Journal of Experimental Social Psychology 84 (2019) 103823

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Escalation of negative social exchange: Reflexive punishment or deliberative deterrence?

James Vandermeer, Christine Hosey, Nicholas Epley*, Boaz Keysar

University of Chicago, United States of America

ARTICLE INFO

This paper has been recommended for acceptance by Rachel Barkan

ABSTRACT

Negative escalation of social exchange exacts significant costs on both individuals and society. Instead of in-kind reciprocity—an eye for an eye—negative reciprocity may escalate, taking two eyes for an eye. We tested two competing mechanisms for negative escalation using a modified dictator game that reliably produces escalating reciprocity to others' negative actions but not to positive actions. According to one mechanism, escalation is strategic: a deliberate attempt to deter future harm. According to another mechanism, escalation is reflexive: an impulsive act of retribution without consideration of future consequences. In seven experiments, we find clear evidence consistent with a reflexive mechanism. Encouraging deliberation reduced negative escalation in one-shot interactions as well as in repeated interactions. Focusing on future consequences decreased escalation whereas disabling deliberation increased escalation. Finally, the explicit goal to punish another's negative behavior increased escalation while the goal of deterring future transgressions did not. These experiments suggest that escalation is a reflexive form of punishment rather than a deliberate act of strategic deterrence. Encouraging, enabling, or training deliberative processes may effectively reduce costly escalation in everyday life.

1. Reflexive escalation in social exchange

Cooperative social exchange is central to a well-functioning society. Around the globe, positive exchanges are guided by a norm of reciprocity whereby kind words and good deeds are reciprocated in kind measure, building trust and strengthening cooperation (Cialdini, 2001; Fehr, Fischbacher, & Gachter, 2002; Gouldner, 1960). Negative exchanges, in contrast, tend to escalate (Anderson, Buckley, & Carnagey, 2008). Insult and injury are not merely repaid, they are repaid with interest; instead of an eye for an eye, escalation demands two eyes for an eye (Keysar, Converse, Wang, & Epley, 2008). Although some retribution may be necessary to maintain cooperative norms, excessive escalation is inefficient and costly for individuals and society (Axelrod & Hamilton, 1981; Tedeschi & Felson, 1994). Understanding how to avoid costly escalation in everyday life requires understanding the psychological processes that produce it.

Whether involving the redistribution of material quantities or abstract qualities like esteem and credibility, social exchange engages considerations of fairness and cooperation. Individuals exhibit striking variation in these areas, and existing literature has elucidated a range of related issues. For instance, cooperation is moderated by state emotions (e.g., Fabiansson & Denson, 2012; Treadway et al., 2014) as well as trait anxiety (Grecucci et al., 2013), and is sensitive to how the context of the exchange is construed (Liberman, Samuels, & Ross, 2004; Rand, 2016). Selfishness and altruism also tend to vary situationally (Winking & Mizer, 2013), though a study of wealthy Tokyo suburbanites found that 7% of individuals had a deeply engrained proclivity for selfishness (Yamagishi, Li, Takagishi, Matsumoto, & Kiyonari, 2014). Research on the punishment of greed and of free riders has explored the frequency and impact of third-party punishment (e.g., Balafoutas, Grechenig, & Nikiforakis, 2014; Crockett, Özdemir, & Fehr, 2014; Denant-Boemont, Masclet, & Noussair, 2007; Nikiforakis, 2008), how punishment affects cooperation (Dreber, Rand, Fudenberg, & Nowak, 2008), and the role of reputation in moderating punishment's effectiveness (dos Santos, Rankin, & Wedekind, 2011). While we have learned a great deal about cooperative behavior and the role of punishment, we still know relatively little about the specific mechanisms that cause negative social exchange to escalate.

Here we do not investigate cooperation or reciprocity per se. Instead, we focus on negative escalation in social exchange. Escalation occurs when people respond to another person's negative act with an even stronger negative act. Escalation is therefore not simply a negative or uncooperative response to another person's action such as refusing to engage, but rather an increase in the negativity of the response to

* Corresponding author.
E-mail address: epley@chicagobooth.edu (N. Epley).

https://doi.org/10.1016/j.jesp.2019.103823
Received 9 March 2019; Received in revised form 3 June 2019; Accepted 10 June 2019

another's action, such as taking two eyes for an eye. We test between two competing, and intuitively plausible, theoretical mechanisms by studying an experimental setting that reliably produces negative escalation.

According to one theory, negative escalation is deliberate and strategic, a reasoned attempt to deter future negative actions (Diekmann, 2004; Gould, 2000). Specifically, negative escalation is a signal that transgressions will be overly costly to any would-be transgressors. A person therefore retaliates harshly in a strategic and reasoned attempt to avoid being harmed again in the future. Rather than provoking further escalation, harsh retaliation is meant to evoke a cost benefit consideration in the offending party that dis-incentivizes future harm. Deliberate thinking about such future consequences takes both time and mental resources. This theory predicts that escalation should *increase* as careful deliberation increases.

According to a competing theory, escalation is a result of a reflexive and impulsive act of retribution. As such, reflexive escalation would arise in response to one's perception of harm with little consideration of future consequences. This assumes that reflexive reactions occur quickly and relatively thoughtlessly, requiring little time or careful thought to initiate. This theory predicts that escalation should *decrease* as deliberation increases.

Either a strategic or a reflexive mechanism could explain escalation. Because negative escalation is difficult to examine experimentally, existing research informs related phenomena but has not examined negative escalation directly nor does it provide a clear explanation for it. For instance, research on the consequences of deliberation has often focused on the tendency to be cooperative or selfish. In some experiments, procedures that promote deliberation reduce cooperation (e.g., Everett, Ingbretsen, Cushman, & Cikara, 2017) and can lead people to act more selfishly or unethically towards others (e.g., Bushman, 2002; Cappelletti, Güth, & Ploner, 2011; Epley, Caruso, & Bazerman, 2006; Ferguson, Maltby, Bibby, & Lawrence, 2014; Rand, Greene, & Nowak, 2012). In other experiments, however, encouraging deliberation decreases people's willingness to act on selfish impulsive urges and increases normatively appropriate behavior (Gunia, Wang, Huang, Wang, & Murnighan, 2012; Shalvi, Eldar, & Bereby-Meyer, 2012; Wolf et al., 2009). Extrapolating from cooperation to escalation, these results are equivocal about the effect of deliberation on negative escalation. The former suggests that deliberation may increase negative escalation whereas the latter suggests the opposite.

Research more closely related to escalation suggests that emotions such as spite and anger guide negative reactions towards others, such as rejecting unfair offers in an ultimatum game (Ben-Shakhar, Bornstein, Hopfensitz, & van Winden, 2007; Fabiansson & Denson, 2012; Pillutla & Murnighan, 1996). Unfair offers are also rejected more often when people are asked to respond quickly and intuitively than when they deliberate more carefully (Grimm & Mengel, 2011; Halali, Bereby-Meyer, & Meiran, 2011; Kirk, Gollwitzer, & Carnevale, 2011; Neo, Yu, Weber, & Gonzalez, 2013; Smith & Silberberg, 2010; cf. Ferguson et al., 2014). This research does not, however, examine whether explicit deliberation would defuse or inflame spontaneous emotional reactions.

These competing theories make distinct predictions on the impact of deliberation on negative escalation. If negative escalation is deliberate and strategic, then escalation should occur when people have the time, mental capacity, and ability to deliberate on another's harmful action and to consider strategic consequences. Negative escalation would then be reduced when people are unable to deliberate about future consequences. In contrast, if negative escalation is reflexive, then it should arise when people respond impulsively, without the time or ability to deliberate about strategic consequences. Reflexive escalation would be guided by impulsive retribution. Negative escalation would then be reduced when people deliberately reflect on the exchange.

Articulating these specific hypotheses is easy, but testing them comprehensively is hard because one cannot study the complexity of escalation as it occurs across varied contexts with severe consequences in everyday life. As a first approximation, we evaluated these competing theories using an experimental procedure that is known to produce escalation in response to negative behavior but not in response to positive behavior. In particular, we examined reactive escalation using a repeated Dictator Game. In this procedure, two players alternate playing the role of a dictator who decides how to divide a pot of money between them. We modified the game to either involve a subjectively positive or negative social exchange (Keysar et al., 2008; List, 2007). In the positive exchange, the pot of money belongs to the "dictator", who decides how much, if any, to *give* the other participant. In the negative exchange, the pot of money belongs to the other participant, and the "dictator" decides how much money to *take* for him or herself, leaving the rest to the other participant. Prior research has demonstrated that people respond differently to identical outcomes in these two exchanges. In particular, when dividing a pot of $100, giving $50 to another person is perceived to be more generous than taking $50 even though the objective outcomes of the decision are identical – $100 is divided equally between two people in both cases (Keysar et al., 2008).

More important, such differential perceptions of generosity seem to guide reciprocity: Participants reciprocate in-kind in the positive exchange, but escalate in the negative exchange (Keysar et al., 2008). For example, a person who had $5 out of $10 given to her in Round 1 would likely respond by giving the other person $5 out of $10 when it was her turn to make the decision in Round 2. In contrast, a person who had $5 out of $10 *taken* from her in Round 1 might respond by taking *more* than $5 from the other person in Round 2. Giving instigated in-kind reciprocity, but taking led to negative escalation across repeated rounds. Although this procedure is uncommon in everyday life, we believe it captures the essential psychological elements involved in reciprocal exchanges and is therefore high in psychological realism (Mook, 1983).

