intergroup bias in third-party punishment is exacerbated by reflexive judgment thus provides convergent evidence that this behavior may have emerged early in human evolutionary development as the result of selection pressures stemming from frequent intergroup competition (Bernhard et al., 2006). By the same token, the fact that deliberation seems to counteract these biases suggests that the egalitarian motives that presumably underlie it are relatively recent additions to social cognition and may involve more elaborate, universalist moral reasoning (Kohlberg, 1969).

Some commentators have wisely cautioned against drawing strong evolutionary conclusions from manipulations of cognitive load (Barrett et al., 2006; Barrett & Kurzban, 2006, 2012). The fact that a behavior arises from reflexive judgment does not necessarily mean that it evolved early in human history. Reflexive judgments may instead be the product of learned behaviors (e.g., stomping instinctively on a car brake). By the same token, deliberative processes may have been influenced by evolutionary forces. For instance, some scientists have argued that evolutionary pressures for group living have influenced cortical growth in humans (Dunbar, 1998). As a result, the fact that cognitive load exacerbates a behavior cannot be used to conclusively demonstrate that it is evolutionarily ancient. Instead, these inferences should converge with the evolutionary record, nonhuman primates, and developmental samples. Ultimately, however, although we cannot definitively rule out the possibility that the intergroup bias that emerges under cognitive load is the product of learned as opposed to evolved behaviors, our data add to a growing body of evidence suggesting that intergroup bias in punishment is rooted in reflexive features of the human mind (Baumgartner et al., 2012; Bernhard et al., 2006; De Dreu et al., 2015; Jordan, McAuliffe, & Warneken, 2014).

Implications for Theories of Group Regulation

The current work speaks to group-regulatory models of third-party punishment, which suggest that people should be most punitive against members of their own group to ward off in-group defection and bolster group reputation (Fearon & Laitin, 1996; Fehr & Gächter, 2002; Marques, Yzerbyt, & Leyens, 1988; Mendoza, Lane, & Amodio, 2014; Otten & Gordijn, 2014; van Prooijen & Lam, 2007; Schmidt, Rakoczy, & Tomasello, 2012). We found no evidence of such regulatory behavior in our experiments. Reflexive judgment appears to produce behaviors more in line with out-group hostility than with in-group policing. This is consistent with recent literature suggesting that third-party punishment may have emerged for other reasons than promoting cooperation. For instance, Dreber and colleagues (2008) found no evidence that costly punishment increases the overall payoffs of a group and that individuals who are most successful in cooperative games are not the ones that willingly incur costs to punish defectors. Other research has shown that people stand to gain from administering third-party punishment because it ultimately enhances their trustworthiness and reputation (Barclay, 2006; Jordan, Hoffman, Bloom, & Rand, 2016). Thus, group regulatory models may not be the most effective in explaining all aspects of the functionality of third-party punishment.

In-Group Love or Out-Group Hate?

Divergent results in Experiments 2 and 3 raise the question as to whether reflexive judgment is more likely to produce out-group hostility or in-group favoritism. Although reflexive judgment made out-group punishment relatively more severe in Experiment 2, in Experiment 3 it made in-group punishment more lenient. These issues have been discussed in previous research. Brewer (1999), for example, suggested that the majority of intergroup bias is driven by in-group favoritism based on mutual dependence and the need for inclusion and assimilation. Furthermore, perceptions of threat or competitions for limited resources may result in explicit antagonism to out-group members. More recently, Halevy et al., (2011) showed that both "in-group love" and "out-group hate" may motivate intergroup bias, although aggressive tendencies toward out-group members were diminished when people had the opportunity to express in-group love. Other work has found that bias in third-party punishment is driven by both in-group favoritism and out-group antagonism (Schiller, Baumgartner, & Knoch, 2014). Consistent with these past theoretical perspectives, we believe that both factors may be at play here and that small differences in the nature of the group identity may determine which effect emerges predominantly. For example, Indian participants may hold reflexive antagonism toward American outsiders, whereas baseball fans may reflexively favor their own team (consistent with work showing that sports fans frequently favor their own team in controversial referee decisions, Mohr & Larsen, 1998). Our data cannot show conclusively whether intergroup bias in punishment is driven primarily by one or the other; it is most consistent with the possibility that both motives play a role.

