Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions.

As Published: 10.1037/PSPI0000186

Publisher: American Psychological Association (APA)

Persistent URL: https://hdl.handle.net/1721.1/136438

Version: Original manuscript: author's manuscript prior to formal peer review

Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions

Jillian J. Jordan and David G. Rand

Citation: Jordan, J. J., & Rand, D. G. (2019). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology.* http://dx.doi.org/10.1037/pspi0000186

Abstract

Moralistic punishment can confer reputation benefits by signaling trustworthiness to observers. But why do people punish even when nobody is watching? We argue that people often rely on the heuristic that reputation is *typically* at stake, such that reputation concerns can shape moral outrage and punishment even in one-shot anonymous interactions. We then support this account using data from Amazon Mechanical Turk. In anonymous experiments, subjects (total n = 8440) report more outrage in response to others' selfishness when they cannot signal their trustworthiness through direct prosociality (sharing with a third party)—such that if the interaction were not anonymous, punishment would have greater signaling value. Furthermore, mediation analyses suggest that sharing opportunities reduce outrage by decreasing reputation concerns. Additionally, anonymous experiments measuring costly punishment (total n = 6076) show the same pattern: subjects punish more when sharing is not possible. And importantly, moderation analyses provide some evidence that sharing opportunities do not merely reduce outrage and punishment by inducing empathy towards selfishness or hypocrisy aversion among non-sharers. Finally, we support the specific role of *heuristics* by investigating individual differences in deliberateness. Less deliberative individuals (who typically rely more on heuristics) are more sensitive to sharing opportunities in our anonymous punishment experiments, but, critically, *not* in punishment experiments where reputation *is* at stake (total n = 3422); and not in our anonymous outrage experiments (where condemning is costless). Together, our results suggest that when nobody is watching, reputation cues nonetheless can shape outrage and—among individuals who rely on heuristics—costly punishment.

Keywords: signaling, third-party punishment, morality, trustworthiness, anger

Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions

Moralistic punishment is a central feature of human nature. Humans react to a wide range of selfish and immoral behaviors with condemnation, and act to punish transgressors—even as third-party observers who have not been directly harmed. Such third-party punishment (TPP) appears universal across cultures (Henrich et al., 2010a; Henrich et al., 2006; Herrmann, Thöni, & Gächter, 2008), has early roots in development (Hamlin, Wynn, Bloom, & Mahajan, 2011; Jordan, McAuliffe, & Warneken, 2014; McAuliffe, Jordan, & Warneken, 2015), is observed in both lab (Fehr & Fischbacher, 2004; FeldmanHall, Sokol-Hessner, Van Bavel, & Phelps, 2014; Goette, Huffman, & Meier, 2006; Jordan, McAuliffe, & Rand, 2015) and field (Balafoutas & Nikiforakis, 2012; Mathew & Boyd, 2011) experiments, and is unique to humans (Jensen, 2010; Riedl, Jensen, Call, & Tomasello, 2012).

Moreover, theoretical research demonstrates that punishment can serve to promote and maintain prosocial behavior by deterring selfishness (Boyd, Gintis, Bowles, & Richerson, 2003; Boyd & Richerson, 1992; Henrich & Boyd, 2001), and empirical evidence supports this claim (Balafoutas, Grechenig, & Nikiforakis, 2014; Balliet, Mulder, & Van Lange, 2011; Charness, Cobo-Reyes, & Jimenez, 2008; Feinberg, Willer, & Schultz, 2014; Jordan et al., 2015; Mathew & Boyd, 2011; Yamagishi, 1986). Thus, moralistic punishment plays a critical role in shaping human morality and supporting prosocial behavior.

However, punishing wrongdoing can be costly. It can take time and effort, and risk physical harm and (physical or non-physical) retaliation (Balafoutas et al., 2014; Dreber, Rand, Fudenberg, & Nowak, 2008; Nikiforakis, 2008). So why do unaffected third parties respond to moral transgressions with condemnation and punishment? While a large body of research has

investigated the "proximate" psychological drivers of moralistic punishment (e.g., Carlsmith, Darley, & Robinson, 2002; Cushman, Dreber, Wang, & Costa, 2009; Haidt, 2001; Horberg, Oveis, Keltner, & Cohen, 2009; Nelissen & Zeelenberg, 2009), in this paper we focus on the "ultimate-level" question of *why* our psychology should drive us to incur costs to punish wrongdoing. In other words, what material benefits might moralistic punishment confer in the long run, such that it is supported by learning or evolutionary processes?

Reputation mechanisms for punishment

A large body of work has investigated mechanisms through which TPP can, in the long-run, be strategically beneficial. Much of this work has focused on mechanisms through which punishment can confer reputational benefits (Kurzban, DeScioli, & O'Brien, 2007). These mechanisms include indirect reciprocity, whereby punishers are rewarded (Ohtsuki, Iwasa, & Nowak, 2009; Raihani & Bshary, 2015b), and signaling, whereby punishers advertise either their prosociality (Barclay, 2006; Horita, 2010; Jordan, Hoffman, Bloom, & Rand, 2016a; Nelissen, 2008; Raihani & Bshary, 2015a) or their willingness to retaliate when harmed *directly* (Delton & Krasnow, 2017; Krasnow, Delton, Cosmides, & Tooby, 2016).

One particular way that TPP may confer reputational benefits is by serving as a *costly signal* (Zahavi, 1975) of trustworthiness (i.e., an individual's propensity to reciprocate cooperation from others). This account is supported by game theoretic modeling that proceeds from the premise that the same mechanisms (e.g., reciprocity, institutions) incentivize people to both (i) cooperate themselves, and (ii) encourage *others* to cooperate by punishing selfishness (Jordan et al., 2016a; Jordan & Rand, 2017). As a result, individuals who face larger incentives to cooperate *also* face larger incentives to punish, such that punishing is less costly for them. TPP can therefore serve as a costly signal that the punisher can be trusted to cooperate.

Critically, however, this theory predicts that punishment should only provide a meaningful window into an individual's underlying trustworthiness in the absence of more informative signals. For example, what would happen if a potential punisher *also* had the opportunity to signal his or her trustworthiness via a "direct" act of prosociality, like helping somebody by sharing a resource with them? If punishment signals trustworthiness because similar incentives encourage both punishment and reciprocal cooperation, we might expect sharing a resource to be an even *stronger* signal of trustworthiness than punishment. The prosociality-promoting mechanisms that incentivize both punishment and reciprocal cooperation should *also* incentivize resource sharing. And typically, the incentive structures underlying resource sharing and reciprocal cooperation may be *more* tightly linked than the incentive structures underlying punishment and reciprocal cooperation, because sharing and reciprocal cooperation (but not punishment) both involve paying a cost to directly benefit another individual.

Relatedly, resource sharing may be a "purer" signal of trustworthiness than punishment. This form of direct helping is unambiguously prosocial—whereas while punishment encourages others to cooperate, it also harms the punished and can thus reflect antisocial or spiteful motivations (Herrmann et al., 2008), or seem wrong or aversive under certain moral frameworks (Baron & Ritov, 1993). Consequently, there are good reasons to expect resource sharing to be a stronger signal of trustworthiness than punishment.

As such, the opportunity to share a resource should undermine the signaling value of punishment. And notably, this should be true both for individuals who *do* and do *not* actually choose to share. After an individual chooses *to* share, she should be perceived as quite trustworthy by others—even if she declines to punish. And after an individual chooses *not* to

share, he should be perceived as quite untrustworthy by others—even if he does punish. Thus, in both cases, the marginal signaling benefit of punishment should decline after a sharing opportunity.

Indeed, experimental evidence supports key predictions of this costly signaling theory. Specifically, when punishment is the only available signal, it is perceived as (Barclay, 2006; Horita, 2010; Jordan et al., 2016a; Nelissen, 2008), and actually is (Jordan et al., 2016a), an honest and reliable signal of trustworthiness. But when potential punishers also have the opportunity to directly help others by sharing a resource with them, the perceived and actual signaling value of punishment declines dramatically (while sharing is perceived as, and actually is, a very strong signal of trustworthiness) (Jordan et al., 2016a). And, most critically, potential punishers are less likely to punish when sharing is possible (i.e., when a more informative signal is available). In other words, rates of punishment are influenced not merely by the transgression itself, but also by the value of punishment as a signal of trustworthiness.

Thus, there is clear evidence that people strategically use TPP to build their reputations. However, a framework that merely views moralistic punishment as a "strategic", reputation-focused phenomenon seems limited in many ways. First, beyond enacting punishment, people often respond to wrongdoing by experiencing genuine *moral outrage*. Moral outrage is often discussed primarily as an affective reaction to wrongdoing, consisting of moralistic anger towards the transgressor (Batson et al., 2007; Haidt, 2003; M. L. Hoffman, 2001; Montada & Schneider, 1989), but other discussions of moral outrage also consider cognitive (e.g., beliefs that the transgressor has bad moral character) and behavioral (e.g., a drive towards or support for punishing the transgressor) components (Fiske & Tetlock, 1997; Tetlock, Kristel, Elson, Green, & Lerner, 2000). And while outrage is proposed to serve the ultimate function of motivating

punishment (Carlsmith et al., 2002; Darley & Pittman, 2003; Fessler & Haley, 2003; Fiske & Tetlock, 1997; Goldberg, Lerner, & Tetlock, 1999; Jordan et al., 2015), from a proximate psychological perspective, it is notable that moral emotions and judgements are usually not caused by reasoning (Haidt, 2001) and do not feel "strategic". Rather, introspection suggests that outrage feels like a private response to immorality that simply tracks the magnitude of wrongdoing that has occurred—and certainly does not shut off in contexts where there is no opportunity for punishment to confer reputation benefits.

Moreover, people sometimes punish wrongdoing even in contexts where punishment cannot confer reputation benefits (Crockett, Özdemir, & Fehr, 2014; Fehr & Fischbacher, 2004; Jordan et al., 2015; Nelissen & Zeelenberg, 2009). In other words, people punish in contexts where there are no observers who can link their behavior to their identity and may interact with them in the future (or transmit information about their behavior to someone who will interact with them in the future). Throughout this paper, we refer to these contexts as "one-shot anonymous interactions", or interactions where "reputation is not at stake".

On first inspection, it may seem that because reputation is not at stake, the ultimate explanation for punishment in these contexts cannot involve reputation. However, in this paper, we challenge this idea. We argue that reputation theories are *not* exclusively relevant to moralistic punishment and outrage in contexts where reputation is at stake. Rather, we argue that because people often rely on the heuristic that reputation is *typically* at stake, a reputation framework that incorporates heuristics—while drawing on the theory that punishment serves to signal trustworthiness—can shed light on when and why people experience outrage and enact punishment even in one-shot anonymous interactions.

A reputation heuristics hypothesis for one-shot anonymous punishment

Our work is based on the premise that it is a good rule of thumb to behave, by default, as if reputation is at stake (i.e., as if your behavior will be observed and linked to your identity, influencing the way that others treat you in the future). One reason such an approach could be optimal is "error management" (Delton, Krasnow, Cosmides, & Tooby, 2011), although see (Zefferman, 2014; Zimmermann & Efferson, 2017). According to this account, even if one attempts to evaluate whether reputation is at stake and concludes that it likely is not, there is still uncertainty—and so it may pay to nonetheless behave as if reputation is at stake. However, our work is based on a *different* reason it may pay to behave, by default, as if reputation is at stake. This reason is not that it is optimal to behave as if reputation is at stake after determining that it appears not to be, but rather that it can sometimes be optimal *not to evaluate* whether reputation appears to be at stake. In other words, it can pay to sometimes rely on the *heuristic* that reputation is typically at stake.

In human social life, reputation is frequently at stake, and determining whether the current situation is an exception may be effortful (e.g., even when nobody seems to be watching, one may need to evaluate whether there are hidden observers). Consequently, evaluating whether reputation is at stake can have cognitive costs (such evaluation takes time and effort (Bear & Rand, 2016; Kahneman, 2011; D. G. Rand, Tomlin, Bear, Ludvig, & Cohen, 2017)) as well as social costs (those who appear calculating in their moral decision-making may be seen negatively by others (Critcher, Inbar, & Pizarro, 2013; M. Hoffman, Yoeli, & Nowak, 2015; Jordan, Hoffman, Nowak, & Rand, 2016b)). To avoid these costs, it may be beneficial to rely on

the heuristic that reputation is typically at stake instead of constantly calculating whether this is currently the case (Bear, Kagan, & Rand, 2017; Bear & Rand, 2016).

If (some) people employ the heuristic that reputation is typically at stake, the theory that punishment serves to signal trustworthiness may help explain when and why people punish in contexts where reputation is not at stake. Specifically, a reputation heuristics hypothesis makes the prediction that even in one-shot anonymous interactions, people's punishment decisions should be sensitive to cues of the *potential* signaling value of punishing (if reputation *were* at stake)—and that this sensitivity should be greater among less deliberative decision-makers, who should be more prone to rely on heuristics. Moreover, because outrage is proposed to adaptively motivate punishment (despite being experienced genuinely), a reputation heuristics hypothesis may predict that when reputation is not at stake, outrage also increases in contexts where punishment *would* confer larger reputation benefits if reputation *were* at stake.

An illustrative example

To illustrate our reputation heuristics hypothesis, imagine the following example. One day, your workplace holds a fundraiser for a local charity that fights homelessness. To collect funds, they ask for donations in the break room during lunchtime. However, you happen to be in a meeting when the funds are collected, so you have no opportunity to donate. Afterwards, a well-off colleague tells you, and several other colleagues, that he thinks homeless people are lazy and makes it a rule to never help them.

How outraged do you feel, and how likely are you to chastise your colleague? A signaling theory predicts that in general, condemning him could confer reputation benefits by demonstrating to other colleagues that *you* are not selfish, and do not have disdain for the homeless. It also predicts that in this particular situation, you might be especially driven to

punish. Because you were out of the room when donations were collected, you were unable to donate to the charity—and consequently, you missed an opportunity to send a more direct signal that you are not selfish and have positive attitudes towards the homeless. Thus, the signaling value of punishing may be especially high, as compared to the counterfactual in which you had the opportunity to donate.

In this example, punishing would be observed by a host of people you know, and could therefore confer genuine reputational benefits. But now consider the case where after missing the opportunity to donate, you leave the office and see a stranger insult a homeless person on the street. If you choose to chastise the stranger, you will *not* be observed by people you know, and thus will not actually gain reputation benefits. However, insofar as you behave by default as if reputation is at stake, your reaction might nonetheless be influenced by reputation-relevant cues. Specifically, your reaction might be influenced by the fact that you missed the opportunity to donate to the office fundraiser—so punishing the stranger *would* serve as a relatively strong signal of your morality if you *were* observed by somebody from work. We thus predict that even in this anonymous context, you might feel heightened outrage, and be more likely to chastise the stranger (as compared to the counterfactual in which you were present when donations were solicited).

Overview of analyses

To test our reputation heuristics account of outrage and punishment in one-shot anonymous interactions, we used five analyses of twelve different experiments. See Table 1 for a summary of our analyses, and Table 8 for a summary of the experiments included in them.

Across our first two analyses, we began by testing the prediction that moral outrage is sensitive to reputation cues in contexts where reputation is not at stake. In Analysis 1, we

investigated seven experiments that measured moral outrage in one-shot anonymous interactions (total n = 8440). (Six experiments measured outrage using a three-item scale designed to tap the affective, cognitive, and behavioral components of outrage, and one used a single item designed to tap only the affective component of outrage). We tested the prediction that outrage would be greater in contexts where punishment *would* serve as effective signal of trustworthiness, if observed. Specifically, we predicted that subjects would report more outrage towards selfishness when they could not signal their trustworthiness via direct helping (sharing a resource with a third party)—and thus if punishment were observed, it would have greater signaling value.