This procedure also enables precise measurement of escalation in response to another's action. We therefore believe it serves as a useful theoretical proxy for testing between competing accounts of negative escalation. To make sure that this experimental context was not systematically biased in favor of one explanation over another, we conducted a brief pilot survey to test the extent to which the theoretical accounts provide an intuitive account of escalation in this paradigm. It would be problematic if we just happened to select a context that was obviously biased in favor of one particular mechanism or another. A survey of 179 participants suggests this was not the case. This survey provided participants with a full description of the procedures used in Experiment 1. Specifically, participants learned that Player 1 was given the opportunity to divide $10 between themselves and Player 2. In the giving condition, Player 1 received $10 and then decided how much to give to the other player. In the taking condition, Player 2 received $10 and Player 1 then decided how much to take from the other player. In both cases, Player 2 was left with $4 because of the action of Player 1, who either gave $4 out of $10, or took $6. Participants then predicted how Player 2 would respond when the roles reverse and Player 2 receives a new pot of $10. Participants learned that Player 2 was either instructed to deliberate or not: to take time and choose carefully or to respond within 2 s and "just go with your gut." Participants in our survey predicted that Player 2's reaction would be more selfish, consistent with escalation, in the taking game than the giving game, leaving for the self on average $6.80 as opposed to $6.35, F (1, 193) = 4.165, p = .04, η^2 = 0.02. However, participants did not predict a significant difference in Player 2's response after deliberating (M = $6.70) than when making a quick "gut" response (M = $6.45), F (1, 193) = 1.50, p = .22, η^2 = 0.01. If anything, participants predicted a response more consistent with deliberative and strategic escalation, expecting the most selfish escalation from Player 2 in the taking game after deliberation (M = $6.91). Given that prior research using this paradigm found consistent escalation following a negative response, we believe this paradigm is useful for adjudicating between two theories that appear plausible both theoretically and intuitively.

Most experiments reported in this paper are conducted in person with real financial incentives. This increases the experiment's psychological realism but requires using locally available participants, thereby restricting the representativeness of our samples. With the exception of Experiment 7, we conducted our experiments in a large Midwestern city in the U.S. (Chicago), with samples composed of people recruited from public places on the University of Chicago campus (Experiments 1, 2, and 6), from a large community sample in downtown Chicago (Experiments 3 and 5), or from a University of Chicago participant pool (Experiment 4). These samples are somewhat diverse, including university students, people recruited at a popular tourist attraction, and community members of varying age, ethnicity, and level of education participating in experiments for money. They are, however, also limited in the sense of being overwhelmingly U.S. citizens. We speculate in the General Discussion on how culture may or may not moderate the results we report.

We report all measures, manipulations, and exclusions in all surveys and experiments reported in this paper.

2. Experiment 1: the effect of deliberation on negative escalation

We test between the strategic and reflexive mechanisms of escalation in Experiment 1 by having participants play a repeated dictator game in which people are either taking resources from each other (the taking game) or giving resources to each other (the giving game). Even when the outcomes are objectively identical, participants construe exchanges of taking as more negative and harmful than exchanges of giving (Keysar et al., 2008). Consistent with prior research, we predict escalation only in cases where another person's behavior was construed as harmful or negative, because both strategic and reflexive accounts presume that perceptions of harm are essential for triggering escalation (to deter future harm in the strategic case, and as retributive punishment for harm in the reflexive condition). We therefore predict negative escalation only in the taking game and predict no escalation in any condition of the giving game.

If escalation is strategic, then participants should escalate more when they deliberate in the taking game than when they respond intuitively. If escalation is reflexive, then we would expect the opposite pattern: more escalation in the taking game when participants respond intuitively.

2.1. Method

2.1.1. Participants

Pedestrians around the University of Chicago campus participated in exchange for the money they earned in the experiment. Because the expected effect size was unknown, we began by recruiting 120 participants (30 per condition). Results showed a reliable pattern,[1] and we therefore decided to double our sample size to a target of 240 participants (60 per condition) to maximize power and to ensure that our result was robust. We continued running participants through a regularly scheduled session, yielding 263 participants. Of these, twenty-nine participants were excluded from analysis for the following reasons, stemming mainly from the challenges of collecting data in a field setting: suspecting the other participant was not real (2), failing to follow instructions (10), failing to respond within the deadline in the reflexive condition (10), declining payment (1), having worked on research in the area (2), or due to communication problems over the walkie-talkies (4). This left 234 participants (104 female, 130 male) included in the analyses.

2.1.2. Procedure and design

Participants played a dictator game with what they believed was another participant located elsewhere on campus. The participant and the confederate could not see each other, and the two experimenters communicated the decisions back and forth using walkie-talkies. They used the terms “Player 1” and “Player 2” while communicating with participants to avoid gender effects (e.g. “Player 1 takes $6”). Participants kept the money they earned in the game. We restricted decisions to whole dollar amounts in order to simplify payment.

In all cases, participants learned that they had been randomly chosen to be Player 2, and that Player 1 (the confederate) would make a decision about how to divide $10 between the two of them. Participants were randomly assigned to either the “take” game or the “give” game condition. In the take game, the participant first received $10, and the confederate decided how much of that sum to take. In the give game, the confederate received the $10 and decided how much to give to the participant. The decisions of the confederate were fixed, such that the participants were always left with $4, either because the confederate took $6 away from them in the first turn of the take game, or gave them $4 in the first turn of the give game.

The critical question was how the participant would divide a new pot of $10 when experimenters subsequently initiated a second turn with reversed roles. To prevent pre-planned strategies, participants were not told at the outset that there would be a second turn, thus the opportunity to reciprocate came as a surprise. In this second turn of the taking game, the confederate received $10 and the participant decided how much to take away. In the second turn of the giving game, the participant received $10 and decided how much to give the confederate. Before beginning the second turn, participants were randomly assigned to either the reflexive or deliberative condition. In the reflexive condition, the experimenter urged participants to make a decision as soon as possible after learning of the decision of the confederate in the first turn. Specifically, as soon as the confederate radioed Player 1's decision, which was audible to the participant because the experimenter held the walkie-talkie between them, the experimenter would immediately prompt the participant to make a decision regarding the new pot of $10. The experimenter silently counted off 2 s, and if the participant had not responded by this point, the experimenter gave further instructions to “just go with your gut.” If the participant did not respond within another 2 s, the experimenter recorded this in the experimental notes and that participant was later excluded on the grounds that the five or more seconds of deliberation disqualified him or her from the reflexive condition.

In the deliberative condition, the experimenter urged the participant to deliberate about the decision of the first dictator before making a reciprocal decision. Specifically, as soon as the experimenter radioed the confederate's decision, the experimenter handed the participant a pen and clipboard containing an evaluation scale. The participant evaluated the decision using a simple Likert scale from −5 (“extremely selfish”) to +5 (“extremely generous”). After 30 s, the experimenter took the clipboard and prompted the participant to make a decision on the new pot of money. Participants reported their decision verbally to the experimenter. Participants were then probed for suspicion that the other participant was a confederate.

We defined a decision as escalation if the second dictators allotted more money to themselves than the $6 the first dictator allotted to himself. If negative escalation is a reasoned, strategic decision, then escalation should be more pronounced in the deliberative than the reflexive condition of the take game. If negative escalation is reflexive and intuitive, then it should be more pronounced in the reflexive condition of the take game. The give game served as control, as giving typically produces in-kind reciprocity rather than escalation (Keysar et al., 2008).

[1] At this point in the experiment, after excluding participants for reasons noted in this section, a one-way ANOVA on the amount claimed showed a significant effect of condition, F (3, 101) = 4.639, p = .004, η^2 = 0.06. In the taking game, participants took more money when responding reflexively ($7.13), escalating selfishness from the first turn, than when responding deliberatively (M = $5.57), t (50) = 2.87, p < .01, d = 0.81.

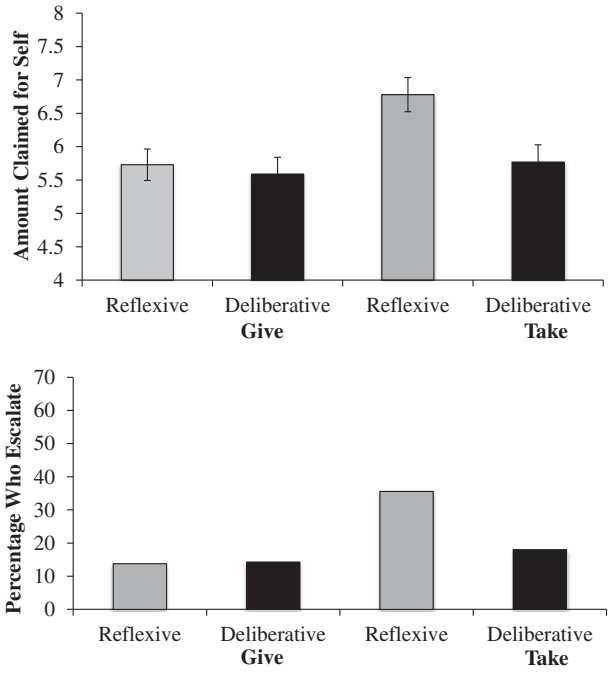

Fig. 1. Average amount claimed (Top panel), and the percentage of participants who escalated (Bottom panel), in response to positive (give) and negative (take) acts under reflexive and deliberative conditions (Experiment 1).