Impact and Reliability

The fact that our studies were carried out exclusively in an online environment raises the question as to whether participants found the scenario believable and impactful, and whether the manipulations of cognitive load were effective. First, we note that online experimentation has yielded replications of a wide variety of psychological phenomena (Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013; Hilbig, 2015). Furthermore, a closer examination of the data confirms that participants believed the experimental cover story, were indeed under cognitive load, and thought they were dealing with real interaction partners with genuine monetary consequences. The vast majority of participants in Experiments 2 and 3 said they had not written down the code to be memorized, and the data with just those participants who said they had not recorded the code remained unchanged. Furthermore, the fact that participants might have recorded the code yields an even more conservative test of our hypothesis, because the effect of the cognitive load manipulation would be, if anything, diluted.

To more thoroughly examine the question of whether participants believed they were in a real interaction, we assigned an independent rater blind to the experiment's hypothesis to code participants' free responses concerning the identity of their interaction partners. In Experiment 1, about 4% of participants gave statements suggesting that they believed the interaction partner was not real; in Experiments 2 and 3, less than 1% of participants did so. Furthermore, participants' greetings in the interaction room overwhelmingly reflected the belief that they were interacting with real people (e.g., "Nice to see another Chiefs logo," and "Hey. let's give everything and not take anything . . . I think that would be better for all"). Reanalyzing the data from each of the experiments excluding those participants who expressed the belief that their interaction partner was a computer did not change any of the reported results.

Future Directions

This research opens many avenues for further work on the nature of third-party punishment in intergroup contexts. One of the most important issues that remain to be resolved is what moderating variables influence the extent to which third-party punishment is subject to intergroup bias versus in-group policing. Here are some potential variables that may influence the results.

Group identification. One variable that may influence whether people punish in-group or out-group members more severely is the extent to which they are identified with the group. Indeed, some research (Mendoza, Lane, & Amodio, 2014) found that high group identifiers punish in-group deviants more harshly—a finding consistent with work on the black sheep effect (Marques, Yzerbyt, & Leyens, 1988). We measured in-group identification in all three experiments, but found that, in our data at least, this factor had no moderating effect on people's punishment decisions (all $ps > .3$). In other words, those who were strongly identified with the group showed just as much reflexive intergroup bias as those who were weakly identified. It is therefore possible that intergroup bias in punishment may hold even when identification with the group is weak.

Victim identity. Another important moderator that may guide whether intergroup bias or group regulation motivates third-party punishment is victim identity. In this research, the victim was always a member of an unrelated third party. If the victim were a member of the punisher's own group, more stringent punishment of in-group transgressors may emerge, consistent with the notion that the betrayal of one's own kin is among the worst forms of treachery (Alighieri, 1973; Bernhard et al., 2006; Goette et al., 2010).

Repeated games. Whereas the current research examined people's punishment behavior in one-shot dilemmas, research suggests that people's behavior may change under conditions of repeated interaction (e.g., Bó, 2005; Denant-Boemont, Masclet, & Noussair, 2007). When people know they will have repeated interactions with certain individuals, they may be more willing to punish in-group deviance to serve a pedagogical goal. Thus, it will be important to continue to investigate third-party punishment in repeated interactions. Furthermore, research has shown that reputational concerns impact people's willingness to sustain pro-social behavior (Jordan, Hoffman, Bloom & Rand, 2016; Kurzban, DeScioli, & O'Brien, 2007; Milinski, Semmann, & Krambeck, 2002). Thus, a fruitful line of investigation might consider the extent to which such reputation spreads among members of a group, and the effect this has on decisions to sanction in-group and out-group transgressors.