In Analysis 2, we tested the prediction that helping opportunities influence reported outrage via reputation concerns. To this end, we investigated mediation through two reputation-relevant constructs. First, in one subset of our outrage experiments (n = 2434), we measured the perceived reputation benefits of punishment. This construct was intended to be a mediator, and we predicted that (i) when subjects did not have the opportunity to help, they would report that punishment would have greater reputations benefits, (ii) the perceived reputation benefits of punishment would correlate positively with outrage, and (iii) the perceived reputation benefits of punishment would mediate the effect of helping opportunities on outrage. Second, in a partially-overlapping subset of our outrage experiments (n = 2432), we measured general reputation concerns. This construct was initially intended to be a moderator; however, it was measured after our helping opportunities manipulation and we found evidence that it was influenced by helping opportunities, so we analyzed it as a mediator. We thus investigated whether (i) subjects reported being more generally concerned with their reputations when they did not have the opportunity to help, (ii) general reputation concerns correlated positively with outrage, and (iii) general reputation concerns mediated the effect of helping opportunities on outrage.

In Analysis 3, we tested the prediction that helping opportunities *also* influence costly punishment in contexts where reputation is not at stake. We investigated a set of four experiments that measured costly punishment in one-shot anonymous interactions (total n = 6076). We predicted that subjects would be more likely to punish when they did not have the opportunity to help.

In Analysis 4, we tested the deflationary hypothesis that helping opportunities reduce outrage and punishment merely by inducing empathy towards selfishness or hypocrisy aversion among subjects who decline to help. This deflationary hypothesis predicts that helping opportunities only reduce outrage and punishment among subjects who choose *not* to help when given the opportunity, and not among subjects who choose *to* help. In contrast, our signaling hypothesis predicts that helping opportunities should reduce outrage and punishment among all subjects, regardless of whether or not they choose to help when given the opportunity.

To test our signaling hypothesis, we sought to tap individual differences in the likelihood of helping, when given the chance. To this end, after one of our experiments (which manipulated helping opportunities and measured affective outrage and punishment), we conducted a follow-up experiment that gave all subjects the opportunity to help. We treated follow-up experiment helping as an index of an individual's propensity to help when given the chance. Then, we tested (i) whether follow-up experiment helping moderated the effects of helping opportunities on affective outrage and punishment in our original experiment, and (ii) if so, whether these effects were driven solely by follow-up experiment non-helpers, or also held among follow-up experiment helpers. We predicted that the negative effects of helping opportunities on affective outrage and punishment would not be driven solely by non-helpers.

Together, Analyses 1-4 tested the predictions that moral outrage and costly punishment are influenced by the potential signaling value of punishment, even when reputation is not at stake. Finally, in Analysis 5, we specifically tested our reputation *heuristics* explanation for these predictions. Based on the premise that deliberative individuals tend to rely more on heuristics, we investigated the potential moderating role of deliberativeness. We did so by investigating two indicators of deliberativeness: performance on questions assessing comprehension of the incentives in our experiment, and performance on the Cognitive Reflection Task (Frederick, 2005).

As per our reputation heuristics hypothesis, we predicted that less deliberative subjects would be more likely to enact one-shot anonymous punishment when helping was not possible, while more deliberative subjects would punish at relatively lower rates regardless of helping opportunities. Moreover, we predicted that that deliberativeness would *not* moderate the influence of helping opportunities on punishment in a set of experiments (total n = 3422) where reputation *was* actually at stake, and thus attending to reputation cues actually had strategic value. Finally, we also investigated whether deliberativeness would moderate the effect of helping opportunities on outrage in our one-shot anonymous outrage experiments.

		Experiments included											
Analysis	**Key questions and predictions**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	When reputation is not at stake, how do helping opportunities influence outrage, and vice versa? -Predict: helping opportunities reduce outrage (Exps 1-7) -Predict: the opportunity to rate outrage does not reduce helping (Exp 1)	X	X	X	X	X	X	X					
2	When reputation is not at stake, do two reputation-relevant constructs mediate the effect of helping opportunities on outrage? -Predict: the Perceived Reputation Benefits of Punishment (PRBP) mediate the effect of helping opportunities on outrage (Exps 2,4,5) -Investigate whether General Reputation Concerns (GRC) mediate the effect of helping opportunities on outrage (Exps 3-5)		X	X	X	X							
3	When reputation is not at stake, how do helping opportunities influence punishment, and vice versa? -Predict: helping opportunities reduce punishment (Exps 6, 8-10) -Predict: punishment opportunities do not reduce helping (Exps 8-9)						X		X	X	X		
4	When reputation is not at stake, does follow-up experiment helping moderate the effects of helping opportunities on affective outrage and						X						

	punishment? And if so, are these effects driven solely by non-helpers, or do they also hold among helpers? -Predict: the negative effects of helping opportunities on affective outrage and punishment are not driven solely by non-helpers (Exp 6)												
5a	When reputation is not at stake, does deliberativeness moderate the effect of helping opportunities on punishment? -Predict: Deliberativeness attenuates the effect of helping opportunities on punishment (Exps 6, 8-10)						X		X	X	X		
5b	When reputation is at stake, does deliberativeness moderate the effect of helping opportunities on punishment? -Predict: Deliberativeness does not attenuate the effect of helping opportunities on punishment (Exps 9-12)									X	X	X	X
5c	When reputation is not at stake, does deliberativeness moderate the effect of helping opportunities on punishment? -Explore this question without a directional prediction (Exps 1-7)	X	X	X	X	X	X	X					

Table 1. Overview of analyses. For each analysis, we report the key questions and predictions, and the experiments included.

Analysis 1

In Analysis 1, we tested the prediction that moral outrage is influenced by cues of the potential signaling value that punishment. To this end, we considered seven experiments investigating whether people respond to selfishness with more moral outrage in situations where they lack the opportunity to directly help others.

As discussed previously, there are theoretical reasons that direct helping should typically be a stronger signal of trustworthiness than punishment. And indeed, empirical evidence from a context where reputation *is* at stake suggests that the expected signaling value of punishment is larger when helping is not possible (and thus a better signal is not available) (Jordan et al., 2016a). Moreover, Jordan et al. find that helping opportunities reduce punishment (as predicted by the observation that helping is a stronger signal than punishment), while punishment opportunities do *not* reciprocally reduce helping (as predicted by the observation that punishment is a weaker signal than helping).

When designing the seven experiments analyzed here, we adapted the design of this previous work to test the hypothesis that moral outrage is sensitive to cues of punishment's potential signaling value. Across all seven experiments, we tested the prediction that helping

opportunities would decrease *moral outrage* (rather than punishment) in a context where reputation was *not* actually at stake. Additionally, we tested the prediction that the opportunity to express moral outrage would not reciprocally decrease helping.

Methods

Design. We designed a "Third-Party Condemnation Game" (TPCG), which we used in all seven experiments. The TPCG had three players, and involved an incentivized economic game decision with no deception. In this game, subjects had the opportunity to earn money that was paid out in a "bonus payment", on top of the show-up fee they earned for participating. Specifically, one subject (the Helper) was endowed with money (30¢) and decided whether or not to split it evenly with (i.e., help) another subject (the Recipient). Then, a third subject (the Condemner) rated their moral outrage towards the Helper. (Specifically, we always measured outrage towards a *selfish* helper; see Procedure for details.) The TPCG met our definition of a one-shot anonymous interaction, in which reputation was not at stake. It was conducted online in privacy, with anonymous strangers, and there was no potential for any of the players to base their game play on other players' past actions. Moreover, while we (i.e., the experimenters) could observe subjects' responses, we could not link them to subjects' identities. Thus, there was no strategic reason for subjects to care about how their responses were perceived by others.

In all seven experiments, target subjects read about all roles in the TPCG, and then we manipulated the role(s) they were assigned to play. In the *Condemnation Only* condition, we assigned target subjects to play the TPCG once, in the role of the Condemner. In the *Condemnation+Helping* condition, by contrast, we assigned target subjects to play twice, with two different sets of other players: once in the role of Condemner, and once in the role of Helper.

While our experiments were anonymous, what would happen if target subjects in these conditions were actually being judged by observers? In the Condemnation+Helping condition, an observer would have access to a very strong signal of a target's trustworthiness: whether or not the target chose to help. Therefore, if the observer were to also find out whether the target punished selfishness, we would expect this second (weaker) signal to have limited influence on the observer's judgement. In contrast, in Condemnation Only, the observer would not have information about target helping, so we would expect punishment to carry more weight. Thus, despite the fact that our experiments were anonymous, helping opportunities served to undermine the *potential* signaling value that punishment *could* confer if observed. We predicted that this would influence outrage—such that subjects in Condemnation Only would report more outrage than subjects in Condemnation+Helping.

We note that, importantly, when target subjects participated as the Condemner, they rated their outrage towards a Helper who had behaved selfishly towards a Recipient. In contrast, when target subjects participated as the Helper, they decided whether to help a Recipient who had *no* previous experience with the game. In other words, they decided whether to help a completely neutral party—they were *not* paired with a Recipient who had previously been mistreated in any way. Thus, while Condemners had the opportunity to express outrage in response to a selfish transgression, Helpers were not reacting to injustice or compensating victims. As such, our experimental design falls outside the purview of the moral psychology literature on compensation versus punishment as modes of restorative justice (e.g., Darley & Pittman, 2003; Gromet, Okimoto, Wenzel, & Darley, 2012; Lotz, Okimoto, Schlösser, & Fetchenhauer, 2011).

While the above-described design was constant across our seven outrage experiments, some details varied (see Table 8 for an overview of differences). First, in Experiment 1, to rule

out the possibility that subjects might express less outrage in Condemnation+Helping simply because they had two different response options, we also included a Helping Only condition (in which we assigned target subjects to play the TPCG once, in the role of Helper). We predicted that while helping opportunities would attenuate reported outrage, condemnation opportunities would *not* reciprocally attenuate rates of helping. In other words, we predicted that rates of helping would be similar in the Helping Only and Condemnation+Helping conditions. As described below, we found support for this prediction; thus, in Experiments 2-7, we focused only on the effect of helping opportunities on outrage, and did not include Helping Only conditions.

Relatedly, in Experiment 1, we counterbalanced the order in which subjects in Condemnation+Helping made their Condemner and Helper decisions. (Subjects in Condemnation+Helping always knew that they would make both a Condemner and a Helper decision, but we randomized the order in which these decisions were made). This counterbalancing allowed for a symmetrical test of the effect of helping opportunities on outrage, as compared to the effect of condemnation opportunities on helping. In contrast, in Experiments 2-7, we always assigned subjects in Condemnation+Helping to make their Helper decisions before their Condemner decisions. This fixed order was used to increase the salience of helping opportunities, given our exclusive focus on the effect of helping opportunities on outrage. (For analyses of order effects within the Condemnation+Helping condition of Experiment 1, see SM.)

Additionally, Experiments 2-5 investigated the mechanism behind the effect of helping opportunities on outrage by measuring two candidate mediators (see Analysis 2 for more details).

Next, Experiment 6 differed from Experiments 1-5 in several ways. In Experiments 1-5, we framed the outrage-rating task as “making a judgement about the Helper’s moral character”, asked subjects to complete this task *imagining* that the Helper chose not to help (without

knowing what the Helper actually did), and measured outrage using a three-item scale designed to tap the affective, cognitive, and behavioral components of outrage; in Experiment 6, we framed the outrage-rating task more neutrally, told subjects that the Helper chose not to help, and measured outrage with a single item designed to tap only the affective component of outrage (see Procedure for details). Additionally, unlike Experiments 1-5, Experiment 6 (i) administered the Cognitive Reflection Task before assigning subjects to an experimental condition (see Analysis 5 for more details), (ii) after measuring outrage, administered a filler task and then measured costly punishment (see Analysis 3 for details), (iii) added a few post-experimental questions (see Procedure and Analysis 2 for details); and (iv) approximately two weeks after completing data collection, conducted a follow-up experiment that Experiment 6 subjects were invited to complete (see Analysis 4 for details).

Finally, in Experiment 7, we returned to our procedure from Experiments 1-5, and thus used our three-item outrage scale, and framed our outrage-rating task as "making a judgement about the Helper's moral character"; however, as in Experiment 6, we again told subjects that the Helper chose not to help (as opposed to asking them to imagine that the Helper chose not to help). We also included a slightly modified version of one of the post-experimental questions included in Experiment 6 (see Procedure for more details).

Subjects. In each of Experiments 1-5, we requested a target of $n = 400$ subjects per condition from Amazon Mechanical Turk (AMT) (i.e., a total of $n = 1200$ subjects in Experiment 1, which included a Helping Only condition, and $n = 800$ subjects in each of Experiments 2-5, which did not). In Experiment 6, we decided (prior to data collection) to request a larger sample size of $n = 1500$ subjects per condition (i.e., a total of $n = 3000$ subjects) for greater power, particularly because this experiment involved inviting subjects to complete a follow-up

experiment (see Analysis 4 for more details) and we were concerned about the potential for low response rates. Finally, in Experiment 7, we decided (prior to data collection) to request a sample of $n = 750$ subjects per condition. We selected this sample size to provide high power to confirm that the effect of helping opportunities on outrage would be comparable to the effect observed in Experiments 1-6, despite Experiment 7 being the only experiment in which we both used our three-item outrage scale and told subjects that the Helper chose not to help (rather than asking them to imagine the Helper not helping).

In our final samples for analysis, we included all subjects who completed all dependent variables and had a unique IP address and AMT ID; when we encountered duplicate IPs or IDs, we included only the observation that was completed chronologically first. This process sometimes resulted in final samples that were slightly larger than the target number requested on AMT (as some subjects completed our survey, but did not indicate this to AMT).

Throughout our paper, we report and plot results from all subjects, regardless of performance on comprehension questions (see Supplementary Materials (SM) for statistics on performance); then, in Analysis 5, we investigate the influence of comprehension on our results. Aggregating across our seven experiments, our final samples have $n = 8847$ subjects ($n = 4228$ in Condemnation Only, $n = 4212$ in Condemnation+Helping, and $n = 407$ in Helping Only), $M_{age} = 35.98$ years, $SD_{age} = 11.66$ years, 43% male. For demographics by experiment, see SM.

Procedure. We began by providing subjects with instructions explaining the TPCG and their role(s) in it (as determined by condition). In Experiments 1-5 and 7, we described the Condemner's role as "making a judgement about the Helper's moral character, in the event that the Helper decides not to help". In contrast, in Experiment 6, we described the Condemner's role more neutrally, as "rating their reaction towards the Helper".

Next, we provided subjects with two comprehension questions evaluating their understanding of the incentive structure of the TPCG helping decision (for all experiments in this paper, see SM for full stimuli). Then, subjects in Helping Only made their helping decisions, subjects in Condemnation Only rated their outrage, and subjects in Condemnation+Helping made both decisions. To measure helping, we reminded subjects that they had 30¢, and that their job was to decide whether to pay 15¢ to share with the Recipient. We then asked them to make a decision, which we subsequently repeated back to them.

To measure moral outrage, we reminded subjects of their role as Condemner (using the language described above). Then, in Experiments 1-5, we instructed subjects to imagine that the Helper decided not to share, and in Experiments 6-7, we told subjects that the Helper did not share. Next, we presented our moral outrage scale. In Experiments 1-5 and 7, we used a three-item scale that we designed to tap the affective, cognitive, and behavioral components of outrage. This scale was conceptually similar to other moral outrage scales designed to tap these three components of outrage (Salerno & Peter-Hagene, 2013; Skitka, Bauman, & Mullen, 2004), and was designed to use language appropriate for the relatively minor transgression our experiments focused on. In our scale, we asked subjects (i) how angry they felt towards the Helper, (ii) how much the Helper deserved to be punished, and (iii) how morally bad the Helper was (in that fixed order); then, we computed moral outrage scores as the average response across our three scale items.