2.2. Results

As can be seen in the top half of Fig. 1, the results supported the reflexive theory of escalation. A one-way ANOVA revealed a significant effect of condition on the average amount claimed, F (3, 230) = 4.66, $p < .005$, $\eta^2 = 0.06$. More important, planned contrasts indicated that participants in the deliberative condition of the taking game claimed more than participants in the other three conditions, t (230) = 3.90, $p < .001$, $d = 0.51$. Participants in the taking game claimed significantly more in the reflexive condition than in the deliberative condition, t (230) = 2.84, $p < .01$, $d = 0.37$. Participants in the giving game responded similarly in the reflexive and deliberative conditions, t (230) = 0.15, $p = .88$, $d = 0.02$. Compared to the amount originally claimed by the initial participant ($6), only participants in the reflexive condition of the taking game escalated significantly by taking more in return, t (60) = 2.90, $p = .005$, $d = 0.75$.

The percentage of participants who escalated shows a similar pattern, as the percentages of those who claimed less (38%), the same (41%), and more (20%) than was originally claimed on the first round varied by experimental condition, $\chi^2(1) = 13.82$, $p < .001$. As shown in the bottom half of Fig. 1, participants escalated most frequently in the reflexive condition of the taking game, with participants in the remaining three conditions escalating much less.

This pattern provides evidence consistent with the reflexive account of negative escalation. Had escalation been the product of deliberate or strategic thinking of any kind, then participants should have taken more in the deliberative condition because they would have been able to use the time and effort to strategically generate a deterring response. In fact, exactly the opposite happened, as deliberation reduced escalation in the taking game. The absence of escalation in either condition of the giving game indicates that escalation is not a reflexive response to *any* action from another person. Instead, it suggests that escalation is a reflexive response to another's harmful or negative action.

One problem with this experiment is that not all participants followed our experimental instructions to respond quickly enough in the reflexive condition. It is possible that these participants would not have escalated had they responded more quickly, thereby potentially inflating the degree of escalation observed in the reflexive condition. However, including these participants in the analyses would not be reasonable because their failure to follow instructions means that they could not have been affected by our experimental manipulation. We therefore addressed this potential problem in two ways. First, we designed Experiments 2 & 3 to reduce attrition. Indeed, fewer participants are excluded from analyses there. Second, we used a different procedure in Experiment 5 that does not include a deadline, and hence would yield no attrition.

3. Experiment 2: cooling off or deliberate defusing?

Negative escalation was defused in the taking game of Experiment 1 when participants responded deliberatively. Experiment 2 tested two different explanations for this result: 1) that active deliberation defuses escalation or, 2) that the mere passage of time reduces an impulsive reaction regardless of deliberation. Experiment 2 tested these possibilities by replicating the taking game with an additional condition that allowed the passage of time but did not allow deliberation. If deliberation is actually required to defuse escalation, then the mere passage of time should not reduce escalation. We did not include the giving game conditions in this experiment because we were specifically interested in testing the role of delay in negative escalation, and participants do not escalate in the giving game conditions (Experiment 1; Keysar et al., 2008).

3.1. Method

3.1.1. Participants

Because Experiment 1 involved a large sample collected from pedestrian traffic around the University of Chicago campus, there was a chance that subsequent, similar experiments in the same area would be contaminated by word of mouth. Experiments 2 and 3 were therefore conducted in a busy tourist area in downtown Chicago with substantial pedestrian traffic. Because these experiments involved only take games, and the p-value in take games was well under α in Experiment 1, the target sample sizes of these subsequent experiments was reduced from 60 to 40 per condition, with an actual total of 131 participants. Two participants could not render decisions due to communication problems, leaving a total of 129 participants (56 female, 73 male) included in the analyses. We incentivized participants by allowing them to keep the money they earned during the game.

3.1.2. Procedure and design

Experiment 2 used the taking game procedure from Experiment 1 with three changes: 1) Each round used $5 instead of $10, 2) the confederate took away $3 instead of 6, and 3) there was a delay condition that allowed no deliberation.

In designing the delay condition we considered allowing participants to simply delay their response while doing nothing else. But such a procedure would have allowed deliberation as participants would naturally ponder what Player 1 did and how they should respond. Given that we wanted to evaluate the effect of deliberation per se and separate it from the time it takes to deliberate, it was crucial to allow the passage of time without allowing deliberation. Therefore, the delay condition included an irrelevant task that required the participants' attention. Instead of being handed a clipboard with the scale, participants in the delay condition were handed a clipboard containing a word search puzzle. The puzzle was a standard letter matrix containing the names of the fifty states of the US. Participants were instructed to "circle the first three states you find. This is not a race. We're just interested in which names pop out to you." In Experiment 1, participants often took less than the full 30 s to complete the scale in the deliberation condition. We conducted a norming study with 16 participants, which determined that the average completion times of the delay task is slightly < 30 s (M = 25.7 s, SD = 6.3). Therefore, the task was temporally equivalent

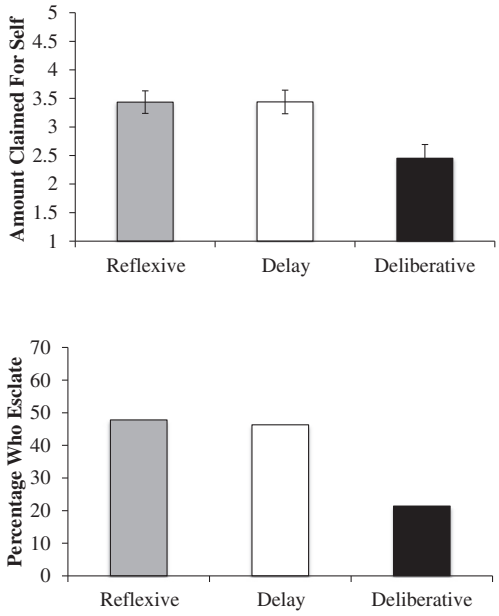

Fig. 2. Average amount claimed (top panel), and the percentage of participants who escalate (bottom panel), in the reflexive, delay, and deliberative take conditions (Experiment 2).

to the deliberative condition.

This allowed us to test whether the passage of time without direct reflection on the action of the other is sufficient for defusing escalation. As in Experiment 1, participants could only take an amount in whole dollars.

3.2. Results

Fig. 2 shows that participants escalated more in the reflexive and delay conditions than in the deliberation condition. A one-way ANOVA on the amount taken revealed a significant effect of condition F (2, 126) = 6.87, $p < .005$, $\eta^2 = 0.1$. Planned contrasts revealed that participants again claimed significantly more in the reflexive condition ($M = 3.43$) than in the deliberative condition ($M = 2.45$), t (126) = 3.26, $p = .001$, $d = 0.70$. Adding a delay alone did not significantly reduce escalation. Participants in the delay condition ($M = 3.44$) claimed the same amount as did participants in the reflexive condition, t (126) = 0.01, *ns*, but claimed significantly more than those in the deliberative condition, t (126) = 3.18, $p = .002$, $d = 0.71$. Compared to the amount the confederate initially took from them, participants in the delay and reflexive conditions significantly escalated, taking on average more than $3 from the confederate, *ts* (40 & 45) = 2.12 & 2.19, respectively, *ps* < 0.05, *ds* = 0.67 & 0.65. In contrast, participants in the deliberative condition actually acted more generously than the confederate, taking significantly less than the $3 originally taken from them, t (41) = −2.27, $p = .03$, d 0.71.

We also analyzed the percentage of participants who escalated by taking more than the confederate initially took from them. Whereas almost half of the participants escalated in the reflexive condition, only 21.4% escalated in the deliberative condition. Adding a time delay without deliberation did not reduce escalation, as just as many escalated in the delay condition as in the reflexive condition, χ^2 (2, $N = 129$) = 7.90, $p < .05$.

This pattern again suggests that negative escalation is reflexive, not strategic. Defusing escalation seems to require deliberation rather than a simple "cooling off" period. These results may seem at odds with a recent finding that a cooling off period reduces rejection of unfair offers in the ultimatum game (Neo et al., 2013). At this point we simply note that our results involve a different type of interpersonal action, namely escalating a seemingly negative action in a repeated interaction. In the General Discussion we will address what we believe are important differences between seemingly similar interpersonal actions.

Although this experiment suggests that the mere passage of time does not defuse escalation, it does not identify what aspect of deliberation defuses escalation. Impulsive interpersonal actions are typically egocentric, focused on one's own perspective without taking the time or expending the effort necessary to consider another person's point of view (Epley, Keysar, VanBoven, & Gilovich, 2004; Epley, Morewedge, & Keysar, 2004; Lin, Keysar, & Epley, 2010; Roxβnagel, 2000; Wardlow, 2013). It might be that a broader perspective on the exchange, rather than an egocentric focus on one's own outcome, contributed to defusing escalation in Experiments 1 and 2. However, it is also possible that specifically asking participants to focus particularly on others' selfishness may account for the results because it made participants more self-aware about the potential selfishness of their own response. If so, then considering others' selfishness alone could explain the effects of deliberation on reducing negative escalation. Experiment 3 tested this possibility.