Costly punishment. The current research used a procedure in which punishment was costless. Thus, an open question is the extent to which these findings can be applied to instances of costly (or altruistic) punishment. Some work suggests there may be theoretical and physiological distinctions between these behaviors; for instance, costly punishment draws more heavily on brain regions such as the ventromedial prefrontal cortex and medial orbitofrontal cortex (de Quervain et al., 2004). However, there is also extensive work examining costly punishment behavior on its own (e.g., Buckholtz et al., 2008; Carlsmith, Darley, Robinson, 2002; De Castella, Platow, Wenzel, Okimoto, & Feather, 2011; Roos, Gelfand, Nau, & Carr, 2014; Whitson, Wang, See, Baker, & Murnighan, 2015), demonstrating the theoretical merit of examining this behavior. Indeed, many instances of punishment behavior, including those administered by juries, are costless. Overall, although we hypothesize that these effects would hold in conditions of costly punishment, we leave this question open to future research.

Leadership. Another factor determining the ultimate impact of punishment decisions is the perceived role of the punisher in the group (Abrams, Randsley de Moura, & Travaglino, 2013; Podsakoff, Todor, Grover, & Huber, 1984). Punishers who feel themselves to be in leadership roles may feel an added responsibility for regulating the behavior of group members, and so may exhibit greater tendencies toward in-group policing than average group members. Likewise, lower-status group members may wish to police other members as a way of demonstrating commitment to the group and shoring up group loyalty. Thus, the issue of leadership on punishment decisions demands further investigation.

Practical Implications

Our findings may have important implications in several policy domains. One such case is in the legal domain. Research has shown that out-group members are often subject to more stringent punishments, even in cases involving serious punishment. For instance, faces that are stereotypically "blacker" are more likely to receive death sentences in capital punishment cases (Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006). Moreover, all-white jurors are significantly more likely to punish black, but not white, defendants (Anwar, Bayer, Hjalmarsson, 2011). Our research may provide at least one explanation as to why this is the case. Often cases are subject to considerable complexity, putting jurors under cognitive load and eliciting more reflexive punishment decisions. This notion that justice decisions may be influenced in a nonoptimal way by extraneous factors is consistent with well-documented phenomena in courtroom decisions (Danziger, Levav, & Avnaim-Pesso, 2011). Our research suggested that group manipulations as trivial as sports fans can elicit discrepancies in the administration of justice; thus, extra effort must be made to avoid such biases in the courtroom. If authorities such as judges are made aware of the aversive effects of cognitive load on intergroup biases, they may be able to circumvent or at least mitigate such biases by encouraging jurors to be aware of these tendencies and head them off before they shape judgment.

Conclusions

The human capacity for third-party punishment is essential for long-term cooperation, and cooperation in turn promotes cultural advancement and technological progress. Although punishment may be used to regulate the members of the in-group, the evidence presented here suggests that it is often driven by intergroup bias—leading to harsher punishment of out-group members. Although humans value fairness and equity, under conditions of reflexive judgment they use punishment as a tool to exacerbate intergroup differences. Thus, third-party punishment can be used to promote cooperative behavior as well as hostility toward the "other."

References

Abrams, D., Randsley de Moura, G., & Travaglino, G. A. (2013). A double standard when group members behave badly: Transgression credit to ingroup leaders. *Journal of Personality and Social Psychology, 105,* 799–815. http://dx.doi.org/10.1037/a0033600

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Atlanta, GA: Sage.

Alighieri, D. (1973). *The divine comedy*. New York, NY: Grolier.