In Experiment 6, we replaced this three-item scale with one item that specifically measured the affective component of outrage. Our goal was to investigate whether our results were robust to a context in which only affective outrage was measured, in order to provide a stronger case for an effect on an affective process, and thus to connect our work to the

psychological literatures on affective outrage, moral emotions, and emotion regulation (e.g., Batson et al., 2007; Brady, Wills, Jost, Tucker, & Van Bavel, 2017; Gross, 1998b; Haidt, 2003; Hutcherson & Gross, 2011; Nelissen & Zeelenberg, 2009; Tangney, Stuewig, & Mashek, 2007). To this end, we presented only the anger item from our three-item scale.

In Experiment 1, Condemners made ratings using Likert scales ranging from 10 to 100 in 10-point increments, with extreme anchors reading *Not at all* and *Very much*. In Experiments 2-7, we modified these scales to range from 0 to 100. Then, to facilitate comparison across experiments, we rescaled Experiment 1 responses (which originally ranged from 10 to 100) by subtracting 10 and then multiplying by 10/9 (such that they ranged from 0 to 100, like in Experiments 2-6). In Experiments 6-7, for grammatical correctness, we changed the wording on the extreme anchor from *Very much* to *A lot.*

After subjects made their decisions, they completed a post-experimental survey including some demographic and other questions. Of relevance to Analysis 1, in both Experiments 6 and 7, we included one post-experimental question investigating subjects' beliefs about whether other players could influence their payoffs. These questions were designed to investigate whether, to the extent that subjects were sensitive to reputation cues in our one-shot anonymous experiments, this reflected a mistaken explicit belief that other players really *could* observe their behavior and then influence their payoffs. Specifically, in Experiment 6, we asked subjects who, if anyone, could influence their payoffs, and provided response options of the Helper, the Recipient, both, and neither; responses of "neither" were considered correct. In Experiment 7, we modified the wording slightly to avoid suggesting to subjects that other player(s) could influence their payoffs. We asked subjects whether, while rating their outrage, they believed that any of the other players had the ability to influence their payoffs, and provided response options of Yes or No; then,

(only) if subjects selected "Yes", we asked them to pick between the response options offered in Experiment 6. Responses of "No" were considered correct.

Finally, after all data was collected, we used ex-post matching to pair Helpers and Recipients and calculate their bonuses.

Results

We begin by noting that all of our data, and a script for reproducing all our analyses, is available online at https://osf.io/7z8b6/.

Next, we report aggregated analyses of moral outrage across Experiments 1-7. These analyses aggregated average responses across our three-item outrage scale in Experiments 1-5 and 7 with responses to our single item affective outrage measure in Experiment 6; however, we subsequently report analyses by experiment to demonstrate robustness across both measures. We note that throughout our analyses, we used linear regressions to predict continuous variables and logistic regressions to predict binary variables, and in all analyses that pooled data from multiple experiments, we included experiment dummies.

Collapsing across our six experiments that used our three-item outrage scale, we found that this scale was reliable ($\alpha = 0.88$). All three items were strongly correlated with each other: anger and deserved punishment ($r = .73$, $p < .001$), anger and badness of person ($r = .72$, $p < .001$), badness of person and deserved punishment ($r = .71$, $p < .001$). Additionally, as predicted and as illustrated in Figure 1a, we found that subjects across Experiments 1-7 reported significantly more outrage in Condemnation Only ($M = 35.18$, $SD = 29.50$) than Condemnation+Helping ($M = 30.22$, $SD = 30.06$), $B = 0.08$, $t = 7.68$, $p < .001$, $n = 8440$. Thus, when subjects had the opportunity to signal their trustworthiness via direct helping, they reported less outrage in response to selfish behavior.

As predicted and illustrated in Figure 1b, conversely, we observed comparable rates of helping in Helping Only (66%) and Condemnation+Helping (67%) in Experiment 1 (which included the Helping Only control condition), $OR = 0.94$, $z = -.39$, $p = .693$, $n = 797$. Thus, while helping opportunities reduced outrage across Experiments 1-7, condemnation opportunities did not reduce helping in Experiment 1. We also bolster this conclusion by directly comparing the effect of helping opportunities on outrage to the effect of condemnation opportunities on helping within Experiment 1. Investigating only this experiment, we used linear regressions to predict both outrage and helping as a function of condition, and found that the standardized condition coefficient was significantly larger when predicting outrage ($B = .10$, $SE = .04$, $p = .006$) than helping ($B = -.01$, $SE = .04$, $p = .694$), $z = 2.24$, $p = .025$.

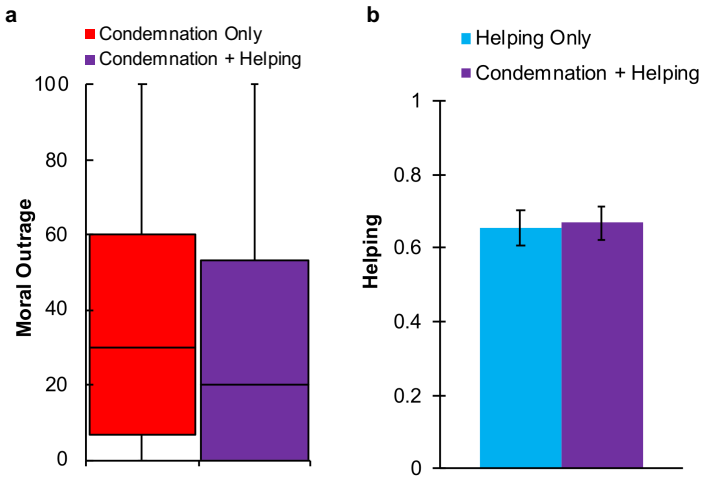

Figure 1. Helping opportunities reduce moral outrage (while condemnation opportunities do not reduce helping) in one-shot anonymous interactions. In **a,** we show box plots (drawing lines at the 25th, 50th, and 75th percentiles, and illustrating the minimum and maximum values) for outrage as a function of helping opportunities across Experiments 1-7. In **b,** we plot the proportion of subjects helping as a function of condemnation opportunities in Experiment 1; error bars are 95% CIs.

Next, we report our results by experiment. In Table 2 we report, for each experiment and overall, the reliability of, and effect of helping opportunities on, our three-item outrage scale (measured in Experiments 1-5 and 7), as well as the effect of helping opportunities on our affective outrage item (anger) specifically (measured in all seven experiments).

Our results are quite robust across experiments: in all six experiments measuring outrage via our three-item scale, we found that the scale was reliable, and in five out of those six, we observed significantly more outrage in Condemnation Only than Condemnation+Helping. Importantly, this effect was significant in Experiment 7, demonstrating that the effect of helping opportunities on outrage was robust to telling Condemners that the Helper did not share (rather than asking them to imagine the Helper not sharing). Additionally, across all seven experiments, we always observed directionally more affective outrage in Condemnation Only, and this effect was significant in three experiments and overall, as well as marginally significant in two experiments. Importantly, the effect was significant in Experiment 6, providing further evidence that the effect of helping opportunities on outrage was robust to telling Condemners that the Helper did not share, and also demonstrating that it was robust to framing the Condemner's role more neutrally, and measuring affective outrage only.

Statistic	**Exp. 1** $n = 788$	**Exp. 2** $n = 819$	**Exp. 3** $n = 817$	**Exp. 4** $n = 811$	**Exp. 5** $n = 804$	**Exp. 6** $n = 2924$	**Exp. 7** $n = 1477$	**Aggregate** (n varies)
Reliability of three-item outrage scale	.88	.88	.86	.88	.88	N/A	.90	.88 (Exps 1-5, 7, $n = 5516$)
Effect of Condemnation Only dummy on three-item outrage scale	$B = .10$ $p = .006$	$B = .07$ $p = .036$	$B = .04$ $p = .274$	$B = .09$ $p = .011$	$B = .12$ $p = .001$	N/A	$B = .08$ $p = .001$	$B = .08$ $p < .001$ (Exps 1-5, 7, $n = 5516$)
Effect of Condemnation Only dummy on affective outrage (anger item)	$B = .07$ $p = .050$	$B = .05$ $p = .123$	$B = .02$ $p = .635$	$B = .06$ $p = .094$	$B = .09$ $p = .012$	$B = .09$ $p < .001$	$B = .06$ $p = .031$	$B = .07$ $p < .001$ (Exps 1-7, $n = 8440$)

Table 2. Analysis 1 results, by experiment. Sample sizes indicate subjects for whom outrage was measured (i.e., the Experiment 1 sample size excludes the Helping Only condition).

Finally, we consider a deflationary account of our results. Were our subjects sensitive to helping opportunities in our one-shot anonymous experiments simply because they held the mistaken explicit belief that other players really *could* observe their behavior and then influence their payoffs? To address this question, we analyzed responses to the post-experimental questions in Experiments 6 and 7 that measured subjects' explicit beliefs about whether other players could influence their payoffs. Excluding subjects who reported that other player(s) *could* influence their payoffs, we still found that helping opportunities reduced affective outrage in Experiment 6 ($B = 0.09$, $t = 3.74$, $p < .001$, $n = 1703$) and outrage in Experiment 7 ($B = 0.08$, $t = 2.86$, $p = .004$, $n = 1227$). Thus, our results do not seem to have been driven by subjects who held the mistaken explicit belief that other players could observe their behavior and influence their payoffs.

Because the one-shot anonymous nature of our experiments is a critical design feature, we also examined the absolute percentage of subjects who correctly indicated that other players could not influence their payoffs. We found that 58.24% of subjects answered correctly in Experiment 6, while 83.07% answered correctly in Experiment 7. At face value, the percentage of correct responses in Experiment 6 seems worryingly low. However, given the Experiment 6 question wording, and the Experiment 7 result, we suspect that this percentage over-estimates the frequency with which Experiment 6 subjects, while rating their outrage and making their punishment decisions, actually believed that other players could influence their payoffs.

Recall that Experiment 6 asked subjects who, if anyone, could influence their payoffs, and then presented four response options, three of which indicated that other player(s) could

influence their payoffs. We worried that this setup may have suggested to subjects that other players could influence their payoffs, inducing this belief among subjects who had not previously held it while making their outrage and punishment decisions. We also considered that random responses (provided by hurried or inattentive subjects) would be incorrect 75% of the time. For these reasons, in Experiment 7, we asked a "yes or no" question of whether, while rating their outrage, subjects believed that any of the other players could influence their payoffs. Using this wording, we found a substantial increase in correct responding, consistent with the possibility that our Experiment 6 wording was suggestive.

Of course, it is difficult to completely avoid suggestion while measure subjects' beliefs about whether other players could influence their payoffs. Our data cannot decisively reveal the true percentage of subjects in each experiment who held this belief prior to us asking about it. Nonetheless, we see the comprehension rate in Experiment 7 as encouraging, and consistent with our general prior that it would be relatively unlikely for subjects to explicitly believe that other players could influence their payoffs (as this would require confabulating an addition component of the game that did not exist). And regardless of the true comprehension rate, our key hypothesis—that reputation heuristics can shape outrage in one-shot anonymous interactions—is supported by the fact that our results hold among subjects who explicitly understood that other players could not influence their payoffs. Of course, our results may have been driven by subjects who *implicitly* believed that other players could influence their payoffs—a possibility that is consistent with our reputation heuristics hypothesis.

Discussion

Analysis 1 supports our prediction that in one-shot anonymous interactions, subjects who did not have the opportunity to signal their trustworthiness via direct helping—and thus for

whom punishment had greater potential signaling value—reacted to selfishness with more moral outrage. This finding does not appear to reflect a general mechanism whereby any response is less likely when two response options are available: as predicted, while helping opportunities reduced reported outrage, the opportunity to report outrage did not reduce helping.

Our Analysis 1 results are aligned with the previous results that, in a context where reputation *was* actually at stake, helping opportunities reduced rates of costly punishment, while the reverse was not true (Jordan et al., 2016a). Analysis 1 extends this pattern to the context of reported moral outrage in one-shot anonymous interactions.

We note that the observed effect of helping opportunities on outrage was relatively small. However, it is theoretically significant that helping opportunities—a proposed reputation cue—had *any* effect on outrage, given that the transgression in question was identical across conditions. This result provides support for the proposal that, as an adaptive motivator of punishment, moral outrage is not *just* an objective indicator of the magnitude of wrongdoing that has occurred. Rather, despite being experienced genuinely, our data suggest that outrage can be influenced by the potential signaling value of punishment. This conclusion has the implication that in daily life, other reputation cues could also influence outrage—and more broadly, our results support the theory that a reputation framework can shed light on our moral psychology, even in contexts where reputation is not at stake.

However, while our results are consistent with the hypothesis that helping opportunities influenced the subjective experience of moral outrage (i.e., that subjects genuinely felt more morally outraged when helping was not possible), we note that an alternative interpretation of our results is also possible. Specifically, it is possible that subjects who did not have the opportunity to help had the same subjective experience of moral outrage, but were driven to rate

themselves as more morally outraged. In other words, helping opportunities may not have influenced *feelings* of moral outrage, but the drive to *express* those feelings—in this case, via ratings on our moral outrage scale, which subjects may have treated as an opportunity for verbal condemnation (rather than a precise barometer of their subjective experience). It is difficult to discriminate between these possibilities, which are not mutually exclusive. An increase in self-reported outrage can always reflect an increase in the experience of outrage, or the drive to express it—but it is difficult to measure the subjective experience of outrage without self-report.

We do find it notable that helping opportunities reduced reported outrage even in Experiment 6, where we focused on the affective component of outrage (by measuring only anger), and framed the outrage-rating task more neutrally (by telling subjects to "rate their reaction towards" rather than "make a moral judgement about" the Helper). It seems possible that these changes reduced the extent to which subjects viewed our outrage-rating task as an opportunity to verbally condemn selfishness, such that Experiment 6 served as a purer measure of subjects' true affective experience. Nonetheless, it is of course still possible that helping opportunities reduced reported affective outrage in Experiment 6 merely by modifying subjects' drive to express their (unchanged) experience of affective outrage. Future work should seek to differentiate between these possibilities. Even if helping opportunities only influenced expressions of outrage, however, our results would still imply that the basic drive to express outrage—in a context where expressions are completely anonymous—is shaped by reputation cues. Either interpretation thus suggests that a reputation framework can help explain a broad set of expressions of moral outrage and acts of punishment, even when reputation is not at stake.

Analysis 2

In Analysis 2, we aimed to provide more direct support for a reputation-based interpretation of our Analysis 1 results. Specifically, we sought to test the hypothesis that helping opportunities influenced reported outrage *because* they served as a cue of the potential signaling value of punishment. To this end, we conducted mediation analyses to investigate the mechanisms through which helping opportunities influenced reported outrage, and tested the prediction that they did so insofar as they were seen as a reputation-relevant cue.

Two candidate mediators

We investigated two reputation-relevant candidate mediators. First, in Experiments 2, 4, and 5, we measured the *perceived reputation benefits of punishment*. According to our theory, (i) because helping is such a diagnostic signal of trustworthiness, the potential reputation benefits of punishment should decline after a helping opportunity, and (ii) moral outrage (or the drive to express outrage) should be sensitive to the potential reputation benefits of punishment. Thus, helping opportunities should influence outrage—insofar as they are, in fact, seen as relevant to punishment's potential reputation value. In other words, the perceived reputation benefits of punishment should mediate the effect of helping opportunities on outrage.

This pattern could reflect that when helping is possible, outrage declines insofar as people have learned that helping opportunities are a reliable cue that punishment will have limited reputation value. Alternatively, outrage might decline insofar as people respond to helping opportunities by computing, in the moment, that the potential reputation value of punishment is relatively small (Crockett, 2013; Cushman, 2013). Even under this second possibility, it seems unlikely that our subjects would *consciously* compute the reputation value of punishment, or that such reasoning would *consciously* influence outrage: in our outrage experiments, subjects did not actually make punishment decisions, and reputation was not actually at stake. Thus, we saw it as

more likely that helping opportunities would *unconsciously* influence outrage insofar as they were *implicitly* seen as a reliable reputation cue, or influenced *implicit* computations about punishment's reputation value. However, we reasoned that these implicit processes could likely be accessed by explicitly asking subjects to evaluate the reputation value of punishment. We thus directly asked subjects how a hypothetical act of punishment would be perceived by others, and treated this measure as our first candidate mediator.