4. Experiment 3: self-focus or general evaluation?

If deliberation defuses escalation because it broadens one's perspective beyond an egocentric focus on one's own outcome, then the specific content of deliberation should matter less than the broader act of deliberating about another's action. To test this idea, Experiment 3 adds a deliberation condition in which participants evaluate another aspect of the other person's behavior, its commonality, rather than its selfishness. If defusing escalation stems from providing a broader perspective on the decision at hand, then this act of deliberation should defuse escalation as well.

4.1. Method

4.1.1. Participants

We conducted Experiment 3 in a popular tourist area in downtown Chicago to continue sampling from a more diverse population than we could obtain on a university campus. We targeted 40 participants per condition, continuing data collection through the end of a scheduled session as we neared that target. We again incentivized participants by allowing them to keep the money they earned during the game. In total, 123 people participated, but 10 were excluded due either to communication problems (1), not accepting payment (7), not following instructions (1), or not answering within 2 s of the "go with your gut" prompt (1), leaving 113 participants (45 female, 68 male) in our final analyses.

4.1.2. Procedure and design

The procedure was identical to Experiment 2 except that we replaced the delay condition with a condition using a different evaluation scale than the deliberative condition. This second scale was identical except that it was nonnumeric, and it focused on a different aspect of the action. In particular, participants in this condition evaluated how common the first dictator's action was, rather than how selfish it was. This experiment therefore included three conditions of the take game: reflexive, deliberative (selfish), and deliberative (common).

4.2. Results

As predicted by the reflexive theory, participants were more likely to escalate in the reflexive condition than in the two deliberation conditions, which did not differ from each other. As shown in the top panel of Fig. 3, participants took more money on average in the reflexive condition than in the two deliberative conditions. An ANOVA on the average amounts the participants claimed showed a main effect for condition, F (2, 110) = 7.38, $p < .005$, $\eta^2 = 0.12$. Orthogonal

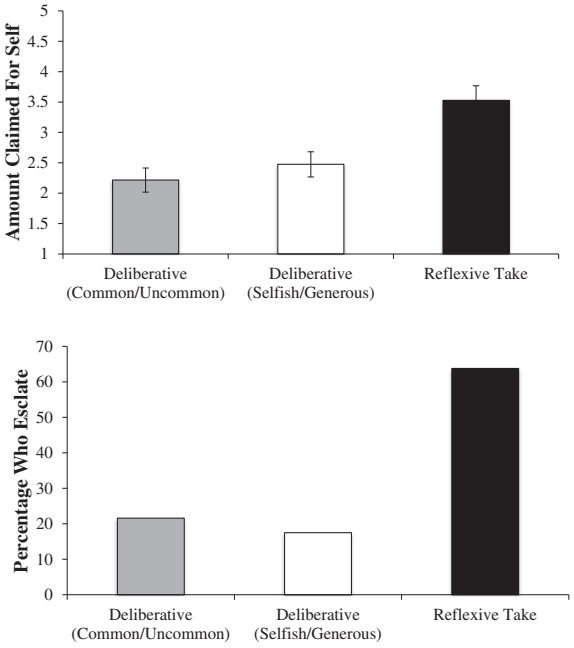

Fig. 3. Average amount claimed (top panel), and the percentage of participants who escalated (bottom panel), when participants deliberate about the commonality or selfishness of the action compared to when they respond reflexively (Experiment 3).

contrasts indicated that amount taken in the reflexive condition differed significantly from both deliberative conditions (uncommon: t (70) = 3.47, $p < .005$, $d = 0.84$; selfish: t (75) = 3.17, $p < .01$, $d = 0.76$). As in Experiment 2, participants in the two deliberative conditions were somewhat more generous than Player 1, taking significantly less than $3 on average, ts (39 & 36) = −2.29 & −2.89, selfish & uncommon respectively, $ps < 0.05$, $ds = 0.73$ & 0.95. In contrast, participants in the reflexive condition took significantly more than $3, on average, t (35) = 2.16, $p = .04$, $d = 0.73$. Likewise, the bottom panel of Fig. 3 shows that many more participants escalated in the reflexive condition (64%) than in the two deliberative conditions (selfish, 22%; uncommon, 18%), χ^2 (2, N = 113) = 21.9, $p < .001$ (See Panel B of Fig. 3). These results indicate that deliberation may defuse escalation because it triggers a broader consideration of the exchange beyond one's own egocentric perspective (Gross, 1998), rather than being the product of the specific content we used during deliberation in Experiments 1 and 2.

Experiments 1–3 suggest that deliberation before reacting to a negative action defuses escalation. These results are consistent with a reflexive theoretical account of escalation and inconsistent with a deliberative or strategic account. We provide additional tests of the reflexive versus deliberative nature of escalation in Experiments 4–7. First, if the reflexive account is correct, then deliberation should reduce escalation even when the exchange develops over time through repeated interactions. Experiment 4 investigates this possibility. Second, Experiment 5 evaluates the predictions of the two accounts without explicit instructions for deliberation, but instead by manipulating the availability of cognitive resources. Third, Experiment 6 provides the most direct test of the strategic account's assumption that deliberation over future consequences leads to escalation by asking participants to explicitly focus on future consequences. Finally, Experiment 7 manipulates goals explicitly to test whether escalation results from an attempt to deter or to punish.

5. Experiment 4: escalation in repeated exchange

Experiments 1–3 found escalation only in a single response to another person's action, but at least some (if not most) escalation in daily life occurs over repeated exchanges from different starting points with some ambiguity in the degree of harm involved. Experiment 4 investigated whether these results generalize to repeated exchanges without a fixed opening action. In this procedure, participants played a 5-round dictator game in which they divided a pot of coins with another person either by giving or by taking. Participants in the take condition took some amount of $100 in coins out of the bowl of another participant over a series of repeated exchanges. Participants were literally taking money from another person, right in front of their eyes. Participants in the give condition, in contrast, gave coins to the other person by pouring them into the other participant's bowl. If the results of Experiments 1–3 generalize to a repeated exchange, then negative escalation should occur in the reflexive condition of the taking game, but not in the deliberative condition.

5.1. Method

5.1.1. Participants

One hundred seventy-eight participants completed this experiment in pairs (89) in a university laboratory in exchange for $3 and an entry into a raffle, wherein they could win the money they earned in a random round of the game. All pairs were composed of strangers. We targeted 20–25 pairs per condition based on the sample size of previous experiments using a similar repeated game (Keysar et al., 2008).

5.1.2. Procedure and design

We did not use a confederate in this experiment to allow natural exchanges between rounds. Participants were recruited in pairs and each pair was randomly assigned to either a give or a take game and to either a deliberative or reflexive condition. Participants sat at a table with a barrier that allowed them to see only the other participant's hands. They were not allowed to communicate. In the take condition, the experimenter put a bowl filled with quarters totaling $100 in front of one participant and the other participant took from it as much as he or she wanted by reaching under the barrier. Participants in the deliberative condition then evaluated the other's action before moving to the next turn. Participants in the reflexive condition did not evaluate the other person's action and instead moved to the next turn immediately. In the give condition, the experimenter placed the bowl of quarters in front of one participant who then chose to give as many coins to the other participant as he or she wanted by reaching across the barrier to put coins in the other participant's bowl. As in Experiment 1, these giving conditions serve as a control because we do not expect giving to trigger escalation, and therefore do not expect a difference between the deliberative and reflexive conditions.

Each pair played 5 rounds of back and forth, for a total of 10 individual turns. The decision in each of the ten turns involved a new pot of $100 worth of quarters. Participants were told to take (or give) as much as they wanted by moving coins from the full bowl to the empty bowl. Whereas participants in previous experiments were not initially informed that they would have the opportunity to reciprocate, due to the extended number of exchanges in this experiment, participants were told in advance that they would be making alternating decisions. However, to prevent end game effects, they were not told the number of turns they would be taking.

5.2. Results

We divided the ten individual turns to five rounds and averaged both players' turns in each round. Fig. 4 shows the average amount participants ended up with for themselves across the five rounds. Escalation across rounds occurred only when responding reflexively in the take condition, with pairs in this condition beginning fairly generously, claiming an average of only $34 in coins in round 1, but escalating to $50 in round 5. The predicted three-way interaction was significant, F

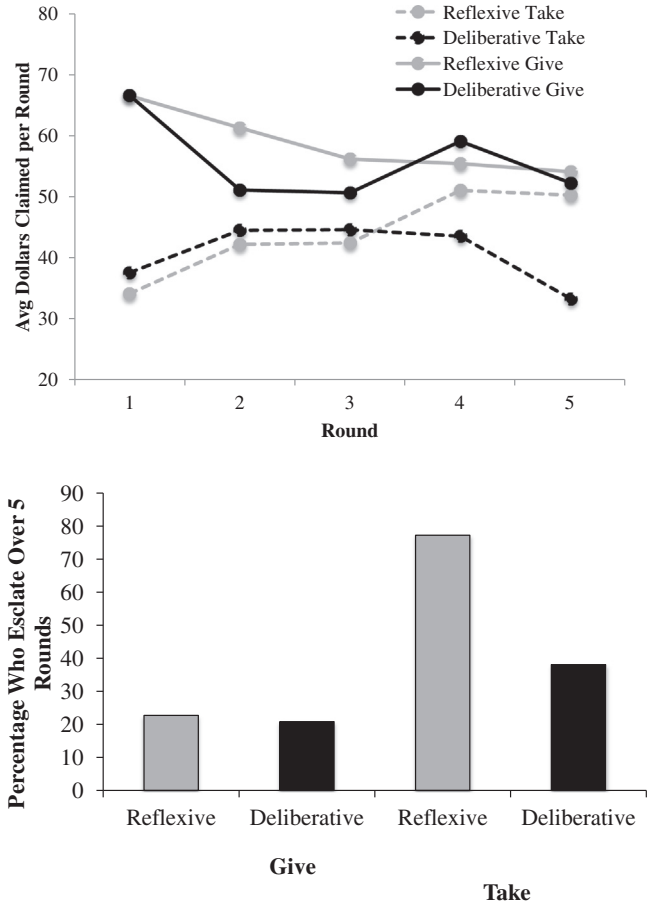

Fig. 4. Average amount claimed (top), and the percentage of pairs that escalate between the first and last decision (bottom), in positive (give) and negative (take) multiple-round games under reflexive and deliberative conditions (Experiment 4).