Anwar, S., Bayer, P., & Hjalmarsson, R. (2012). The impact of jury race in criminal trials. *The Quarterly Journal of Economics, 127,* 1017–1055. http://dx.doi.org/10.1093/qje/qjs014

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin, 140,* 1556–1581. http://dx.doi.org/10.1037/a0037737

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27,* 325–344. http://dx.doi.org/10.1016/j.evolhumbehav.2006.01.003

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54,* 462–479. http://dx.doi.org/10.1037/0003-066X.54.7.462

Barrett, H. C., Frederick, D. A., Haselton, M. G., & Kurzban, R. (2006). Can manipulations of cognitive load be used to test evolutionary hypotheses? *Journal of Personality and Social Psychology, 91,* 513–518. http://dx.doi.org/10.1037/0022-3514.91.3.513

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review, 113,* 628–647.

Barrett, H. C., & Kurzban, R. (2012). What are the functions of System 2 modules? A reply to Chiappe and Gardner. *Theory & Psychology, 22,* 683–688. http://dx.doi.org/10.1177/0959354312455469

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 570–585. http://dx.doi.org/10.1037/0278-7393.33.3.570

Baumgartner, T., Götte, L., Gügler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping, 33,* 1452–1469. http://dx.doi.org/10.1002/hbm.21298

Bernhard, H., Fehr, E., & Fischbacher, U. (2006). Group affiliation and altruistic norm enforcement. *The American Economic Review, 96,* 217–221. http://dx.doi.org/10.1257/000282806777212594

Bó, P. D. (2005). Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. *The American Economic Review, 95,* 1591–1604. http://dx.doi.org/10.1257/000282805775014434

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology & Sociobiology, 13,* 171–195. http://dx.doi.org/10.1016/0162-3095(92)90032-Y

Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions Royal Society B: Biological Sciences, 364,* 3281–3288. http://dx.doi.org/10.1098/rstb.2009.0134

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues, 55,* 429–444. http://dx.doi.org/10.1111/0022-4537.00126

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron, 60,* 930–940. http://dx.doi.org/10.1016/j.neuron.2008.10.016

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. http://dx.doi.org/10.1177/1745691610393980

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83,* 284–299. http://dx.doi.org/10.1037/0022-3514.83.2.284

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York, NY: Guilford Press.

Choi, J.-K., & Bowles, S. (2007). The Coevolution of Parochial Altruism and War. *Science, 318,* 636–640. http://dx.doi.org/10.1126/science.1144237

Coolidge, F. L., & Wynn, T. (2001). Executive functions of the frontal lobes and the evolutionary ascendancy of *Homo sapiens*. *Cambridge Archaeological Journal, 11,* 255–260. http://dx.doi.org/10.1017/S0959774301000142

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83,* 1314–1329. http://dx.doi.org/10.1037/0022-3514.83.6.1314

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8,* e57410. http://dx.doi.org/10.1371/journal.pone.0057410

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition, 25,* 736–760. http://dx.doi.org/10.1521/soco.2007.25.5.736

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17,* 273–292. http://dx.doi.org/10.1177/1088868313495594

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 6889–6892. http://dx.doi.org/10.1073/pnas.1018033108

De Castella, K., Platow, M. J., Wenzel, M., Okimoto, T., & Feather, N. T. (2011). Retribution or restoration? Anglo–Australian's views towards domestic violence involving Muslim and Anglo–Australian victims and offenders. *Psychology, Crime & Law, 17,* 403–420. http://dx.doi.org/10.1080/10683160903292253

De Dreu, C. K., Dussel, D. B., & ten Velden, F. S. (2015, May 6). In intergroup conflict, self-sacrifice is stronger among pro-social individuals, and parochial altruism emerges especially among cognitively taxed individuals. *Frontiers in Psychology*.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory, 33,* 145–167. http://dx.doi.org/10.1007/s00199-007-0212-0

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

DeScioli, P., & Kurzban, R. (2013). A Solution to the Mysteries of Morality. *Psychological Bulletin, 139,* 477–496. http://dx.doi.org/10.1037/a0029065

D'Esposito, M., Postle, B. R., & Rypma, B. (2000). Prefrontal cortical contributions to working memory: Evidence from event-related fMRI studies. *Experimental Brain Research, 133,* 3–11. http://dx.doi.org/10.1007/s002210000395

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science, 305,* 1254–1258. http://dx.doi.org/10.1126/science.1100735

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5–18. http://dx.doi.org/10.1037/0022-3514.56.1.5

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452,* 348–351. http://dx.doi.org/10.1038/nature06723

Duckitt, J., Wagner, C., Du Plessis, I., & Birum, I. (2002). The psychological bases of ideology and prejudice: Testing a dual process model. *Journal of Personality and Social Psychology, 83,* 75–93.