Our second candidate mediator, which was measured in Experiments 3-5, was the extent to which subjects reported being generally concerned with their reputations (i.e., being a person who tends to desire positive social evaluation, and fear negative social evaluation). We initially intended for this measure to be a *moderator,* and thus selected a scale that was designed to assess the general *trait* of concern with one's reputation across contexts. However, we always collected this measure after our manipulation of helping opportunities, and found some evidence that it was influenced by helping opportunities (with a significant effect observed in Experiment 3, and a marginally significant effect observed in aggregated analyses). Thus, we concluded that it would be inappropriate to treat this measure as a moderator of the effect of helping opportunities.

Moreover, the evidence that our manipulation impacted reported trait reputation concerns suggests that to some degree, subjects were also reporting on their state reputation concerns. Thus, we chose to investigate general reputation concerns as a candidate *mediator*. Thus far, we have proposed that when helping is not possible, outrage is elevated insofar as subjects implicitly see punishment as having more reputation value. But by treating general reputation concerns as a second mediator, we could also ask whether subjects who did not have the chance to help felt more concerned with their reputations, and whether such concerns might have shaped outrage.

Methods

As discussed above, Experiments 2-5 each measured at least one of our candidate mediators. In each of these experiments, we always measured our mediators *after* measuring outrage, and thus avoided activating reputation concepts before measuring outrage. This decision has an important advantage: we can be confident that evidence of mediation does not merely reflect that we induced subjects to think about the reputation value of punishment, or their general reputation concerns, before measuring outrage. However, it also has a disadvantage: responses to our outrage scale could have causally influenced responses to our mediator scales. This possibility is worth keeping in mind in the interpretation of our mediation analyses. However, we note that if we had measured our mediators before measuring outrage, while the act of *reporting* outrage could not have causally affected ratings of our mediators, the outrage subjects *experienced* (prior to being asked to report it) *could* still have causally affected these ratings. Thus, we view a possible causal path from our dependent variable to our mediating variables as an inherent issue that would be necessary to keep in mind, regardless of order.

Perceived reputation benefits of punishment. To measure the perceived reputation benefits of punishment, we instructed subjects to imagine that, instead of being asked to make a judgement about the Helper, they had instead been given the opportunity to punish the Helper with a financial fine. Specifically, we instructed subjects to imagine that (i) the Helper did not share with the Recipient, and (ii) they were given 30¢, and had the opportunity to punish the Helper by paying 5¢ to deduct 15¢ from the Helper's payoff. Then, subjects answered six questions, which measured their beliefs that punishing—if observed—would have positive reputation consequences, as compared to not punishing.

Models of punishment as a signal of trustworthiness show that, depending on the context, the act of punishing can be a *positive* signal (i.e., it can increase the punisher's perceived

trustworthiness), and the act of not punishing can be a *negative* signal (i.e., it can decrease the punisher's perceived trustworthiness) (Jordan et al., 2016a; Jordan & Rand, 2017). For this reason, we asked three questions about the likely positive reputation consequences of punishing, and three questions about the likely negative reputation consequences of not punishing.

Specifically, we first asked subjects, if they were to punish the Helper, (i) how morally good this would make them look in the eyes of others, (ii) how much this would benefit their reputation, and (ii) how positively others would see this. Next, we asked subjects, if they were not to punish the Helper, (i) how immoral this would seem to somebody else, (ii) to what extent this would make them look bad, and (ii) how much this would reflect negatively on their reputation. These questions were presented in that fixed order, and were each answered on a Likert scale that ranged from 0 to 100 in 10-point increments, with extreme anchors reading *Not at all* and *Very much.* As our composite measure of the perceived reputation benefits of punishment, we took the average value across the six items in our scale (although we note that results were qualitatively equivalent when using only the positive or negative items).

We also note that our items were not neutrally framed (i.e., they suggested that punishing would be perceived neutrally or positively but not negatively, and that not punishing would be perceived neutrally or negatively but not positively). While this is likely to have affected absolute ratings of the perceived reputation value of punishment, we do not believe that it is likely to have interacted with our helping opportunities manipulation in order to produce the predicted mediation pattern. Finally, we note that in Experiment 4 (but not Experiments 2 or 5), we added an extra item to the end of our perceived reputation benefits of punishment scale, which was designed to measure subjects' valuation of those benefits (see stimuli in SM for details). We found no condition effect on this item, and thus do not report analyses of it.

General reputation concerns. To measure general reputation concerns, we used a sixteen-item scale. Eight of the items were the eight straightforwardly-worded items on the brief fear of negative evaluation scale (BFNE). The BFNE (Leary, 1983) is based on the fear of negative evaluation scale (FNE) (Watson & Friend, 1969), which was designed to measure the extent to which people are afraid of being evaluated negatively by others, and predicts behaviors like working hard to gain approval in the eyes of others, as well as traits like social approval seeking. The eight straightforwardly-worded BFNE items have been shown to correlate more strongly with theoretically related measures than reverse-worded items (Rodebaugh et al., 2004). The other eight items in our general reputation concerns scale were designed by us to mirror these eight BFNE items, but measure the desire for *positive* evaluation.

All 16 items were measured as in the BFNE: with 1-5 Likert scales with anchors at every item, ranging from *Not at all characteristic of me* to *Extremely characteristic of me*. We presented the 16 items in a pseudorandom order across two pages. For all subjects, each page had the same four fear of negative evaluation items and four desire for positive evaluation items, but we randomized the order of the items within each page across subjects. As our composite general reputation concerns measure, we took the average value across our 16 scale items (although the results were qualitatively equivalent when using only the positive or negative items).

Finally, we note that in Experiments 4-5, which measured both mediators, we randomized the order in which they were measured between subjects.

Results

Perceived reputation benefits of punishment (PRBP). We began by investigating our first candidate mediator in Experiments 2, 4, and 5. Collapsing across these three experiments, we found that our six-item PRBP scale was reliable ($\alpha = 0.92$). Before testing for mediation, we

also estimated the total effect of helping opportunities on outrage in Experiments 2, 4, and 5 (which was slightly different from the results reported in Analysis 1, because it excluded Experiments 1 and 3). Within these experiments, we observed significantly more outrage when helping was not possible, $B = 0.09$, $t = 4.57$, $p < .001$, $n = 2434$.

Next, we tested for mediation, and found the predicted pattern (Figure 2a). First, helping opportunities attenuated the perceived reputation benefits of punishment. Subjects in Condemnation Only reported that punishment would have significantly greater reputational benefits ($M = 3.77$, $SD = 2.43$) than subjects in Condemnation+Helping did ($M = 3.24$, $SD = 2.47$), $B = 0.11$, $t = 5.42$, $p < .001$, $n = 2434$. This suggests that subjects did, in fact, treat helping opportunities as a cue of punishment's reputation value. Second, predicting outrage as a function of condition and PRBP, we found a significant effect of PRBP, $B = 0.51$, $t = 29.47$, $p < .001$, $n = 2434$. This suggests that individuals who believed that punishing would confer larger reputation benefits experienced more outrage, which is consistent with the theory that outrage functions to motivate punishment and thus is sensitive to its perceived reputation value.

Finally, we investigated the indirect effect of helping opportunities on outrage through PRBP, and the direct effect of helping opportunities on outrage. For all analyses, we calculated indirect and direct effects using standardized Beta coefficients and Preacher and Hayes's (2008) bootstrapping procedure with 5,000 resamples. We found a significant indirect effect of .06 [.04, .08], and a significant direct effect of 0.04 (.002, .07). Comparing the direct effect to the total effect of helping opportunities revealed that 61% of the total effect was mediated by PRBP.

General reputation concerns. We next investigated our second candidate mediator in Experiments 3-5. Collapsing across these three experiments, we found that our sixteen-item GRC scale was reliable ($\alpha = 0.96$). Before testing for mediation, we also estimated the total effect of

helping opportunities on outrage in Experiments 3-5. Within these experiments, we observed significantly more outrage when helping was not possible, $B = 0.08$, $t = 4.01$, $p < .001$, $n = 2432$.

Next, we tested for mediation, and found equivocal evidence (Figure 2b). First, helping opportunities had a marginally significant effect on general reputation concerns. Subjects in Condemnation Only reported being marginally significantly more concerned with their reputations ($M = 2.99$, $SD = 0.96$) than subjects in Condemnation+Helping did ($M = 2.92$, $SD = 0.97$), $B = 0.04$, $t = 1.79$, $p = .073$, $n = 2432$. This suggests that having the chance to help may have reduced the extent to which subjects felt concerned with their reputations. Second, predicting outrage as a function of condition and GRC, we found a significant effect of GRC, $B = 0.20$, $t = 10.19$, $p < .001$, $n = 2432$. This suggests that individuals with greater general reputation concerns reported more outrage. This correlation is consistent with the theory that reputation concerns drive punishment, and thus shape outrage as a motivator of punishment.

Finally, we estimated the indirect effect of helping opportunities through GRC, as well as the direct effect of helping opportunities. We observed a marginally significant indirect effect of .01 [-.001, .02], and a significant direct effect of 0.07 [.04, .11]. Comparing the direct and total effects revealed that 9% of the total effect was mediated by GRC.

Multiple mediation. Finally, we simultaneously investigated both mediators in Experiments 4-5. First, within these two experiments, we investigated the total effect of helping opportunities on outrage. We observed significantly more outrage when helping was not possible, $B = 0.10$, $t = 4.14$, $p < .001$, $n = 1615$.

Next, we conducted a multiple mediation analysis (Figure 2c). First, we found that helping opportunities significantly influenced both mediators. In Condemnation Only, subjects both reported that the reputation value of punishment was higher ($B = 0.09$, $t = 3.71$, $p < .001$, n

= 1615) and that they were more concerned with their reputations (B = 0.07, t = 2.68, p = .007, n = 1615). Second, predicting outrage as a function of condition, PRBP, and GRC, we found significant effects of both PRBP (B = .49, t = 22.82, p < .001, n = 1615) and GRC (B = .12, t = 5.71, p < .001, n = 1615). This result suggests that the perceived reputation benefits of punishment and general reputation concerns may have had independent effects on outrage.

Finally, we estimated the indirect effects of each mediator, as well as the direct effect of helping opportunities. We found significant indirect effects through both PRBP (.05 [.02, .07]) and GRC (.01 [.001, .01]), resulting in a significant total indirect effect (.05 [.03, .08]). We also found a significant direct effect of .05 [.01, .09]. Comparing the direct and total effects revealed that 52% of the total effect was mediated by our mediators. (Note that this percentage is smaller than what was reported above for PRBP alone because in Experiment 2, which only measured PRBP, PRBP mediated considerably more of the total effect than it did in Experiments 4-5.)

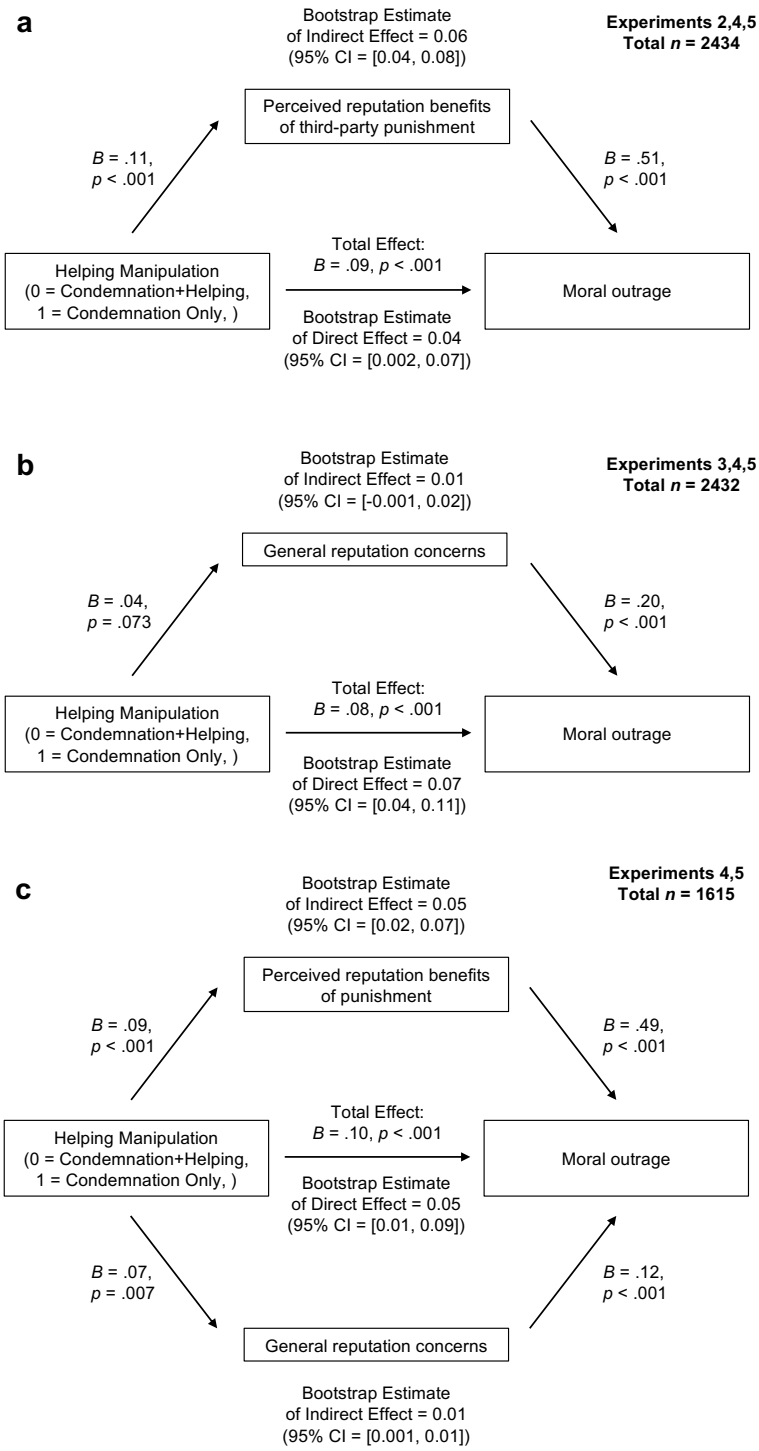

Figure 2. Reputation constructs mediate the effect of helping opportunities on outrage. We illustrate the effect of our helping opportunities manipulation on moral outrage, as mediated by (a) the perceived reputation benefits of punishment (in a single mediation analysis of

Experiments 2, 4, and 5), (b) general reputation concerns (in a single mediation analysis of Experiments 3-5), and (c) both candidate mediators (in a multiple mediation analysis of Experiments 4-5).

Mediation results by Experiment.

Finally, we conducted mediation analyses separately by experiment (Table 3). All three experiments measuring PRBP showed a consistent pattern: when subjects did not have the opportunity to help, they reliably reported increased PRBP, which reliably predicted increased outrage—so we reliably observed indirect effects through PRBP.

In contrast, across the three experiments measuring GRC, we saw a mixed pattern. In all three experiments, GRC reliably predicted increased outrage. However, we found equivocal evidence regarding the effect of the Condemnation Only condition on GRC, and thus the indirect effect through GRC. Specifically, the marginally significant positive indirect effect in our aggregated analysis was driven most strongly by the significant positive effect in Experiment 5. It was also consistent with the directionally positive effect in Experiment 4—but not the directionally negative effect in Experiment 3. In interpreting this pattern, it is perhaps worth noting that Experiment 3 was the one experiment in which we did not observe a significant effect of helping opportunities on outrage (see Table 2); this might suggest that for some reason (e.g., a randomization failure), the effect of our manipulation on both outrage and GRC was meaningfully different in Experiment 3.