(4, 340) = 2.96, $p = .02$, $\eta^2 = 0.03$, and paired *t*-tests between initial and final rounds of taking conditions shows that reflexive taking escalates significantly, paired $t\ (21) = -2.7$, $p < .05$, $d = 0.74$, whereas deliberative taking does not, $t < 1$. In the give condition, we observed no significant escalation in either the deliberative or reflexive condition. In fact, participants in the give condition became more generous between round 1 and round 5 in both the deliberative and reflexive conditions, paired *ts* (21, 23) = 2.85 & 2.16, respectively, both *ps* < 0.05. This same pattern of differences emerges when looking simply at the percentage of pairs that escalate the amount kept for oneself between the initial behavior on round 1 through the last behavior on round 5, $\chi^2\ (3, N = 89) = 19.27, p < .001$. In the take condition, 77% of pairs escalated in the reflexive condition compared to only 38% in the deliberative condition. In the give condition, only 21% and 23% of pairs escalated in the deliberative and reflexive conditions, respectively. These results in the take conditions are particularly striking because participants in the deliberative condition had ample opportunities over multiple rounds to escalate strategically, and yet we observed no negative escalation at all in these conditions. These results are consistent with the theory that escalation emerges as a relatively thoughtless and impulsive reaction to a perceived slight, but inconsistent with the theory that escalation is caused by a deliberate act of strategic deterrence.

6. Experiment 5: escalation and cognitive resources

Experiments 1–4 found that deliberation can mitigate escalation, suggesting that escalation occurs reflexively rather than strategically. Experiment 5 provides another test between competing mechanisms of escalation by manipulating available cognitive resources. If escalation occurs deliberatively, then it should require cognitive resources, and should be reduced when a person lacks the cognitive resources necessary for deliberative thinking. However, if escalation occurs reflexively, then negative escalation should be more pronounced when a person *lacks* the cognitive resources necessary for deliberative thinking. Experiment 5 tests these predictions by studying whether people escalate more or less when placed under cognitive load.

Experiment 5 also allows us to evaluate if the preceding results are due to a potential artifact of the experimental manipulation. Specifically, participants in the reflexive conditions of the preceding experiments were not only asked to respond quickly, but were also informed about how to make their decision (e.g., “going with the gut”). It is possible that these instructions artificially induced the effect. Experiment 5 avoids such potential concern by simply manipulating cognitive resources without any further instruction about how to make the decision.

6.1. Method

6.1.1. Participants

A total of one hundred ninety-eight participants who were recruited to a laboratory in downtown Chicago completed this experiment as pairs in exchange for $5 and a 1 in 36 chance (rolling two 1's with a single roll of two dice) to win the money they accrued during the game. We targeted 20–25 pairs per condition based on the sample size of the previous experiments using a similar task (Keysar et al., 2008). We excluded one pair from all analyses because one of the participants acted strangely during the experiment, repeatedly apologizing for his poor memory performance purportedly caused by a long history of drug abuse. Participants in each pair were strangers.

6.1.2. Procedure and design

Participants were told that they would be performing a series of small tasks under conditions of distraction. Each pair of participants played two turns of the give or take game in a modified version of the coin procedure used in Experiment 4. Because grabbing handfuls of coins from a bowl can entail multiple actions, the physical effort of this task may affect results, especially for participants under high cognitive load. We therefore had participants give or take coins by pouring them from a wide-mouthed jug into buckets, using dimes instead of quarters because they poured more easily. Instead of including a deliberative condition, this experiment manipulated the degree to which participants were under cognitive load during the interaction. In the take game, the first dictator took from the other participant's $100 sum of dimes, and then the second dictator had an opportunity to reciprocate by taking from the first dictator's new $100 sum of coins. We again included a give game condition to serve as a comparison.

We manipulated the availability of cognitive resources by adding a secondary task. Half the pairs were in the high cognitive load condition, and the other half in the low cognitive load condition. The high load procedure reliably reduces performance on tasks requiring deliberation (Halford, Bain, & Mayberry, 1984; Logan, 1979). The procedure consists of three parts: practice, capacity determination, and the actual game. All three parts had a similar basic format. The goal of the first part was simply to acquaint the participants with the operation of the cognitive load task. The load manipulation used a standard symbol string task, in which a symbol string appeared on the screen, and participants had 12 s to study it. After the string disappeared, participants were instructed to perform a manual task. In the actual experiment, the dictator game replaced the manual task. In the practice phase, the manual task was to draw specified shapes on a pad of paper. When finished with the manual task, participants pressed a computer key, at which point they were instructed to enter the string of symbols they had studied before the manual task. The symbol strings in the practice

period were only two symbols long.

The goal of the second part was to assess the cognitive capacity of each participant in order to calibrate the level of required load during the critical period of the experiment. To do so, the symbol string increased in length by one symbol after every trial the participant got correct during the second period. This was repeated until the participant failed to recall the string correctly. This enabled a rough approximation of cognitive capacity. The manual task was also made more difficult to better approximate the decision task in the actual game, where participants would have to move coins between containers. Specifically, participants' manual task was to open a numbered envelope specified by the program, remove a set of playing cards that it contained, place the cards in an order specified by the program, and then proceed to recall the symbol string they had seen previous to the manual task prompt.

The third part of the procedure was the repeated dictator game. In the high load condition, participants received symbol strings that were the same length as the longest symbol string they successfully completed in the second part. In the low load condition, participants received a two-symbol string. Each pair was randomly assigned to either high load or low load condition. To ensure that the second dictators, the reciprocators, were not differentially prevented by load from perceiving the actions of the first dictators, participants were placed under cognitive load only when it was their turn to make the decision. This produced the following procedure. The first dictator received a string to memorize, then made an allocation decision over the $100 sum of coins. The first dictator then recalled the string. The second dictator then received a string to memorize, made an allocation decision over a new set of coins, and recalled the string.

6.2. Results

Eleven participants in the role of the second dictator failed to recall the number string in the cognitive load manipulation correctly, indicating a failure of our key manipulation among these participants. We therefore excluded them from the following analyses.

Based on the results of Experiments 1–4, we predicted that cognitive load would increase escalation in response to another person's negative action, consistent with a reflexive account of escalation. Fig. 5 shows that this prediction was confirmed: second dictators in the take game escalated more when under high load by taking more money for themselves ($69.56) than was initially taken from them ($45.09), *paired* t (19) = −3.38, p = .003, d = 0.79. Participants under low load in the take game, if anything, took slightly less from the first dictator ($45.73) than was initially taken from them ($48.79), *paired* $t < 1$. No escalation occurred under either load condition of the give game, both $ts < 1.1$, again indicating that escalation is a reflexive response to another's perceived negative action (taking rather than giving). The predicted three-way interaction in an ANOVA was significant, F (1, 82) = 4.67, p = .03, η^2 = 0.03. A significant two-way interaction in the taking game alone indicates that negative escalation occurred only when participants reciprocated taking under high cognitive load, F (1, 38) = 6.04, p = .02, η^2 = 0.04. Likewise, the bottom panel of Fig. 5 shows more participants in the taking game escalated when under high load compared to low cognitive load, but that the frequency of escalation was unaffected by cognitive load in the giving game.

It might seem unexpected that Player 2 would escalate after Player 1 in the high load condition took slightly less than half the total amount. By splitting the money roughly in half, Player 1 may appear to act very fairly. However, participants in past experiments evaluated someone who *took* half of the total from them as behaving in a relatively unfair and negative way, even more so than a person who *gave* them only 30% of the total amount (Keysar et al., 2008). In these cases, taking half is evaluated negatively, thereby explaining why we would expect escalation even in response to what is an objectively generous allocation.

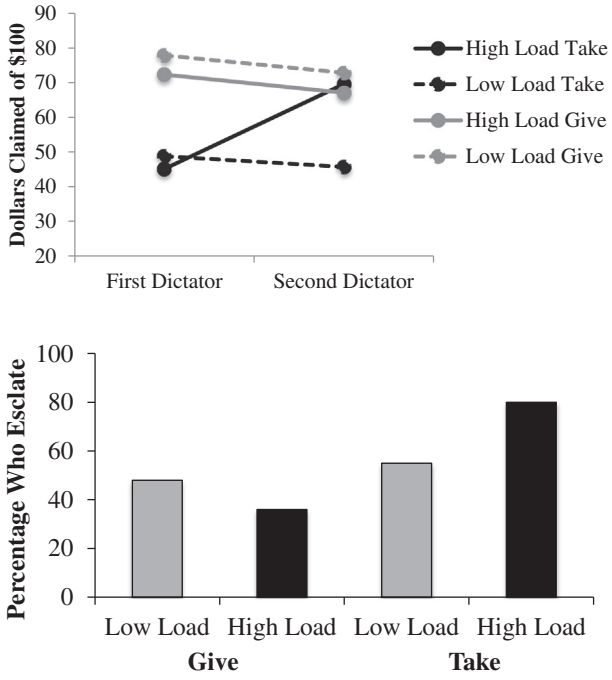

Fig. 5. Average amount claimed by the first and second dictator, and the percentage of second directors who escalate, in positive (give) and negative (take) conditions under high versus low cognitive load (Experiment 5).