Dunbar, R. I. (1998). The social brain hypothesis. *Foundations in Social Neuroscience, 5,* 178–190.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science, 17,* 383–386. http://dx.doi.org/10.1111/j.1467-9280.2006.01716.x

Efferson, C., Lalive, R., & Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science, 321,* 1844–1849. http://dx.doi.org/10.1126/science.1155805

Evans, A. M., Dillon, K. D., & Rand, D. G. (2015). Fast but not intuitive, slow but not reflective: Decision conflict drives reaction times in social dilemmas. *Journal of Experimental Psychology: General, 144,* 951–966. http://dx.doi.org/10.1037/xge0000107

Fearon, J. D., & Laitin, D. D. (1996). Explaining interethnic cooperation. *The American Political Science Review, 90,* 715–735. http://dx.doi.org/10.2307/2945838

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences, 8,* 185–190. http://dx.doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *The Journal of Economic Perspectives, 14,* 159–181. http://dx.doi.org/10.1257/jep.14.3.159

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415,* 137–140. http://dx.doi.org/10.1038/415137a

Fowler, J. H. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 7047–7049. http://dx.doi.org/10.1073/pnas.0500938102

Gardner, A., & West, S. A. (2004). Cooperation and punishment, especially in humans. *American Naturalist, 164,* 753–764. http://dx.doi.org/10.1086/425623

Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology, 57,* 940–949. http://dx.doi.org/10.1037/0022-3514.57.6.940

Goette, L., Huffman, D., Meier, S., & Sutter, M. (2010). Group membership, competition, and altruistic versus antisocial punishment: Evidence from randomly assigned army groups. Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1682710

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences, 11,* 322–323. http://dx.doi.org/10.1016/j.tics.2007.06.004

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York: Penguin Press.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6,* 517–523. http://dx.doi.org/10.1016/S1364-6613(02)02011-9

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107,* 1144–1154. http://dx.doi.org/10.1016/j.cognition.2007.11.004

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109,* 3–25. http://dx.doi.org/10.1037/0033-295X.109.1.3

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–1480. http://dx.doi.org/10.1037/0022-3514.74.6.1464

Halevy, N., Weisel, O., & Bornstein, G. (2011). "In-group love" and "out-group hate" in repeated interaction between groups. *Journal of Behavioral Decision Making, 25,* 188–195. http://dx.doi.org/10.1002/bdm.726

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 19931–19936. http://dx.doi.org/10.1073/pnas.1110306108

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2006). Costly punishment across human societies. *Science, 312,* 1767–1770.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology, 53,* 575–604. http://dx.doi.org/10.1146/annurev.psych.53.100901.135109

Hilbig, B. E. (2015). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*. http://dx.doi.org/10.3758/s13428-015-0678-9

Jackson, J. W. (1993). Realistic group conflict theory: A review and evaluation of the theoretical and empirical literature. *The Psychological Record, 43,* 391–419.

Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences of the*

United States of America, 104, 13046–13050. http://dx.doi.org/10.1073/pnas.0705555104

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530,* 473–476. http://dx.doi.org/10.1038/nature16981

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 12710–12715. http://dx.doi.org/10.1073/pnas.1402280111

Kahneman, D. (2011). Thinking, fast and slow. New York, NY: Macmillan.