Finally, in both experiments measuring both PRBP and GRC, our multiple mediation analyses produced fairly consistent results. Both experiments showed a significant positive indirect effect of PRBP, and a directionally positive indirect effect of GRC (although this effect was only significant in Experiment 5).

	Exp. 2 *n* = 819	**Exp. 3** *n* = 817	**Exp. 4** *n* = 811	**Exp. 5** *n* = 804	**Aggregate** (*n* varies)
Single mediation by Perceived Reputation Benefits of Punishment (PRBP) (aggregate *n* = 2434)					
Effect of Condemnation Only (CO) dummy on PRBP	*B* = .14 *p* < .001		*B* = .09 *p* = .007	*B* = .09 *p* = .011	*B* = .11 *p* < .001
Effect of PRBOP on outrage (controlling for CO)	*B* = .53 *p* < .001		*B* = .50 *p* < .001	*B* = .51 *p* < .001	*B* = .51 *p* < .001
Indirect effect of CO on outrage via PRBP	.07 [.04, .11]		.05 [.01, .08]	.05 [.01, .08]	.06 [.04, .08]
Single mediation by General Reputation Concerns (GRC) (aggregate *n* = 2432)					
Effect of CO on GRC		*B* = -.02 *p* = .505	*B* = .05 *p* = .157	*B* = .08 *p* = .018	*B* = .04 *p* = .073
Effect of GRC on outrage (controlling for CO)		*B* = .21 *p* < .001	*B* = .18 *p* < .001	*B* = .21 *p* < .001	*B* = .20 *p* < .001
Indirect effect of CO on outrage via GRC		-.005 [-0.02, 0.01]	.01 [-0.004, 0.02]	.02 [0.001, 0.03]	.01 [-0.001, 0.02]
Multiple mediation by both PRBP and GRC (aggregate *n* = 1615)					
Indirect effect of CO on outrage via PRBP in multiple mediation			.05 [.01, .08]	.04 [.01, .08]	.05 [.02, .07]
Indirect effect of CO on outrage via GRC in multiple mediation			.01 [-.003, .01]	.01 [.001, .02]	.01 [.001, .01]

Table 3. Analysis 2 by Experiment. For a and b paths, we show standardized coefficients and p-values, and for indirect effects, we show standardized coefficients and 95% CIs.

Discussion

Together, Analysis 2 supports the hypothesis that helping opportunities shape reported outrage insofar as they serve as a reputation-relevant cue. We found robust evidence for partial mediation through the perceived reputation benefits of punishment. When subjects did not have the chance to help, they saw punishing as having greater reputation value—and insofar as this was true, their outrage was heightened. This pattern is consistent with our theory that helping opportunities influence outrage because helping is a stronger signal of trustworthiness than punishment, and outrage is sensitive to the potential signaling value of punishment.

We also found equivocal evidence for partial mediation through general reputation concerns. When subjects did not have the chance to help, our results suggest that they may have

felt somewhat more concerned with their reputations. This effect was only marginally significant despite our large sample sizes, and the effect size was very small. The relative weakness of this effect could reflect that we designed our reputation concerns scale as a trait measure, and that it measured *general* reputation concerns, rather than *moral* reputation concerns specifically. But it might also reflect that helping opportunities genuinely do not have much effect on reputation concerns. However, while the effect of helping opportunities on general reputation concerns was marginal, we did observe a robust correlation between general reputation concerns and outrage. This correlation supports our reputation framework for moral outrage.

Reputation in the eyes of who? Our reputation heuristics theory proposes that even when reputation is not at stake, people may engage in reputation-relevant computations that make them sensitive to reputation cues. But what kind of reputation-relevant computations? Our candidate mediators focused on reputation in the eyes of vaguely-described "others". We asked subjects to report on how *others* would evaluate them if they chose to punish (or not), and how concerned they were with *others* evaluating them positively (or negatively). We reasoned that a plausible mechanism through which our subjects may have implemented a reputation heuristic, and shown a sensitivity to helping opportunities, is by engaging in (likely implicit) computations about their hypothetical reputation in the eyes of other (but absent) individuals.

However, other mechanisms are also plausible: subjects may have engaged in reputation-based computations that did *not* concern reputation in the eyes of absent others. For example, subjects may have conducted computations about their reputation in their *own* eyes. People can always observe their own behavior, and are strongly driven to view themselves as morally good—and a large body of work demonstrates the importance of self-concept management in shaping our moral psychology and behavior (Aquino & Reed, 2002; Mazar, Amir, & Ariely,

2008; Merritt, Effron, & Monin, 2010; Monin & Miller, 2001; Perugini & Leone, 2009; Sachdeva, Iliev, & Medin, 2009; Young, Chakroff, & Tom, 2012). Likewise, despite the fact that the experimenter could not link subjects' responses to their identities, subjects may have conducted computations about their reputations in the eyes of the *experimenter*. Such computations could reflect a general heuristic to care about what people will think of your behavior, even if they cannot identify you or will not interact with you in the future.

Moreover, when measuring our candidate mediators, we may have tapped these alternative reputation computations. Specifically, when measuring our first candidate mediator, we asked subjects how "others" would perceive the choice to (or not to) punish selfishness; however, their responses may have reflected how *they* would have perceived their *own* choice, or how the experimenter would have perceived it. Likewise, when measuring our second candidate mediator, we asked subjects how concerned they typically are with the way others evaluate them; however, their responses may have reflected concerns with their own self-evaluations, or evaluations from the experimenter.

These different reputation-based computations are theoretically distinct, but teasing them apart empirically is a challenge: they are not mutually exclusive, and may often be strongly positively correlated. Moreover, all of these reputation-based computations could ultimately function to implement reputation heuristics in anonymous interactions. For these reasons, we did not attempt to discriminate between them in our mediation analyses.

However, in Experiment 6, we did collect some exploratory data designed to investigate the extent to which subjects reported being concerned with signaling to others, themselves, and the experimenter. These items were retrospectively measured after outrage, punishment, and the post-experimental question about whether other players could influence subjects' bonuses. We

observed strong positive correlations between them (*B*s ranging from .66 to .83, all *p*s < .001), and none of them were influenced by our manipulation of helping opportunities; thus, we did not treat them as mediators. However, descriptive statistics about these variables may provide some interesting and suggestive information about the mechanisms through which reputation heuristics operate in the context of our experiments.

Specifically, in our post-experimental survey (i.e., after we measured both outrage punishment), we asked subjects to report the extent to which they had been concerned with whether their decisions would (i) make them look like a good person in the eyes of others (other-signaling concerns), (ii) make them look like a good person in the eyes of the "HIT requestor" (i.e., the experimenters) (experimenter-signaling concerns), and (iii) make them think that they were a good person (self-signaling concerns), using 1-7 Likert scales ranging from *Not concerned at all* to *Very concerned*. We randomized the order of these three questions between subjects, and in the Condemnation+Helping condition, we specifically asked subjects about the extent to which they had held such concerns while in the role of the condemner. We observed moderate levels of all three types of signaling concerns, with somewhat higher levels for self-signaling concerns ($M = 3.38$, $SD = 2.07$) than other-signaling concerns ($M = 2.87$, $SD = 1.96$) (paired-sample t-test: $t = 18.51$, $p < .001$, $n = 2924$), and somewhat higher other-signaling concerns than experimenter-signaling concerns ($M = 2.74$, $SD = 1.93$) (paired-sample t-test: $t = 5.89$, $p < .001$, $n = 2924$). These results suggest that all of these reputation concerns are plausible mechanisms through which subjects may have implemented reputation heuristics in the context of our one-shot anonymous experiments, and should be investigated in future research.

Analysis 3

Together, Analyses 1 and 2 supported our proposal that moral outrage is sensitive to the potential signaling value of punishment—and thus is influenced by helping opportunities even in one-shot anonymous interactions. In Analysis 3, we tested the prediction that *costly punishment* decisions are also sensitive to helping opportunities in one-shot anonymous interactions.

Methods

Design. To this end, we conducted three additional experiments (Experiments 8-10) that measured costly punishment in one-shot anonymous interactions. Their design was very similar to that of Experiments 1-7, except that the dependent variable was costly punishment, not moral outrage. It was thus also very similar to previous research showing that helping opportunities reduce punishment in a situation where reputation was at stake (Jordan et al., 2016a), except that we modified the design so that reputation was not at stake.

Experiments 8-10 thus employed the Third-Party *Punishment* Game (TPPG) from Jordan et al., 2016a. The TPPG was identical to the TPCG described previously, except that the Condemner was replaced with a Punisher. The Punisher thus made an incentivized decision that was similar to the hypothetical punishment decision described to subjects in our perceived reputation benefits of punishment scale. As in Experiments 1-7, subjects in Experiments 8-10 read about the TPPG and their role(s) in it. In the Punishment Only condition, target subjects played once as the Punisher. In Punishment+Helping, they played twice: once as the Punisher, and once as the Helper.

Experiments 8-10 measured punishment via the "strategy method": Punishers were endowed with 20¢, and—without knowing whether or not the Helper chose to share with the Recipient—decided whether or not to commit to punishing the Helper *in the event that they*

chose not share. Specifically, Punishers could commit to paying 5¢ to punish the Helper by deducting 15¢ from their payoff, in the event that the Helper did not share. By using the strategy method, we were able to obtain an incentivized measure of punishment of selfishness for all Punishers, despite the fact that not all Helpers selfishly declined to share. We note that the strategy method is a standard approach for measuring third-party punishment (Fehr & Fischbacher, 2004) and evidence suggests that it does not influence rates of punishment (Jordan et al., 2015).

In addition to Experiments 8-10, as previously described, Experiment 6 also measured costly punishment. Specifically, in Experiment 6, after manipulating helping opportunities and measuring outrage, we explained the punishment decision described above, and then measured punishment. Thus, Experiments 6 and 8-10 all manipulated helping opportunities and measured punishment, and we analyzed them together in Analysis 3. In the context of Analysis 3 (as well as all other punishment analyses in this paper) we refer to the Experiment 6 conditions as "Punishment Only" and "Punishment+Helping" (rather than "Condemnation Only" and "Condemnation+Helping", as in the context of our outrage analyses).

However, recall that in Experiment 6, all subjects of interest were matched with a Helper who did not share, and were told that the Helper did not share before rating their outrage. Thus, Experiment 6 subjects also knew that the Helper did not share before deciding whether to punish; in other words, Experiment 6 did *not* use the strategy method, and helps test whether our results are robust to this methodological distinction. Additionally, in Experiment 6, we presented subjects with a filler task after measuring outrage but before measuring punishment (see Procedure for more details). Our goal was to reduce the probability that measuring outrage influenced punishment ratings via anchoring or consistency effects (which could cause subjects

to match their punishment decisions to their outrage ratings) in order to facilitate comparison between Experiment 6 and our other punishment experiments (which did not measure outrage).

Moreover, several other details varied across our set of punishment experiments (see Table 8 for an overview of differences). First, as described previously, subjects in Experiment 6 who had the opportunity to help always made their helping decision before we measured outrage—and subsequently, punishment. In contrast, within the Punishment+Helping conditions of Experiments 8-10, we always counterbalanced the order of helping and punishment decisions. (For analyses of order effects within the Punishment+Helping conditions of our punishment experiments that employed counterbalancing, see SM.) Second, Experiments 8-9 included a Helping Only condition (in which target subjects played the TPPG once as the Helper).

Third, while reputation was never at stake in Experiments 6 and 8, Experiments 9-10 also included a manipulation of whether reputation was at stake (which we examine in Analysis 5b). Specifically, for half of subjects, like in Experiments 6 and 8, the experiment ended after the TPPG; thus, TPPG decisions had no reputation consequences. These are the subjects who we analyze in Analysis 3, which investigates one-shot anonymous punishment. However, for the other half of subjects, the TPPG was followed by an economic Trust Game (TG), as in Jordan et al., 2016a. In this TG, another AMT worker—who was *not* involved in the TPPG—decided how much money to entrust the target subject with, and could condition this decision on the target subject's TPPG behavior. Thus, reputation *was* at stake. In Analysis 5b, we provide more methodological details about our TG manipulation, and investigate costly punishment when reputation *is* at stake. (Hereafter, we thus refer to Experiments 6 and 8, and the "No Trust Game" conditions of Experiments 9-10, as our "No TG punishment experiments"; and we refer to the

"Trust Game" conditions of Experiments 9-10, as well as two other very similar experiments employing a Trust Game, as our "TG punishment experiments").

Finally, for completeness we note that in Experiments 8-9, after subjects finished their economic game decisions (i.e., punishment and/or helping), they completed some emotion ratings. Specifically, subjects in Punishment Only completed our three-item outrage scale, subjects in Helping Only completed a three-item scale measuring positive emotions towards the Recipient, and subjects in Punishment+Helping completed both scales. This design makes it possible to analyze the effect of helping opportunities on outrage in these experiments; however, we leave this analysis to the SM because, due to a programming error, we failed to counterbalance the order of scale presentation (outrage or positive emotions first) in the Punishment+Helping condition. As such, we confounded the effect of helping before rating outrage with the effect of rating positive emotions before rating outrage, and suspect that these two manipulations may have had countervailing effects. See SM for complete methodological details, analyses, and discussion.

Subjects. As reported above, in Experiment 6 we requested a target of $n = 1500$ subjects per condition from AMT (i.e., a total of $n = 3000$ subjects). In Experiments 8-9, which both included Helping Only conditions, we requested a target of $n = 400$ subjects per condition (i.e., a total of $n = 1200$ subjects in Experiment 8, and across the No TG conditions of Experiment 9). In Experiment 10, which did not include a Helping Only condition, we decided (prior to data collection) to request a larger sample of $n = 775$ subjects per condition (i.e., a total of $n = 1550$ subjects across the No TG conditions) for increased power because this experiment sought to detect an interaction between helping opportunities and the presence of a TG. Our final sample of No TG punishment experiments includes $n = 6863$ subjects ($n = 3066$ in Punishment Only, n

= 3010 in Punishment+Helping, and n = 787 in Helping Only), M_{age} = 35.64 years, SD_{age} = 11.61 years, 45% male.

Procedure. The Experiment 8-10 procedure was analogous to that of our outrage experiments, but with the above-described design changes. Our TPPG instructions were followed by four TPPG comprehension questions. Two tested comprehension of the incentive structure underlying the Helper's decision (like in the TPCG); the other two focused on the Punisher's decision. When measuring punishment, we reminded subjects that they had 20¢, and that their job was to decide whether to pay 5¢ to deduct 15¢ from the Helper, if they Helper chose not to share with the Recipient. We then asked them to make a decision, which we subsequently repeated back to them.

In Experiment 6, after measuring outrage, we presented subjects with a filler task that involved memorizing a list of words. Specifically, we informed subjects that they would be shown a list of 20 words for 60 seconds, and instructed them to try their best to study and remember them without writing them down, before attempting to recall as many as possible on the next screen. Then, we presented 20 neutral words (with no moral content), while a timer counted down from one minute. Finally, the screen advanced and subjects were asked to recall as many words as they could. Subjects were informed that their performance in this task would have no bearing on their bonus payments.

Afterwards, we informed subjects that they would move on to the "next phase" of the game, where they would have the opportunity to make another decision. Then, we explained their punishment decision, and presented them with the two punishment-relevant TPPG comprehension questions. Finally, we measured punishment. Punishment was measured as in Experiments 8-10, except that in Experiment 6, subjects had already been told that the Helper did

not share and were asked whether they wanted to punish (whereas in Experiments 8-10, subjects were asked whether they wanted to punish *if* the Helper did not share).