7. Experiment 6: deliberating about future consequences

Experiments 1–5 suggest that escalation is a reflexive response to another's perceived negative action, rather than a deliberate or strategic attempt to deter future harm. When people respond reflexively we observe no escalation in response to what is perceived to be a positive action (i.e., giving), but do observe escalation in response to a perceived negative action (i.e., taking). None of these experiments, however, test whether escalation *could* occur when people focus on the future. According to the deliberative account, escalation results not just from thinking carefully, but rather from thinking strategically about future consequences. Escalation, on this account, is a deliberate and strategic attempt to deter future harm. This suggests that focusing participants on future actions in an exchange could increase escalation through a deliberative process. Experiment 6 tests this hypothesis directly by instructing participants to deliberate about the future consequences of their action. The reflexive and deliberative theories of escalation once again make opposing predictions. If people escalate as a strategic act of deterrence, then focusing on the future consequences of their action creates conditions that facilitate escalation. The deliberative theory therefore predicts that focusing on the future would increase escalation compared to a reflexive response. However, if escalation is reflexive, then deliberating about future consequences should defuse escalation because it involves more thoughtful deliberation about the situation and this should mitigate a reflexive reaction. The reflexive theory therefore predicts that explicitly considering future consequences would decrease escalation compared to a reflexive response.

7.1. Method

7.1.1. Participants

People recruited from public places (N = 314) on the University of Chicago campus participated in the experiment.

7.1.2. Procedure and design

This experiment was pre-registered. Details can be found here: http://aspredicted.org/dmrm7.pdf.

The procedure was similar to that of Experiments 1–3. Participants believed they were playing with another participant. As before, a confederate Player 1 was the dictator in the first round, and then the participant was the dictator in the second round. In the first round the participant received $10 in singles, and Player 1 took $6 out of it. As in Experiments 1–3, the second experimenter used a walkie-talkie to communicate Player 1's decision, and the experimenter physically removed the $6 from the participant's pot. The participant was then informed that Player 1 received a new pot of $10 and had to decide how much or it to take away. The participants knew that they get to take home the money the other Player leaves them, and the money they take from Player 1.

Participants were randomly assigned to one of the three conditions. Participants in the reflexive condition were required to make a quick decision in no more than 2 s, and were instructed to "just go with your gut." Participants in the deliberate condition were instructed to think carefully about their decision, to not go with their gut but to slow down and reason analytically about their choice. They were provided with a notepad and encouraged to write down their thoughts and to report their decision only after 30 s of careful deliberation. Finally, participants in the future consequences condition received a new deliberation instruction that explicitly encouraged them to think strategically about the future consequences of their actions. Specifically, participants received the notepad and the same instructions as in the deliberative conditions, but were also told to think carefully about the future consequences of their action, how the other person is likely to respond to whatever choice they make, and how the exchange is likely to continue to unfold.

After collecting our primary measure of how much participants took from the other player, we also asked participants to report how many rounds they thought the experiment would continue for, how fair they though their own response was, how fair they thought the other person would rate their response to be, and some demographic variables (gender, age, ethnicity, and occupation). Responses on the items did not vary across conditions, and we did not have any specific hypotheses about them. We therefore do not report analyses of these items.

7.2. Results

One participant was removed from the following analyses due to a data entry error (ID number was mistakenly entered in as the primary dependent variable).

A one-way ANOVA indicated a significant effect of condition on the amount participants claimed for themselves, F (2,312) = 3.27, p = .04, η^2 = 0.02. Explicitly encouraging strategic thinking about future consequences reduced, rather than increased, escalation. Planned contrasts indicated that participants took significantly less from "Player 1" in the future consequences condition (M = 5.42) than both the reflexive condition (M = 6.07), t (310) = 2.55, p = .01, d = 0.32, and the deliberation condition (M = 5.76), t (310) = 2.55, p = .01, d = 0.31. Although participants in the reflexive condition again took more than did participants in the deliberation condition, this difference was statistically nonsignificant, t (310) = 1.36, p = .18, d = 0.15. Similarly, fewer participants escalated by taking more than $6 from the other player in the future consequences condition (20%) than in the reflexive condition (35%).

As the reflexive account of escalation predicts, people took less when they reflected on future consequences than when they reacted with their gut. This experiment also rules out an alternative explanation to Experiments 1–4 that assumes that the particular deliberation instructions we used reduced escalation because they distracted participants from focusing on the future. Even when explicitly directed to evaluate future outcomes, escalation is diffused compared to when people respond reflexively. Unlike the preceding experiments, the deliberation condition did not significantly reduce escalation compared to the reflexive condition. Although the result was in the same direction, this result suggests that a broad instruction to deliberate may not be as robust a method for defusing escalation as deliberating about future consequences more specifically. Future work might identify how the specific content of deliberation affects defusion, and whether focusing on the future is truly a more reliable method for defusing escalation. For now, these results yield only a more modest conclusion that escalation in this situation is not the product of strategic thinking about future consequences.

8. Experiment 7: escalation as goal pursuit

Experiments 1–6 consistently indicate that people do not escalate their response to a negative action when they deliberate carefully. These experiments also indicate that reacting quickly is more likely to lead to escalation than deliberating before reacting. These experiments do not, however, explain why reflexive reactions to negative actions consistently produce escalation, whereas more deliberate responding does not. We suggest that the underlying mechanism that can explain this are the goals that are guiding people's actions: The reflexive response reflects a goal to punish another person whereas the deliberative response reflects a goal to reduce or avoid harmful conflict. If the reflexive theory is correct, then those given a goal to punish another person should be more likely to escalate than those trying to prevent future conflict. We tested this hypothesis directly in a final experiment by manipulating people's goals and measuring their response to another's negative action.

Specifically, we described the multiple round dictator game used in the preceding experiments and asked participants to prepare to respond to another person's action. We explained that in each round one person would get $10 and the other would decide how much to take away. Participants were asked to report how they would react if the other person were to take $6 from them. They were instructed to achieve one of three goals: to (1) punish the other player for taking more than half, (2) deter the other player from taking more than half in the future, or (3) strategically lead the other person to not take more than half in the future. If people escalate strategically in order to deter greedy action then they should escalate when they are explicitly trying to deter future selfish behavior in the last two conditions. If, however, escalation stems from a reflexive goal to punish another person, then they should escalate only when trying to punish the other person.

8.1. Method

8.1.1. Participants

We targeted 100 participants per condition and allowed anyone still completing the survey to finish once we reached this target. This yielded three hundred and six participants from Amazon's Mechanical Turk who completed the survey in exchange for a small monetary fee (51% female).

8.1.2. Procedure

This experiment was pre-registered. Details can be found here: http://aspredicted.org/blind.php?x=fg2ux2.

Participants were told to imagine that they were going to play a multiple-round game with another participant. In each round, one participant would receive $10 and the other participant would be able to take away as much as he or she wanted. They were also told that they will alternate roles between rounds. The instructions mentioned that taking $5 would be a fair amount as it divides the pot equally.

This procedure was meant to mimic contingency planning. Specifically, participants were told that their goal was to prepare for the game by deciding what to do in response to the other participant's action. Each participant was told to imagine that he or she received $10 in the first round of the game, that the other participant took away $6, and that they were to plan how much to take away from the other person out of $10 on the second round of the game. Participants were then

randomly assigned to one of three goal conditions. Participants in the strategic deterrence condition were told to "respond in a way that most strategically deters your partner from responding unfairly to you again by taking more than $5 from you in future rounds." Participants in the fairness condition were told to "respond in a way that makes it most likely that your partner will start responding fairly in future rounds, taking only $5 from you instead of taking more than $5." Finally, participants in the punish condition were told, "You would like to punish your partner for taking more than the fair amount from you on the first round." All participants were then asked to indicate how much out of $10 they would take from their partner in order to best achieve their stated goal. As an attention check, participants were then asked to recall the amount that the other participant took from them in round 1.

8.2. Results and discussion

We excluded 84 participants because they failed the attention check, but including them does not meaningfully affect the results we report below. This left 222 participants in our primary analyses (73 in the fairness condition, 73 in the deterrence condition and 76 in the punish condition).