King, R. D., & Wheelock, D. (2007). Group threat and social control: Race, perceptions of minorities and the desire to punish. *Social Forces, 85,* 1255–1280. http://dx.doi.org/10.1353/sof.2007.0045

Kohlberg, L. (1969). Stage and sequence: The cognitive–developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). Chicago, IL: Rand McNally.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications, 6,* 7455. http://dx.doi.org/10.1038/ncomms8455

Kuo, W. J., Sjöström, T., Chen, Y. P., Wang, Y. H., & Huang, C. Y. (2009). Intuition and deliberation: Two systems for strategizing in the brain. *Science, 324,* 519–522. http://dx.doi.org/10.1126/science.1165598

Kurzban, R., DeScioli, P., & Obrien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28,* 75–84. http://dx.doi.org/10.1016/j.evolhumbehav.2006.06.001

Lancaster, J. B., Kaplan, H. S., Hill, K., & Hurtado, A. M. (2000). The evolution of life history, intelligence and diet among chimpanzees and human foragers. In *Perspectives in Ethology Series: Volume 13. Perspectives in ethology* (pp. 47–72). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4615-1221-9_2

Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology, 5,* 289–305. http://dx.doi.org/10.1177/147470490700500203

Liu, Y., He, N., & Dou, K. (2015). Ego-depletion promotes altruistic punishment. *Open Journal of Social Sciences, 3,* 62–69. http://dx.doi.org/10.4236/jss.2015.311009

Makimura, Y., & Yamagishi, T. (2003). Ongoing group interaction, ingroup favoritism, and reward allocation. *Shinrigaku Kenkyu*. The Japanese Journal of Psychology, *73,* 488–493.

Manoach, D. S., Schlaug, G., & Siewert, B. (1997). Prefrontal cortex fMRI signal changes are correlated with working memory load. *NeuroReport, 8,* 545–549.

Marques, J. M., Yzerbyt, V. Y., & Leyens, J. P. (1988). The "black sheep effect": Extremity of judgments towards ingroup members as a function of group identification. *European Journal of Social Psychology, 18,* 1–16. http://dx.doi.org/10.1002/ejsp.2420180102

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition, 134,* 1–10. http://dx.doi.org/10.1016/j.cognition.2014.08.013

Mendoza, S. A., Lane, S. P., & Amodio, D. M. (2014). For members only Ingroup punishment of fairness norm violations in the ultimatum game. *Social Psychological and Personality Science, 5,* 662–670. http://dx.doi.org/10.1177/1948550614527115

Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the "tragedy of the commons." *Nature, 415,* 424–426. http://dx.doi.org/10.1038/415424a

Mohr, P. B., & Larsen, K. (1998). Ingroup favoritism in umpiring decisions in Australian football. *The Journal of Social Psychology, 138,* 495–504. http://dx.doi.org/10.1080/00224549809600403

Mussweiler, T., & Ockenfels, A. (2013). Similarity increases altruistic punishment in humans. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 19318–19323. http://dx.doi.org/10.1073/pnas.1215443110

Otten, S., & Gordijn, E. H. (2014). Was it one of us? How people cope with misconduct by fellow in-group members. *Social and Personality Psychology Compass, 8,* 165–177.

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38,* 63–71. http://dx.doi.org/10.1207/S15326985EP3801_8

Pochon, J. B., Levy, R., Fossati, P., Lehericy, S., Poline, J. B., Pillon, B., . . . Dubois, B. (2002). The neural system that bridges reward and cognition in humans: An fMRI study. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 5669–5674. http://dx.doi.org/10.1073/pnas.082111099

Podsakoff, P. M., Todor, W. D., Grover, R. A., & Huber, V. L. (1984). Situational moderators of leader reward and punishment behaviors: Fact or fiction? *Organizational Behavior & Human Performance, 34,* 21–63. http://dx.doi.org/10.1016/0030-5073(84)90036-9

Raihani, N. J., Grutter, A. S., & Bshary, R. (2010). Punishers benefit from third-party punishment in fish. *Science, 327,* 171–171. http://dx.doi.org/10.1126/science.1183068

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489,* 427–430. http://dx.doi.org/10.1038/nature11467

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532. http://dx.doi.org/10.1037/0033-2909.114.3.510

Richerson, P. J., & Boyd, R. (2001). The evolution of subjective commitment to groups: A tribal instincts hypothesis. In R. M. Nesse (Ed.), *Evolution and the capacity for commitment* (pp. 186–220). New York, NY: Russell Sage Foundation.