Results

We began by investigating the effect of helping opportunities on punishment in an aggregated analysis of our No TG punishment experiments. As predicted and illustrated in Figure 3a, subjects were more likely to punish in Punishment Only (32%) than Punishment+Helping (27%), $OR = 1.31$, $z = 4.78$, $p < .001$, $n = 6076$. Thus, when subjects had the opportunity to signal their trustworthiness via direct helping, they were less likely to pay to punish—even though reputation was not actually at stake.

Next, as in Analysis 1, we asked whether this effect simply reflected that subjects in Punishment+Helping had two actions available to them. To address this question, we investigated whether punishment opportunities reciprocally influenced helping in the subset of our No TG punishment experiments that included a Helping Only condition. On the contrary, as predicted and illustrated in Figure 3b, subjects in these experiments helped at comparable rates in Helping Only (58%) and Punishment+Helping (57%), $OR = 1.05$, $z = .43$, $p = .669$, $n = 1556$. We also investigated only these experiments, and used linear regressions to predict both punishment and helping as a function of condition. We found that the standardized condition coefficient was marginally significantly larger when predicting punishment ($B = .08$, $SE = .03$, $p = .002$) than helping ($B = .01$, $SE = .03$, $p = .669$), $z = 1.90$, $p = .058$.

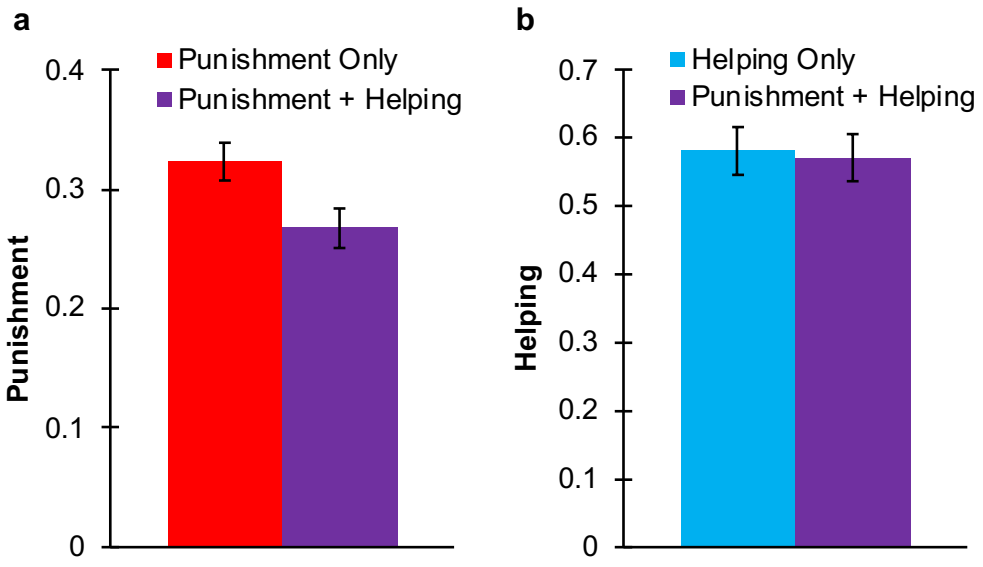

Figure 3. Helping opportunities reduce punishment (while punishment opportunities do not reduce helping) in one-shot anonymous interactions. In **a,** we plot the proportion of subjects punishing as a function of helping opportunities across our No TG punishment experiments. In **b,** we plot the proportion of subjects helping as a function of punishment opportunities across the subset of these experiments with a Helping Only condition. Error bars are 95% CIs.

Next, we investigated the effect of helping opportunities on punishment by experiment (Table 4). In three out of our four experiments, we observed significantly more punishment in Punishment Only than Punishment+Helping, and in the fourth, we observed a marginally significant effect in the same direction. Thus, our results were fairly robust across experiments. In particular, we note that compared to Experiments 8-10, Experiment 6 showed a similar effect, despite its methodological differences. This suggests that that our key result was robust to whether outrage was measured before punishment, and whether the strategy method was used to measure punishment.

Exp. 6 $n = 2924$	**Exp. 8** $n = 772$	**No TG condition of Exp. 9** $n = 799$	**No TG condition of Exp. 10** $n = 1581$	**Aggregate** $n = 6076$
$OR = 1.26$ $p = .005$	$OR = 1.53$ $p = .009$	$OR = 1.32$ $p = .072$	$OR = 1.31$ $p = .014$	$OR = 1.31$ $p < .001$

Table 4. The effect of helping opportunities on punishment in one-shot anonymous interactions, by experiment. Reported sample sizes include only subjects for whom punishment was measured (i.e., sample sizes exclude subjects in Helping Only conditions).

Finally, like in Analysis 1, we investigated whether helping opportunities merely influenced one-shot anonymous punishment among subjects who held the mistaken explicit belief that other players *could* observe their behavior and influence their payoffs. The only punishment experiment in which we measured these beliefs was Experiment 6; thus, we asked whether helping opportunities reduced costly punishment in Experiment 6, excluding all subjects who reported that other player(s) could influence their bonuses. Indeed, we continued to observe more punishment in Punishment Only than Punishment+Helping, $OR = 1.34$, $z = 2.47$, $p = .014$, $n = 1703$. Thus, our punishment results do not seem to be driven by subjects who held the mistaken explicit belief that other players could observe their behavior and influence their payoffs.

Discussion

Analysis 3 provides evidence that in one-shot anonymous interactions, helping opportunities can influence costly punishment. It demonstrates that in settings where reputation is not at stake, reputation cues do not merely have the potential to influence reported outrage (as shown in Analysis 1)—they can also influence the willingness to pay actual costs to punish wrongdoers. As in the context of outrage, this effect is relatively small, but has the important implication that a reputation framework can help explain one-shot anonymous punishment.

Analysis 4

Together, Analyses 1 and 3 provided evidence that helping opportunities reduce outrage and punishment in one-shot anonymous interactions, and Analysis 2 provided evidence that helping opportunities specifically reduce outrage insofar as they serve as a reputation-relevant cue. In Analysis 4, we aimed to further support our reputation-based theory by testing a deflationary explanation for the effects of helping opportunities on outrage and punishment.

As discussed previously, our theory posits that helping opportunities should reduce outrage and punishment for all subjects, regardless of whether or not they chose to help. After having the chance to help, the *positive* reputations of people who *did* help should be relatively established, and the *negative* reputations of people who *did not* help should be relatively established—so for everyone, the potential reputation value of punishing should decline.

However, one might imagine that helping opportunities specifically reduced outrage and punishment among subjects who declined to help, for two reasons. First, declining to help and then condemning another non-helper is hypocritical, and hypocrites are viewed negatively (Barden, Rucker, & Petty, 2005; Effron, Lucas, & O'Connor, 2015; Jordan, Sommers, Bloom, & Rand, 2017); thus, hypocrisy aversion could reduce outrage among non-helpers. Second, declining to help could increase empathy for other non-helpers, reducing outrage.

Our Analysis 2 results provide some evidence that this empathy mechanism is not the sole driver of our results: an empathy-based mechanism would not predict the observed mediation patterns. However, our finding that helping opportunities reduced the perceived reputation value of punishment *could* merely reflect subjects who declined to help perceiving that punishment would be seen as hypocritical, harming their reputations. While this possibility would still support the theory that reputation concerns shape outrage and punishment in one-shot

anonymous contexts, our reputation-based theory is based on a broader reputation mechanism that should extend to both helpers and non-helpers. In Analysis 4, we sought to support our reputation-based mechanism and provide evidence against the deflationary explanation that our results merely reflect empathy or hypocrisy aversion among non-helpers. To this end, we tested our prediction that the effects of helping opportunities on outrage and punishment were *not* solely driven by non-helpers.

To test this prediction, we needed to compare the effects of helping opportunities on outrage and punishment among helpers versus non-helpers. But how? One obvious approach is to simply compare subjects who chose to help to subjects who did not have the opportunity to help (and to compare subjects who chose not to help to subjects who did not have the opportunity to help). However, these comparisons introduce a self-selection effect that violates random assignment and prevents appropriate causal inference: subjects who chose to help might differ from the overall population in their baseline inclination towards outrage and punishment (and likewise for subjects who chose not to help). Consistent with this possibility, across the Condemnation+Helping conditions of our outrage experiments, subjects who helped reported significantly more outrage ($M = 37.78$, $SD = 30.04$, $n = 2937$) than subjects who did not help ($M = 12.82$, $SD = 21.72$, $n = 1275$), $B = 0.39$, $t = 27.38$, $p < .001$. Likewise, across the Punishment+Helping conditions of our punishment experiments (including both our TG and no TG punishment experiments), subjects who helped ($n = 3257$) punished at a significantly higher rate (36%) than subjects who did not help ($n = 1445$) (10%), $OR = 5.06$, $z = 16.89$, $p < .001$. Thus, comparing helpers to those who did not have the opportunity to help likely biases us away from finding the predicted negative effect of helping opportunities on outrage and punishment

(while comparing non-helpers to those who did not have the opportunity to help likely biases us towards finding the predicted negative effects).

To avoid this self-selection issue, we would ideally compare (i) subjects who did help (when given the chance) to subjects who would have helped (if given the chance), and (ii) subjects who did not help (when given the chance) to subjects who would not have helped (if given the chance). However, we do not know which subjects in our Condemnation Only and Punishment Only conditions would have helped, had they instead been assigned to our Condemnation+Helping or Punishment+Helping conditions.

In Experiment 6, we addressed this issue by gathering additional data, in order to obtain a measure of helping for *all* subjects (regardless of condition). One way to do this would have been to give subjects who initially did not have the opportunity to help (i.e., subjects in our Condemnation / Punishment Only condition) an unexpected helping opportunity after we measured their affective outrage and punishment. However, we were concerned about the possibility of order effects in a design like this (i.e., about the possibility that different types of people would choose to help, depending on whether helping was measured at the beginning or end of the experiment). Instead, then, we conducted a follow-up experiment approximately two weeks after Experiment 6 was finished. In this experiment, we measured helping among all subjects, regardless of their Experiment 6 condition.

In Analysis 4, we treated follow-up experiment helping as an index of an individual's propensity to help when given the chance. In the words, we treated it as a proxy for who *would* have helped in Experiment 6, even among subjects who were *not* given the opportunity to help. Through this approach, we attempted to gain insight into whether the effect of helping opportunities on affective outrage and punishment in Experiment 6 was specifically driven by

non-helpers, as predicted by the hypocrisy aversion and empathy mechanisms. We did so by investigating (i) whether helping in the follow-up experiment moderated the effects of helping opportunities on affective outrage or punishment, and (ii) if so, whether these effects were driven solely by follow-up experiment non-helpers, or also held among follow-up experiment helpers. We predicted that either (i) follow-up experiment helping would not moderate the effects of helping opportunities on affective outrage or punishment, or (ii) it would moderate, but the negative effects of helping opportunities would hold among helpers.

Methods

To conduct our follow-up experiment, 13 (12) days after beginning (completing) data collection for Experiment 6, we invited all subjects to participate in an additional experiment, in which everyone was asked to complete the same helping decision that we used in the helping condition of Experiment 6 (and all other experiments). We kept our follow-up experiment survey open to new respondents for 8 days, at which point the rate of new responses had become very low, and we closed the survey. We then linked follow-up survey responses to our Experiment 6 data using AMT Worker IDs; for subjects who completed the follow-up survey more than once, we used their chronologically first response. A total of n = 2056 subjects completed our follow-up experiment (n = 1051 who were assigned to the Condemnation Only condition of Experiment 6, n = 1005 who were assigned to Condemnation+Helping), M_{age} = 37.22 years, SD_{age} = 12.16 years, 44% male.

We designed our follow-up experiment with the goal that few (if any) subjects would remember Experiment 6 clearly enough for their Experiment 6 decisions to influence their follow-up experiment helping. We attempted to facilitate this goal in two ways. First, we did not tell Experiment 6 subjects that there would be a follow-up experiment, and when inviting them to

participate in the follow-up experiment, we did not tell them that it was related to Experiment 6. We hoped that this would limit the extent to which the follow-up experiment reminded them of Experiment 6. Second, we conducted the follow-up experiment after a meaningful time delay, which we hoped would substantially weaken subjects' memories of Experiment 6 (and give them ample opportunity to complete other tasks on AMT, interfering with their memories). Consistent with this goal, at the end of the follow-up experiment we asked subjects to report the approximate number of tasks they had completed on AMT over the last two weeks. Among subjects who answered this question with a number (n = 1994), the median answer was 80 (25th percentile = 30, 75th percentile = 200). We see these numbers as relatively large, and thus find it likely that most subjects did not have a clear memory of Experiment 6.

Results

Validation of our analysis approach. We began by investigating the validity of treating helping in our follow-up experiment as a proxy for helping in Experiment 6. We did so in two ways. First, we asked whether our Experiment 6 manipulation of helping opportunities influenced rates of helping in our follow-up experiment. This question is relevant to whether it may be appropriate to treat helping in our follow-up experiment as a moderator of our Experiment 6 results, even though the follow-up experiment was conducted after Experiment 6. Indeed, we found that subjects who were assigned to the Condemnation Only condition of Experiment 6 did not show significantly different rates of helping in the follow-up experiment (62%), as compared to subjects who were assigned to the Condemnation+Helping condition of Experiment 6 (60%), OR = 1.09, z = 0.94, p = .346, n = 2056. This provides suggestive evidence that the type of individual who helped in the follow-up experiment did not vary by condition, such that follow-up experiment helping may be an appropriate moderator.

Second, we investigated the correlation between helping in the follow-up experiment and helping in Experiment 6, among subjects assigned to the Experiment 6 helping condition. This question is relevant to whether follow-up experiment helping is in fact a reliable predictor of Experiment 6 helping. Indeed, we found that 88% of follow-up experiment helpers (n = 606) helped in Experiment 6, while only 34% of follow-up experiment non-helpers (n = 399) helped in Experiment 6. We thus observed a significant association between helping in Experiment 6 and the follow-up experiment (via linear regression $B = 0.56$, $t = 21.57$, $p < .001$; via logistic regression $OR = 14.64$, $z = 16.25$, $p < .001$, $n = 1005$). In other words, helping in the follow-up experiment strongly predicted helping in Experiment 6, when given the chance.

Do helping opportunities solely reduce affective outrage and punishment among non-helpers? After validating our Analysis 4 approach, we moved to testing our key prediction: that helping opportunities did *not* solely reduce affective outrage and punishment among non-helpers. More specifically, we tested our prediction that either (i) follow-up experiment helping would not moderate the effects of helping opportunities on affective outrage or punishment, or (ii) follow-up experiment helping would moderate, but the negative effects of helping opportunities on affective outrage and punishment would hold among helpers.

We began by investigating affective outrage. We predicted Experiment 6 affective outrage as a function of a Condemnation Only dummy, helping in the follow-up experiment, and their interaction. We found a significant negative interaction, $B = -.10$, $t = -2.45$, $p = .014$, $n = 2056$. In other words, we *did* support the deflationary explanation's prediction that follow-up experiment helping should moderate the effect of helping opportunities on affective outrage.

As such, we moved to investigating whether the negative effect of helping opportunities on affective outrage held among helpers. Critically, we found a significant positive effect of a

Condemnation Only dummy on affective outrage among both follow-up experiment helpers, B = .07, t = 2.42, p = .016, n = 1261, and non-helpers, B = .18, t = 5.19, p < .001, n = 795. Thus, the negative effect of helping opportunities on affective outrage *did* hold among helpers, as predicted by our signaling account and *not* the deflationary explanation. That said, we did find that the effect was significantly stronger among non-helpers, which is consistent with a role of empathy and/or hypocrisy aversion.