A one-way ANOVA indicated that the amount of money participants took from the other player varied across conditions, $F(2,219) = 15.62$, $p < .001$. The pattern of the means clearly supports the reflexive theoretical account and argues against the deliberative account (See Fig. 6). Only participants in the punishment condition escalated the other person's selfish action by planning to take significantly more than the other person took from them ($6), $t(99) = 2.89$, $p < .01$, $d = 0.58$. Participants in the deterrence condition, in contrast, tended to reciprocate the other person's action by taking the same amount that was initially taken from them, $t(101) = -0.76$, $p = .45$, $d = 0.15$. Those in the fairness condition actually took less, on average, than was originally taken from them, $t(103) = -5.24$, $p < .001$, $d = 1.03$. Planned contrasts indicated that those in the punishment condition took more than participants in both the deterrence and fairness conditions, $ts(219) = 3.38$ & 5.54, respectively, $ds = 0.46$ & 0.75. Participants in the deterrence condition also took more than participants in the fairness condition, explicitly attempting to deter future unfair behavior tended to reciprocate others' action, $t(219) = 2.14$, $p < .04$, $d = 0.29$. Likewise, the percentage of participants that planned to escalate varied across conditions, $\chi^2(2) = 29.29$, $p < .001$, with the most participants escalating in the punish condition (41%) compared to the fairness (10%) or deter (20%) condition.

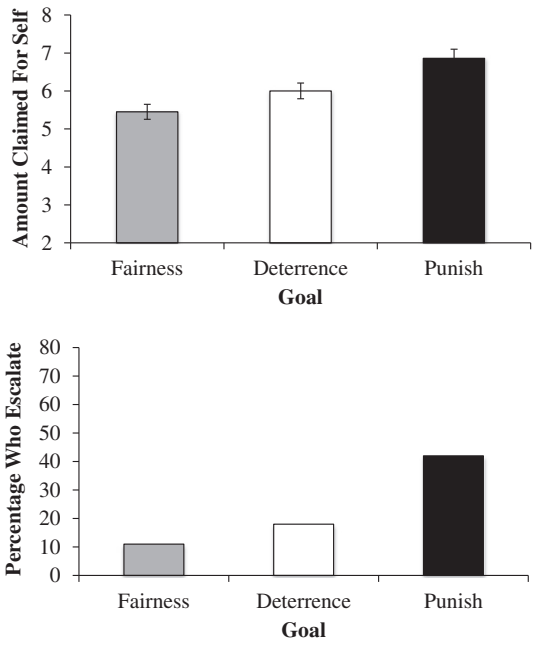

Fig. 6. Average amount participants planned to take (Top), and the frequency of planned escalation (Bottom), in response to another person's relatively negative action in order to achieve a goal of fairness, deterrence, or punishment (Experiment 7).

These results suggest that escalation is more likely to be guided by a goal of punishing another person's negative action rather than by an explicit effort to deter future negative actions. These results also suggest that explicitly trying to deter future harm leads people to withhold escalation in an effort to encourage more positive behavior. Again, these results suggest that escalation is guided by a reflexive act of punishment, rather than a strategic or deliberate act of deterrence.

9. General discussion

Reciprocity is a central mechanism in social exchange: one of the few social norms that seems universal. Negative exchanges, however, are vulnerable to potentially harmful escalation. Instead of reciprocating in equivalent measure, those who've been harmed may reciprocate with increasingly harmful actions, creating an escalating cycle of violence. A better understanding of the psychological factors involved in the early stages of conflict has obvious utility for preventing escalation. Research from a number of related fields points strongly to two possible but distinct mechanisms: escalation could occur as a deliberative strategy intended to deter future harm, or it might result from an impulsive and reflexive act of retribution. Prior research identified an experimental procedure that reliably induces negative escalation. We used this procedure to test competing theories of escalation across seven experiments.

The experiments consistently support the predictions of a reflexive rather than deliberative theoretical account of escalation. Participants tended to escalate a negative action—but not a positive action—from others when they responded reflexively (Experiments 1–4) and when they were unable to deliberate because they were distracted by another task (Experiment 5). Participants did not typically escalate a negative action when they deliberated about some aspect of the action itself, but a delay or "cooling off" period without this deliberation did not defuse escalation (Experiment 3). Experiment 6 provided the most direct test of the deliberative account of escalation by instructing people to focus on future consequences before reacting. Contrary to predictions from the deliberate deterrence account, escalation did not increase but instead decreased. Overall, comparisons between conditions in cases when we expected to observe escalation in Experiments 1–6 produced an average effect size (d) of 0.59.

Finally, Experiment 7 directly tested how people react when they attempt to deter, induce fairness, or punish an unfair action, demonstrating that escalation occurs, on average, only when people are attempting to punish. Taken together, these experiments suggest that escalation, at least in this context, is a reflexive response to perceived harm meant to punish rather than a deliberative choice meant to deter future harm.

Two additional features of our results deserve mention. First, we observed no escalation following subjectively positive actions, that is, when the initial action was giving instead of taking. This is important because it suggests that construal of another's action as either positive or negative is a critical moderator of escalation. This is consistent with escalation being driven by a desire to punish another person rather than an attempt to encourage a specific type of future behavior. Second, the impact of deliberation in our experiments did not depend on its precise content, as we observed similar results when others thought carefully about the selfishness or commonality of the other person's action, and also when they deliberated about the future consequences of their actions. These results suggest that simply thinking more carefully encourages a broader perspective on the decision at hand, and hence a less negative response, than the narrower focus on punishing the other

person when people respond reflexively to another's negative action.

Of course, our results do not demonstrate that escalation *cannot* be deliberate or strategic in specific instances that we have not studied here. Although we believe the experimental procedure utilized in our experiments captures the basic psychological properties that tend to create escalation, it is always possible that there are exceptions. In Experiment 7, for instance, a roughly equivalent percentage of participants planned to take everything ($10) from their partner on the second round of the game in both the punishment and the deterrence conditions, even though escalation was considerably more likely overall in the punishment condition. In addition, 20% of participants in the deterrence condition chose to escalate (compared to 41% in the punish condition). Some people may believe that escalation is an effective deterrent, but our theorizing at this point cannot explain where that belief might come from. Note also that Experiment 7 does suggest a situation where deliberation could encourage escalation: when people have an explicit goal to punish another person. Overall, our results suggest that this explicit goal is less common when people think more deliberately about their response to another's negative action. Future research might identify other contexts that reliably induce negative escalation to test whether similar or different motivations could be operating in these contexts. At this point, we tested the only experimental procedure we know that reliably produces escalation and found it to be driven by reflexive responding.

Although our results clearly suggest that escalation is a reflexive response, the belief that escalation is used for strategic deterrence is present not only in the psychological literature, but also in common intuition. Indeed, the pilot test we presented in the introduction demonstrated that survey responders expected negative escalation in the experimental procedure we utilized, but did not anticipate more escalation in the reflexive condition of this game. If anything, they expected more escalation in the deliberation condition. This pilot test suggests that our results may be at least somewhat inconsistent with people's intuitions, and therefore surprising. These results also suggest that people may interpret escalation as being motivated by strategic deterrence rather than reflexive punishment. Even as people escalate out of a reflexive desire to punish others, they may interpret others behavior or their own as being driven by strategic deterrence.

Several existing findings are consistent with this possibility of a gap between perceived and actual motivations. In criminal law, for instance, extreme punishment for wrongdoing is often said to be motivated by strategic deterrence, even though actual support for extreme punishment seems motivated by impulsive retribution (e.g., Carlsmith, 2006, 2008; Carlsmith, Darley, & Robinson, 2002; Carlsmith & Sood, 2009). Impulsive escalation in response to perceived harm, as we observed, could then masquerade as a seemingly deliberative or strategic act. Understanding how people interpret their own and others' escalating behaviors would be a very promising line for future research. The survey we just described suggests that people may believe that escalation is driven by a motive for deterrence more often than is justified.

Our experiments help clarify a growing body of research examining people's spontaneous versus deliberative inclinations in social exchange. Some research suggests that under certain circumstances cooperation is spontaneous whereas competition, in the form of benefit-destroying behavior, is deliberative (Rand, 2016). In one series of experiments suggesting a spontaneous cooperative response that creates value, people were more likely to cooperate by giving more in a public goods game when they responded quickly but behaved more selfishly after deliberating (Rand et al., 2012). Other research suggests a spontaneous response that destroys value for oneself and others. In ultimatum games, people are less willing to accept unfair offers when they respond reflexively than when they respond more deliberatively (Grimm & Mengel, 2011; Neo et al., 2013; Smith & Silberberg, 2010). These results suggest spontaneous selfishness in the form of impulsive retribution, consistent with our findings. Collectively, these results do not indicate that people are either spontaneously cooperative or competitive, but instead that cooperation varies by context (Rand, 2016). When acting towards others without information about their action, cooperation may be spontaneous. But when *reacting* to another's perceived slight, impulsive retribution that actually increases collective harm seems spontaneous.

The effects we observed suggest that escalation is likely to be moderated by cultural norms and practices that alter the frequency of deliberate versus reflexive thought. For instance, one growing body of research suggests that poverty is characterized by chronic cognitive depletion due to the daily stress of living in a state of scarcity (Mani, Mullainathan, Shafir, & Zhao, 2013; Shah, Mullainathan, & Shafir, 2012). This suggests that socioeconomic status could moderate negative escalation because it affects the capacity to engage in deliberative thought. Although our experiments do not test for moderation by culture, the theory of negative escalation that our experiments support does make clear predictions about cultural moderation. Specifically, those cultural norms and contexts that moderate deliberate thinking should also moderate negative escalation. This research is clearly the critical next step for this research program.