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 14824–14829. http://dx.doi.org/10.1073/pnas.1203179109

Roos, P., Gelfand, M., Nau, D., & Carr, R. (2014). High strength-of-ties and low mobility enable the evolution of third-party punishment. *Proceedings of the Royal Society B: Biological Sciences, 281*(1776), 20132661. http://dx.doi.org/10.1098/rspb.2013.2661

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior, 35,* 169–175. http://dx.doi.org/10.1016/j.evolhumbehav.2013.12.006

Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition, 124,* 325–333. http://dx.doi.org/10.1016/j.cognition.2012.06.004

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33–47). Monterey, CA: Brooks-Cole.

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology, 44,* 1334–1338. http://dx.doi.org/10.1016/j.jesp.2008.03.010

Van Bavel, J. J., & Cunningham, W. A. (2012). A social identity approach to person memory: Group membership, collective identification, and social role shape attention and memory. *Personality and Social Psychology Bulletin, 38,* 1566–1578. http://dx.doi.org/10.1177/0146167212455829

Van Bavel, J. J., FeldmanHall, O., & Mende-Siedlecki, P. (2015). The neuroscience of moral cognition: From dual processes to dynamic systems. *Current Opinion in Psychology, 6,* 167–172. http://dx.doi.org/10.1016/j.copsyc.2015.08.009

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias: A functional magnetic resonance imaging investigation. *Psychological Science, 19,* 1131–1139. http://dx.doi.org/10.1111/j.1467-9280.2008.02214.x

Van Dillen, L. F., Heslenfeld, D. J., & Koole, S. L. (2009). Tuning down the emotional brain: An fMRI study of the effects of cognitive load on the processing of affective images. *NeuroImage, 45,* 1212–1219. http://dx.doi.org/10.1016/j.neuroimage.2009.01.016

van Prooijen, J. W., & Lam, J. (2007). Retributive justice and social categorizations: The perceived fairness of punishment depends on intergroup status. *European Journal of Social Psychology, 37,* 1244–1255. http://dx.doi.org/10.1002/ejsp.421

Wang, C. S., Sivanathan, N., Narayanan, J., Ganegoda, D. B., Bauer, M., Bodenhausen, G. V., & Murnighan, K. (2011). Retribution and emotional regulation: The effects of time delay in angry economic interactions. *Organizational Behavior and Human Decision Processes, 116,* 46–54. http://dx.doi.org/10.1016/j.obhdp.2011.05.007

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record, 58,* 475–482.

Whitson, J., Wang, C. S., See, Y. H. M., Baker, W. E., & Murnighan, J. K. (2015). How, when, and why recipients and observers reward good deeds and punish bad deeds. *Organizational Behavior and Human Decision Processes, 128,* 84–95. http://dx.doi.org/10.1016/j.obhdp.2015.03.006

Wrangham, R. W., & Peterson, D. (1997). *Demonic males: Apes and the origins of human violence*. Boston, MA: Houghton Mifflin Harcourt.

Xiao, Y. J., Coppin, G., & Van Bavel, J. J. (2016). Seeing the world through group-colored glasses: A perceptual model of intergroup relations. *Psychological Inquiry*.

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded generalized reciprocity: Ingroup boasting and ingroup favoritism. *Advances in Group Processes, 16,* 161–197.

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly, 63,* 116–132. http://dx.doi.org/10.2307/2695887

Received August 12, 2015
Revision received May 10, 2016
Accepted May 13, 2016 ■