Next, we turned to punishment. We predicted Experiment 6 punishment as a function of a Punishment Only dummy, helping in the follow-up experiment, and their interaction. We found no significant interaction, OR = .83, z = -.84, p = .403, n = 2056. In other words, we failed to support the deflationary explanation's prediction that helping in the follow-up experiment should moderate the effect of helping opportunities on punishment, and found no statistical justification for investigating helpers and non-helpers separately.

However, because helpers showed a directionally smaller effect than non-helpers and the non-significant interaction could reflect limited power, we nonetheless analyzed each group separately. In these separate analyses, we found a non-significant positive effect of a Punishment Only dummy on punishment among helpers, OR = 1.15, z = 1.19, p = .234, n = 1261, and a marginally significant positive effect among non-helpers, OR = 1.40, z = 1.70, p = .090, n = 795. Thus, while our punishment analyses mirror our affective outrage analyses in terms of the directional effects observed, their more limited power makes them more equivocal: we failed to support the deflationary explanation's moderation prediction, but also were unable to demonstrate a significant effect of helping opportunities on punishment among helpers.

Comparing helpers and non-helpers in the context of our mediators. Finally, we note that our follow-up experiment only included subjects from Experiment 6, which did not measure

either of our reputation-relevant mediators. Thus, we cannot use our Analysis 4 approach to compare the effects of helping opportunities on our reputation-relevant mediators (or the indirect effects via our reputations-relevant mediators) among helpers and non-helpers.

Moreover, the simple approach of investigating these effects among helpers (or non-helpers) by comparing subjects helped (or did not help) to subjects who did not have the opportunity to help creates the same self-selection effect described above in the context of outrage. And like in the context of outrage, we find evidence consistent with the possibility that helpers and non-helpers differ in their baseline perceptions of the reputation value of punishment and general reputation concerns. Across the Condemnation+Helping conditions of our three experiments that measured PRBP, as compared to non-helpers, helpers reported that punishment would have significantly greater reputation value, $B = 0.09$, $t = 3.07$, $p = .002$. And across our three experiments that measured GRC, helpers reported significantly greater general reputation concerns than non-helpers, $B = 0.11$, $t = 3.80$, $p < .001$. Thus, comparing helpers to those who did not have the opportunity to help likely biases us away from finding the predicted negative effect of helping opportunities on our mediators (while comparing non-helpers to those who did not have the opportunity to help likely biases us towards finding the predicted negative effects).

Indeed, across our experiments that measured GRC, subjects in Condemnation Only reported greater general reputation concerns ($M = 2.99$, $SD = 0.96$) than Condemnation+Helping non-helpers ($M = 2.73$, $SD = 0.93$), $B = 0.11$, $t = 4.22$, $p < .001$, $n = 1521$), but not helpers ($M = 2.98$, $SD = 0.98$), $B = 0.004$, $t = 0.19$, $p = .847$, $n = 2127$). However, across our experiments that measured PRBP, subjects in Condemnation Only reported that punishment would have greater reputation value ($M = 3.77$, $SD = 2.43$) than both Condemnation+Helping non-helpers ($M = 2.89$, $SD = 2.44$), $B = 0.15$, $t = 5.99$, $p < .001$, $n = 1557$) *and* helpers ($M = 3.37$, $SD = 2.47$), $B =$

0.08, $t = 3.71$, $p < .001$, $n = 2094$). Because the self-selection effect likely biases us *against* finding this pattern in the context of helpers, this result provides evidence that helping opportunities reduced the perceived reputation value of punishing even among subjects who chose to help. It thus further supports our signaling theory, and its prediction that helping opportunities should reduce outrage and punishment even among helpers.

Discussion

Overall, Analysis 4 supports a role of a signaling-based mechanism for the effects of helping opportunities on affective outrage and punishment, and provides evidence that these effects were *not* solely driven by empathy or hypocrisy aversion among non-helpers. In the context of affective outrage, we supported our prediction that if follow-up experiment helping moderated the negative effect of helping opportunities, this effect would hold among helpers. And in the context of punishment, we did not find significant moderation. Our results thus matched the predictions outlined by our signaling account. We also report evidence suggesting that helping opportunities reduced the perceived reputation value of punishment among helpers. Together, these analyses are supportive of our signaling account.

We note, however, that we did not find a significant effect of helping opportunities on punishment when restricting our analyses to helpers; thus, our conclusions regarding punishment are somewhat equivocal, and future research should attempt to more precisely estimate the effect of helping opportunities on punishment among helpers and non-helpers.

Additionally, it remains possible that helping opportunities reduce outrage and punishment among helpers via mechanism(s) other than the reputation-based ones we have proposed. It seems unlikely that helping opportunities induce hypocrisy aversion among helpers, because helping and condemning others for not helping is not hypocritical. In contrast, however,

it is possible that when people are given the opportunity to help and chose to do so, they gain empathy for the perspective of non-helpers, decreasing outrage towards them.

As noted previously, however, an empathy mechanism would not predict the mediation results from Analysis 2. Furthermore, it is plausible that empathy could go in the *reverse* direction among helpers. Having the chance to help and choosing to do so could make the decision not to help seem *less* relatable, *decreasing* empathy towards non-helpers and thus *increasing* outrage and punishment. This possibility is consistent with evidence that people who have endured a hardship can be less likely to empathize with others enduring the same hardship, as compared to those who have no experience with the relevant situation (Ruttan, McDonnell, & Nordgren, 2015). If helping opportunities reduced empathy towards selfishness among helpers, this effect would actually *suppress* the observed negative effect of helping opportunities on outrage and punishment—such that the reported effects would underestimate the reputation-based mechanism we have proposed.

Adjudicating between these possibilities may be difficult, given that the causal pathway between outrage and empathy is likely bidirectional. If having the chance to help and choosing to do so reduces outrage towards non-helpers for reasons that do *not* relate to empathy (i.e., via our proposed reputation-based mechanism), this could plausibly increase reported empathy for non-helpers, making such a finding difficult to interpret. Nevertheless, future research should attempt to provide further insight into the role of empathy in shaping the effects of helping opportunities on outrage towards and punishment of non-helpers. It should also further investigate whether these effects occur through different processes among helpers and non-helpers.

Together, however, our results from Analyses 1-4 provide support for our theory that helping opportunities reduce outrage and punishment by reducing the signaling value of

punishment, and not merely by inducing hypocrisy aversion or empathy towards selfishness among non-helpers.

Analysis 5

In our fifth and final analysis, we tested the *heuristics* component of our reputation heuristics theory. To this end, we investigated whether deliberativeness moderated the effects of helping opportunities on outrage and punishment.

In general, deliberation allows people to tailor their behavior to the specific situation they are in, and thus can serve to inhibit typically-advantageous responses in atypical contexts where they will be costly (Kahneman, 2011; D. G. Rand et al., 2014; D. G. Rand et al., 2017; Shenhav et al., 2017; Stanovich, 2005). Consequently, when reputation is not actually at stake—but punishment *would* be an effective signal if reputation *were* at stake—we predicted that less deliberative individuals would show elevated levels of costly punishment, while this pattern would be attenuated or eliminated among more deliberative individuals. In other words, we predicted that less deliberative individuals would be more sensitive to helping opportunities in the context of our one-shot anonymous punishment experiments. In contrast, however, we predicted that deliberativeness would not moderate the effect of helping opportunities in contexts where reputation really *was* at stake.

To test these predictions, we investigated individual differences in deliberativeness. We drew on two distinct behavioral indicators of the extent to which subjects were likely to use deliberation during our experiment. First, we considered performance on questions assessing comprehension of incentives in our experiment, following the logic that individuals approaching our experiment more deliberatively should be more likely to carefully consider their current situation and incentives. Second, we considered performance on the Cognitive Reflection Task

(CRT) (Frederick, 2005), a set of math problems with intuitively compelling but incorrect answers designed to measure individual differences in deliberativeness.

Analysis 5a tested the prediction that across both of these indicators, less deliberative subjects would enact one-shot anonymous punishment at higher rates when helping was not possible, while more deliberative subjects would punish at relatively lower rates regardless of helping opportunities. Analysis 5b tested the prediction that deliberativeness would *not* moderate the influence of helping opportunities on punishment in experiments where reputation *was* actually at stake.

Finally, after confirming our prediction from Analysis 5a (that deliberativeness should attenuate the effect of helping opportunities on one-shot anonymous punishment), we sought in Analysis 5c to unpack the mechanism underlying this prediction. To this end, we investigated whether deliberativeness moderated the influence of helping opportunities in our (one-shot anonymous) outrage experiments. If more deliberative subjects are *always* less sensitive to reputation cues in one-shot anonymous interactions, deliberativeness *should* attenuate the influence of helping opportunities on outrage. In contrast, if deliberative subjects are specifically less sensitive to reputation cues when acting on such a sensitivity is *costly,* deliberativeness might *not* moderate the influence of helping opportunities on outrage. Because either possibility seemed consistent with our reputation heuristics theory, we did not approach Analysis 5c with a clear directional prediction.

Analysis 5a

In Analysis 5a, we tested our prediction that the effect of helping opportunities on one-shot anonymous punishment would be driven by relatively less deliberative decision-makers.

Methods. To this end, we investigated whether our two indicators of deliberativeness would moderate the influence of helping opportunities on punishment in our No TG punishment experiments. For our first indicator, we used the comprehension questions included in all of our experiments. For our second indicator, we used performance on the CRT. In Experiment 6, subjects completed the CRT at the beginning of the study. While Experiments 8-10 did not measure the CRT, we took advantage of the observation that CRT scores are fairly stable across time (Stagnaro, Pennycook, & Rand, 2018) to nonetheless obtain CRT scores for some subjects in those experiments by matching AMT IDs to an external dataset compiling other AMT experiments that included the CRT and were conducted by members of our research group (Stagnaro et al., 2018). These experiments all employed a version of the CRT that was conceptually identical to the one presented in Experiment 6, originally published in Frederick, 2005; however, there was some minor variation in the wording for some subset of the questions (e.g., "If it takes 10 seconds for 10 printers to print out 10 documents, how many seconds will it take 50 printers to print out 50 documents?" vs. "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?")

This dataset compiled 11 different sets of experiments, conducted between 2012 and 2017, and included 23,264 unique CRT scores from 17,999 unique subjects (as indexed by AMT IDs). Stagnaro et al. (2018) found that among subjects in this dataset who took the CRT more than once (i.e., because they participated in multiple included experiments), CRT scores increased over time (suggesting learning effects); thus, we considered only the chronologically first CRT score from each subject. Then, we identified matches between subjects in this CRT dataset (as indexed by AMT Worker IDs) and subjects in the No TG conditions of Experiments 8-10. This resulted in a sample of $n = 1672$ matches in the punishment conditions (i.e., excluding

subjects in Helping Only) of these experiments (n = 847 in Punishment Only, n = 825 in Punishment +Helping), M_{age} = 36.95 years, SD_{age} = 11.70 years, 48% male. When including Experiment 6, we had CRT data for a total of n = 4595 subjects in the punishment conditions of our No TG punishment experiments (n = 2313 in Punishment Only, n = 2283 in Punishment +Helping), M_{age} = 36.41 years, SD_{age} = 11.80 years, 46% male.

We note for completeness that in Experiments 6 and 9, our post-experimental survey included one item each from the Faith in Intuition and Need for Cognition scales (Epstein, Pacini, Denes-Raj, & Heier, 1996), which are conceptually related to deliberativeness; however, these single-item self-report measures correlated only weakly with comprehension and CRT performance and did not moderate the effect of helping opportunities on one-shot anonymous punishment. We focus on comprehension and CRT performance here, because (i) as multi-item measures they are more reliable than the single-item measures, and (ii) as behavioral measures of deliberateness, they—unlike the self-report measures—do not rely on subjects' introspection and are not susceptible to self-presentation concerns.

For both of our indicators of deliberativeness, we analyzed both the continuous measure (i.e., number of comprehension questions correct and number of CRT questions correct) as well as a median split on that continuous measure. These median split measures capture (i) whether all comprehension questions were correct (true of 59% of subjects) and (ii) whether at least 2 out of 3 CRT questions were correct (true of 41% subjects for whom we had CRT data). We also note that our continuous indicators were modestly positively correlated, r = .34, $p < .001$, supporting our premise that they are distinct but related indicators of deliberativeness.

Results. Did deliberativeness attenuate the influence of helping opportunities on one-shot anonymous punishment? To address this question, first we separately considered less versus

more deliberative subjects using our two median split indicators, and investigated the effect of helping opportunities on punishment (Table 5 rows 1-2). As predicted, across both indicators, only less deliberative subjects showed a significant effect of helping opportunities on punishment. Next, tested whether our deliberateness indicators significantly moderated the effect of helping opportunities on punishment. For each (continuous or median split) indicator, we separately predicted punishment as a function of condition, deliberativeness, and their interaction (Table 5 row 3). We found a significant negative interaction for all four indicators.

Statistic	**Comprehension** ***n*** **= 6076**		**CRT performance** ***n*** **= 4596**	
	Binary measure	Continuous measure	Binary measure	Continuous measure
Simple effect of Punishment Only (PO) dummy among **less deliberative** subjects	*OR* = 1.59 *z* = 5.50 *p* < .001		*OR* = 1.41 *z* = 4.13 *p* < .001	
Simple effect of PO among **more deliberative** subjects	*OR* = 1.13 *z* = 1.57 *p* = .117		*OR* = 1.06 *z* = .51 *p* = .606	
Interaction between PO and indicator of deliberativeness	*OR* = .71 *z* = -2.96 *p* = .003	*OR* = .85 *z* = -2.74 *p* = .006	*OR* = .75 *z* = -2.13 *p* = .033	*OR* = .89 *z* = -2.02 *p* = .043

Table 5. The effect of helping opportunities on punishment in one-shot anonymous interactions, as a function of deliberativeness. Reported sample sizes indicate the number of subjects for whom punishment and the relevant indicator of deliberativeness were both measured across our No TG punishment experiments.

Next, we further examined these interactions by computing simple slopes for the effect of helping opportunities on punishment at one standard deviation above and below the mean for both of our continuous deliberativeness indicators. We found that helping opportunities had a significant effect on punishment at one standard deviation below the mean on comprehension (*OR* = 1.50, *z* = 5.39, *p* < .001) and CRT (*OR* = 1.42, *z* = 3.97, *p* < .001) performance, but no

significant effect at one standard deviation above the mean on comprehension (OR = 1.11, z = 1.31, p = .191) and CRT (OR = 1.09, z = 0.85, p = .396) performance.

In Figure 4, we illustrate our results. Across our set of No TG punishment experiments, we plot punishment as a function of helping opportunities and our binary measures of comprehension (panel a) and CRT performance (panel b). Across both indicators, we see that less deliberative subjects were more likely to punish when helping is not possible. In contrast, more deliberative subjects were not sensitive to helping opportunities. Instead, they punished at relatively low rates regardless of whether helping was possible.

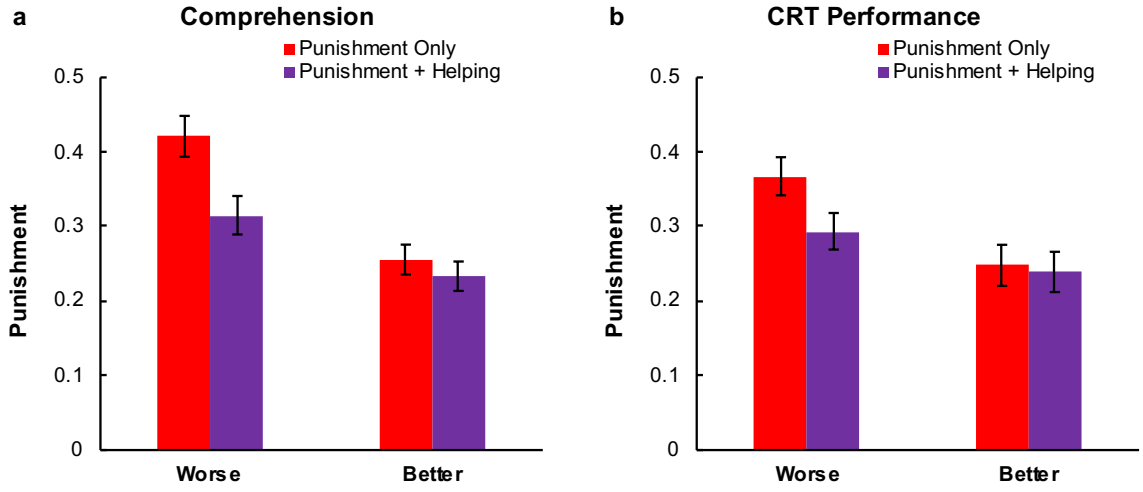

Figure 4. Deliberativeness moderates the influence of helping opportunities on one-shot anonymous punishment. We plot the proportion of subjects punishing as a function of helping opportunities and our median split indicators of deliberativeness. Error bars are 95% CIs.