Whatever the broader significance of an individual's escalation, our experiments suggest insights for those interested in breaking destructive cycles of negative escalation. In particular, defusing escalation might require more than simply slowing down a reflexive response in a "cooling off" period. Instead, it may require deliberate consideration that broadens a person's perspective beyond one's own outcome in the exchange. Such distancing can lead a person to consider another's reasonable perspective in a way that had been overlooked before, and lead a person to reinterpret a transgression in a way that elicits a less retributive response (Mischkowski, Kross, & Bushman, 2012). Our findings demonstrate that in the relatively low stakes reciprocal exchanges that characterize much of everyday life, a natural form of reflection—simply evaluating the other party's action—can defuse negative escalation.

More generally, understanding the mechanism of escalation is essential for avoiding the high cost it can inflict on individuals and societies (Heller, Pollack, Ander, & Ludwig, 2013). Indeed, two large-scale field experiments informed by the results of the experiments we have described found that an intervention aimed at encouraging deliberation among inner-city youth significantly reduced incarceration rates (Heller et al., 2015). This intervention, called Becoming a Man, reduced violent crime arrests by 44% in one experiment and reduced overall arrests by 31% in a second experiment, both large effect sizes compared to existing interventions targeting psychological variables such as education, self-control, or "grit." Psychological interventions aimed at increasing deliberative processing among those most at risk may prove more successful at breaking increasingly destructive cycles of violence because they target the cause of escalation: a reflexive impulse to strike back with an even heavier hand.

Open practices

This article earned Open Materials, Open Data, and Pre-registration badges for transparent practices. Data for all experiment are available at https://osf.io/jkbm9/. Pre-registrations for Experiments 6 and 7 can be found in the following locations, respectively: http://aspredicted.org/dmrm7.pdf, http://aspredicted.org/blind.php?x=fg2ux2.

Acknowledgement

The experiments reported here are based on the first author's doctoral dissertation. We thank Kaushal Addanki, Kate Burke, Xiao Bohannon, Erica Cisneros, Will Craft, and Janet Flores for assistance conducting the experiments, and Leigh Burnett, Sophie Holtzmann, Sheila Sernoff and Donald Lyons for help with the manuscript. This research was funded by the National Science Foundation (SES #1025676), the Chicago's Wisdom Research Project funded by the John

Templeton Foundation, and by the Booth School of Business.

References

Anderson, C. A., Buckley, K. E., & Carnagey, N. S. (2008). Creating your own hostile environment: A laboratory examination of trait aggressiveness and the violence escalation cycle. *Personality and Social Psychology Bulletin, 34*, 462–473.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390–1396.

Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters, 122*, 308–310.

Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., & van Winden, F. (2007). Reciprocity and emotions in bargaining using physiological and self-report measures. *Journal of Economic Psychology, 28*, 314–323.

Bushman, B. (2002). Does venting anger feed or extinguish the flame? Catharsis, rumination, distraction, anger, and aggressive responding. *Personality and Social Psychology Bulletin, 28*, 724–731.

Cappelletti, D., Güth, W., & Ploner, M. (2011). Being of two minds: Ultimatum offers under cognitive constraints. *Journal of Economic Psychology, 32*, 940–950.

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology, 42*, 437–451.

Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research, 21*, 119–137.

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.

Carlsmith, K. M., & Sood, A. M. (2009). The fine line between interrogation and retribution. *Journal of Experimental Social Psychology, 45*, 1–6.

Cialdini, R. B. (2001). *Influence: Science and practice* (4th ed.). New York: Harper Collins.

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General, 143*, 2279.

Denant-Boemont, L., Masclet, D., & Noussair, C. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory, 33*, 145–167.

Diekmann, A. (2004). The power of reciprocity: Fairness, reciprocity, and stakes in variants of the dictator game. *The Journal of Conflict Resolution, 48*, 487–505.

dos Santos, M., Rankin, D., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society, 278*, 371–377.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*, 348–351.

Epley, N., Caruso, E., & Bazerman, M. H. (2006). When perspective taking increases taking: Reactive egoism in social interaction. *Journal of Personality and Social Psychology, 91*, 872–889.

Epley, N., Keysar, B., VanBoven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*, 327–339.

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism and differential correction. *Journal of Experimental Social Psychology, 40*, 760–768.

Everett, J. A. C., Ingbretsen, Z., Cushman, F., & Cikara, M. (2017). Deliberation erodes cooperative behavior — Even towards competitive out-groups, even when using a control condition, and even when eliminating selection bias. *Journal of Experimental Social Psychology, 73*, 76–81.

Fabiansson, E. C., & Denson, T. F. (2012). The effects of intrapersonal anger and its regulation in economic bargaining. *PLoS One, 7*, e51595.

Fehr, E., Fischbacher, U., & Gachter, S. (2002). Strong reciprocity, human cooperation and the enforcement of social norms. *Human Nature, 13*, 1–25.

Ferguson, E., Maltby, J., Bibby, P. A., & Lawrence, C. (2014). Fast to forgive, slow to retaliate: Intuitive responses in the ultimatum game depend on the degree of unfairness. *PLoS One, 9*, e96344.

Gould, R. V. (2000). Revenge as sanction and solidarity display: An analysis of vendettas in nineteenth-century Corsica. *American Sociological Review, 65*, 682–704.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review, 25*, 161–178.

Grecucci, A., Giorgetta, C., Brambilla, P., Zuanon, S., Perini, L., Balestrieri, M., ... Sanfey, A. G. (2013). Anxious ultimatums: How anxiety disorders affect socioeconomic behaviour. *Cognition & Emotion, 27*, 230–244.

Grimm, V., & Mengel, F. (2011). Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters, 111*, 113–115.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*, 271–299.

Gunia, B. C., Wang, L., Huang, L., Wang, J., & Murnighan, J. K. (2012). Contemplation and conversation: Subtle influences on moral decision making. *Academy of Management Journal, 55*, 13–33.

Halali, E., Bereby-Meyer, Y., & Meiran, N. (2011). *When rationality and fairness conflict: The role of cognitive-control in the ultimatum game.* (*Available at SSRN 1868852*).

Halford, G. F., Bain, J. D., & Mayberry, M. T. (1984). Does a concurrent memory load interfere with reasoning? *Current Psychology, 3*, 14–23.

Heller, S. B., Pollack, H. A., Ander, R., & Ludwig, J. (2013). *Preventing youth violence and dropout: A randomized field experiment.* The National Bureau of Economic Research.

Heller, S. B., Shah, A., Guryan, J., Ludwig, J., Mullainathan, S., & Pollack, H. A. (2015). *Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago.* National Bureau Of Economic Research (Working Paper 21178).

Keysar, B., Converse, B., Wang, J., & Epley, N. (2008). Reciprocity is not give and take: Asymmetric reciprocity to positive and negative acts. *Psychological Science, 19*, 1280–1286.

Kirk, D., Gollwitzer, P. M., & Carnevale, P. J. (2011). Self-regulation in ultimatum bargaining: Goals and plans help accepting unfair but profitable offers. *Social Cognition, 29*, 528.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining Prisoner's Dilemma game moves. *Personality and Social Psychology Bulletin, 30*, 1175–1185.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology, 46*, 551–556.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115*, 482–494.

Logan, G. D. (1979). On the use of a concurrent memory load to measure attention and automaticity. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 189–207.

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science, 341*(6149), 976–980.

Mischkowski, D., Kross, E., & Bushman, B. (2012). Flies on the wall are less aggressive: Self-distancing in the heat of the moment reduces aggressive thoughts, angry feelings and aggressive behavior. *Journal of Experimental Social Psychology, 48*, 1187–1191.

Mook, D. G. (1983). In defense of external validity. *American Psychologist, 38*(4), 379–387.

Neo, S. W., Yu, M., Weber, R. A., & Gonzalez, C. (2013). The effects of time delay in reciprocity games. *The Journal of Economic Psychology, 34*, 20–35.

Nikiforakis, N. (2008). Punishment and counter-punishment in Public Goods Games: Can we still govern ourselves? *Journal of Public Economics, 92*, 91–112.

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes, 68*, 208–224.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science, 27*, 1192–1206 (Forthcoming).

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*, 427–430.

Roxβnagel, C. (2000). Cognitive load and perspective-taking: Applying the automatic-controlled distinction to verbal communication. *European Journal of Social Psychology, 30*(3), 429–445.

Shah, A., Mullainathan, S., & Shafir, E. (2012). Some consequences of having too little. *Science, 338*(6107), 682–685.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science, 23*, 1264–1270.

Smith, P., & Silberberg, A. (2010). Rational maximizing by humans (homo sapiens) in an ultimatum game. *Animal Cognition, 13*, 671–677.

Tedeschi, J. T., & Felson, R. B. (1994). *Violence, aggression, and coercive actions.* Washington DC: American Psychological Association.

Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., ... Marois, R. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Neuroscience, 17*, 1270–1275.

Wardlow, L. (2013). Individual differences in speakers' perspective taking: The roles of executive control and working memory. *Psychonomic Bulletin & Review, 20*(4), 766–772.

Winking, J., & Mizer, N. (2013). Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior, 34*, 288–293.

Wolf, S. T., Cohen, T. R., Kirchner, J. L., Rea, A., Montoya, R. M., & Insko, C. A. (2009). Reducing intergroup conflict through the consideration of future consequences. *European Journal of Social Psychology, 39*, 831–841.

Yamagishi, T., Li, Y., Takagishi, H., Matsumoto, Y., & Kiyonari, T. (2014). In search of Homo economicus. *Psychological Science, 25*, 1699–1711.