Finally, we investigated the possibility that the results among our less deliberative subjects were driven exclusively by subjects who held the mistaken explicit belief that other players *could* observe their behavior and influence their payoffs. We again focused on Experiment 6, our only punishment experiment measuring these beliefs, and investigated the

effect of helping opportunities on punishment among below-median comprehension and CRT performers, excluding subjects who reported that other players could influence their payoffs.

In this analysis, we observed marginally significantly more punishment in Punishment Only than Punishment+Helping, both among below-median comprehension performers, $OR = 1.43$, $z = 1.75$, $p = .080$, $n = 538$, and among below-median CRT performers, $OR = 1.34$, $z = 1.79$, $p = .073$, $n = 863$. While these results are only marginally significant, we note that restricting this analysis to less deliberative subjects substantially reduces power. Thus, in conjunction with our Analysis 3 finding that our overall Experiment 6 punishment result is robust to excluding subjects who reported that other players could influence their payoffs, it seems unlikely that our less deliberative subjects were only sensitive to our manipulation because they held this mistaken explicit belief.

Together, Analysis 5a presents evidence that when reputation is not at stake, deliberativeness attenuates the influence of helping opportunities on costly punishment. This evidence is consistent with our theory that in one-shot anonymous interactions, subjects are sensitive to helping opportunities insofar as they rely on reputation heuristics.

Analysis 5b.

Our theory also predicts that when reputation *is* actually at stake, even more deliberative individuals—who rely less on heuristics—should be sensitive to helping opportunities. In such contexts, punishment really can confer reputation benefits, and really is more likely to do so when helping is not possible. Thus, deliberativeness should not moderate the effect of helping opportunities. To test this prediction, we analyzed a set of experiments where reputation *was* at stake, because the opportunity to punish (and/or help) was followed by a Trust Game where another player decided how much to trust the subject (based on his or her TPPG decisions).

Methods.

Design. Specifically, in Analysis 5b, we analyzed data from four experiments (see Table 8 for an overview of their designs). In each of these experiments, we employed the design from our No TG punishment experiments, except that the TPPG was followed by a Trust Game (TG). The TG involved two players: a Sender and a Receiver. The Sender was a new AMT worker who did *not* participate in the TPPG, and the Receiver was the target subject from the TPPG (i.e., the player who we focus on in this paper).

In the TG, the Sender was endowed with 30¢, and decided how much, if anything, to send to the Receiver; anything sent was tripled by the experimenter. Then, the Receiver decided how much of the amount sent to return to the Sender. In this game, Senders had an incentive to send more to Receivers who they trusted to return more. And critically, Senders could condition their sending on the Receiver's TPPG decision(s). Thus, TPPG decisions had reputation consequences. And these reputation consequences were financially meaningful to Receivers: the more money the Sender trusted them with, the more money they could potentially take home.

The first experiment we analyze in Analysis 5b, which we refer to here as Experiment 11, is the previously-mentioned (and published) experiment in Jordan et al, 2016a. The second experiment, which we refer to here as Experiment 12, is a previously unpublished exact replication of Experiment 11—albeit with a smaller sample size (determined prior to data collection). And the final two experiments are the TG conditions of Experiments 9-10, which were very similar to Experiments 11-12, except that in Experiment 10, (i) there was no Helping Only condition, and (ii) we showed subjects an example screenshot of how their TPPG decision(s) might be conveyed to the TG Sender.

Subjects. As noted previously, in Experiment 9 we requested a target of n = 400 subjects per condition (i.e., a total n =1200 subjects across the TG conditions), and in Experiment 10, we requested a target of n = 775 subjects per condition (i.e., a total n = 1550 subjects across the TG conditions). In Experiment 11 we also requested a target of n = 400 subjects per condition (i.e., a total of n =1200 subjects), and in Experiment 12, we requested a target of n = 200 subjects per condition (i.e., a total of n = 600 subjects). Our final sample of TG punishment experiments includes n = 4418 subjects (n = 1730 in Punishment Only, n = 1692 in Punishment+Helping, and n = 996 in Helping Only), M_{age} = 34.07 years, SD_{age} = 11.41 years, 46% male.

Procedure. The procedure was analogous to that of our No TG Punishment Experiments, but with the above-described design changes. After reading about the TPPG, subjects read about the TG and answered three TG comprehension questions. When they subsequently made their TPPG decision(s), they were reminded that the TPPG Sender would see these decision(s) before deciding how much to send them. Afterwards, they decided the percentage of the amount they were sent by the TG Sender to return to them (without actually learning this amount).

Indicators of deliberativeness. To investigate whether deliberativeness would moderate the effect of helping opportunities on punishment in our TG punishment experiments, we used the same two indicators as in Analysis 5a. When investigating comprehension, we thus considered only questions about the TPPG (which were identical to the comprehension questions in our No TG punishment experiments) and *not* questions about the TG. None of our TG punishment experiments directly measured CRT; thus, we relied on n = 1446 matches between the punishment conditions of our TG punishment experiments and the external CRT dataset (n = 714 in Punishment Only, n = 732 in Punishment +Helping), M_{age} = 36.56 years, SD_{age} = 11.73 years, 47% male. Our median split indicators of deliberativeness again captured (i) whether all

comprehension questions were correct (true for 61% of subjects) and (ii) whether at least 2 out of 3 CRT questions were correct (true of 49% subjects for whom we had CRT data). We also again found a moderate correlation between our continuous indicators of comprehension and CRT performance, $r = .31$, $p < .001$.

Results. Before investigating whether deliberativeness moderated the effect of helping opportunities on punishment, we asked whether there was a main effect of helping opportunities on punishment across our TG punishment experiments. Indeed, subjects in these experiments were significantly more likely to punish in the Punishment Only conditions (39%) than the Punishment+Helping conditions (30%), $OR = 1.55$, $z = 6.04$, $p < .001$, $n = 3422$. We also confirmed that punishment opportunities did not reciprocally influence helping in the subset of our TG punishment experiments that included a Helping Only condition. Indeed, subjects in these experiments helped at comparable rates in Helping Only (82%) and Punishment+Helping (81%), $OR = 1.04$, $z = .30$, $p = .766$, $n = 1927$. We also investigated only these experiments, and used linear regressions to predict both punishment and helping as a function of condition. We found that the standardized condition coefficient was significantly larger when predicting punishment ($B = .10$, $SE = .02$, $p < .001$) than when predicting helping ($B = .01$, $SE = .02$, $p = .766$), $z = 2.78$, $p = .006$.

Thus, within our TG punishment experiments, we replicated the findings that helping opportunities reduced punishment, but punishment opportunities did not reduce helping. Next, we tested our key prediction that deliberativeness should *not* moderate the influence of helping opportunities on punishment in these experiments. We used the same approach as in Analysis 5b, reported our results in Table 6, and illustrated them in Figure 5. As predicted, for both of our median split indicators of deliberativeness, both less and more deliberative subjects were more

likely to punish when helping was not possible. Furthermore, we observed no significant interactions between helping opportunities and any indicator of deliberativeness. And when we computed simple slopes for the effect of helping opportunities on punishment at one standard deviation above and below the mean for both of our continuous deliberativeness indicators, we found significant effects at one standard deviation below the mean on comprehension (OR = 1.53, z = 4.06, p < .001) and CRT (OR = 1.51, z = 2.54, p = .011) performance, and at one standard deviation above the mean on comprehension (OR = 1.57, z = 4.42, p < .001) and CRT (OR = 1.47, z = 2.41, p = .016) performance.

Statistic	**Comprehension** ***n*** **= 3422**		**CRT performance** ***n*** **= 1446**	
	Binary measure	Continuous measure	Binary measure	Continuous measure
Simple effect of Punishment Only (PO) dummy among **less deliberative** subjects	*OR* = 1.49 *z* = 3.39 *p* = .001		*OR* = 1.49 *z* = 2.53 *p* = .012	
Simple effect of PO among **more deliberative** subjects	*OR* = 1.59 *z* = 5.07 *p* < .001		*OR* = 1.47 *z* = 2.40 *p* = .016	
Interaction between PO and measure of deliberativeness	*OR* = 1.07 z = .46 *p* = .647	*OR* = 1.01 *z* = .18 *p* = .858	*OR* = .98 *z* = -.09 *p* = .929	*OR* = .99 *z* = -.12 *p* = .906

Table 6. The effect of helping opportunities on punishment when reputation is at stake, as a function of deliberativeness.

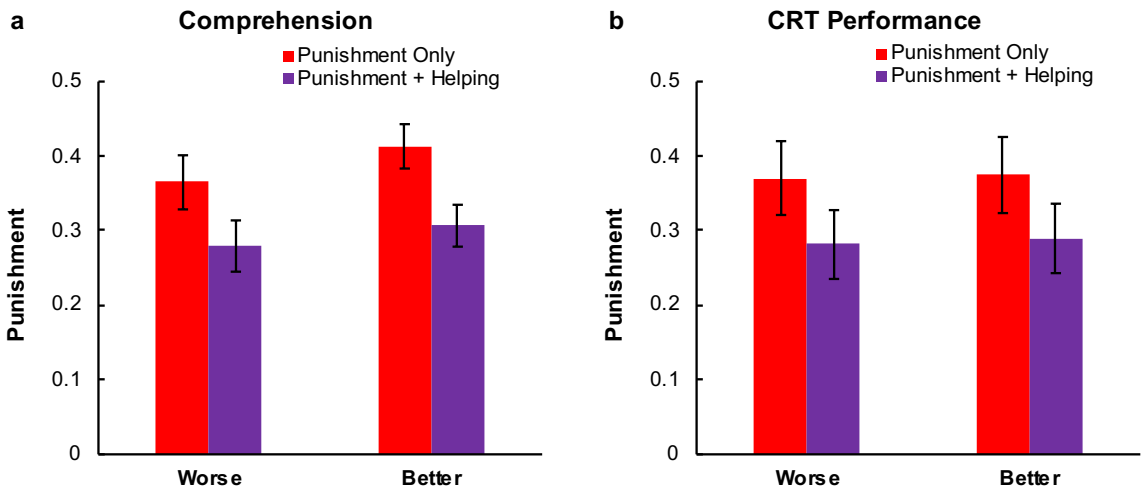

Figure 5. Deliberativeness does not moderate the influence of helping opportunities on punishment when reputation is at stake. We plot the proportion of subjects punishing as a function of helping opportunities and our median split indicators of deliberativeness. Error bars are 95% CIs.

Thus, deliberativeness did *not* undermine the influence of helping opportunities on costly punishment when there was an explicit strategic reason to appear trustworthy. Rather, all subjects were more likely to punish when helping was not possible, regardless of deliberativeness. This finding rules out the possibility that more deliberative subjects are never sensitive to helping opportunities, and documents their sensitivity in a context where it can confer strategic benefits: our TG punishment experiments.

We can also directly compare our TG and no TG punishment experiments by investigating the three-way interactions between helping opportunities, deliberativeness, and the presence of a Trust Game. For each (continuous or median split) indicator of deliberativeness, we predicted punishment as a function of helping opportunities, the deliberativeness indicator, a dummy indicating whether there was a TG, and all two- and three-way interactions. We found that the three-way interaction term was in the predicted direction for all measures, and was significant for our median split measure of comprehension ($OR = 1.51$, $z = 2.17$, $p = .030$, $n =$

9498), but only marginally significant for our continuous measure of comprehension (*OR* = 1.19, $z = 1.75$, $p = .081$, $n = 9498$), and non-significant for our median split (*OR* = 1.31, $z = 1.01$, $p = .311$, $n = 6042$) and continuous (*OR* = 1.11, $z = 0.91$, $p = .362$, $n = 6042$) measures of CRT.

Thus, we found some (albeit weak) evidence of the three-way interaction implied by the significance of the two-way interactions in No TG condition, and the non-significance of the two-way interactions in the TG condition, reported above. The non-significant three-way interaction terms here may reflect a lack of power—even though our sample sizes seem very large, a great deal of power is needed to detect a three-way interaction in the context of a relatively small simple effect on a binary dependent variable. Moreover, because we do not have CRT data for roughly 1/3 of our subjects, we have less power to detect a three-way interaction for our CRT indicators of deliberativeness than for the comprehension indicators.

This possibility is supported by power calculation simulations (described in detail in the SM) investigating our ability to detect a three-way interaction between helping opportunities, our binary deliberativeness measures, and the presence of a Trust Game. These simulations indicate that, even if more deliberative subjects in the No TG condition showed *no* more punishment in Punishment Only than in Punishment+Helping, rates of punishment would have to be 14 percentage points higher in Punishment Only among all other groups of subjects (i.e., less deliberative subjects in No TG, and all subjects in TG) in order to generate 80% power at $n = 750$ per cell (roughly the sample size in our CRT analyses). This would be a rather sizeable simple effect to be entirely eliminated among deliberative subjects when reputation is not at stake: as illustrated in Figures 4-5, we generally observed a baseline of about 30 percent of subjects punishing in Punishment+Helping, and therefore a 14 percentage point simple effect would be an almost 50% increase in punishment.

Thus, the non-significant results for CRT suggest that it is unlikely that the true three-way interaction effect is that large. However, observing our data would not be especially surprising if there *was* actually a three-way interaction, but either the baseline simple effect of helping opportunities was smaller than 14 percentage points and/or that effect was not completely attenuated among more deliberative individuals in the No TG condition. Thus, although our results do not provide strong evidence in support of the hypothesized three-way interaction, they do not provide strong evidence against a meaningfully sized three-way interaction. Furthermore, our power simulations reveal that even with a larger sample size of n = 1200 per cell (roughly our sample size for our comprehension analyses), we are not that well-powered to the three-way interaction (see SM for details). Thus, like with CRT, the relatively weak evidence for a three-way interaction for comprehension does not place an especially small upper bound on the possible true effect size.

Overall, we argue that the results provided in this section provide tentative support for the hypothesis that when making costly punishment decisions, deliberative individuals are specifically less sensitive to reputation cues in contexts where reputation is not at stake.

Analysis 5c

But why were deliberative individuals insensitive to reputation cues in the context of one-shot anonymous punishment, as observed in Analysis 5a? To address this question, we returned to our (one-shot anonymous) outrage experiments, and investigated whether our indicators of deliberativeness moderated the influence of helping opportunities on outrage. If more deliberative individuals are *always* insensitive to reputation cues in one-shot anonymous interactions, deliberativeness should also have attenuated the influence of helping opportunities on reported outrage. In contrast, if deliberative individuals specifically inhibit their sensitivity to

reputation cues when acting on such a sensitivity is *costly,* it is possible that deliberativeness did *not* attenuate the influence of helping opportunities on outrage, which was costless to express in our experiments. Because either possibility seemed consistent with our reputation heuristics theory, we did not approach Analysis 5c with a clear directional prediction.

Methods. We used the same two deliberativeness indicators as in Analyses 5a-b, except that we only considered the *two* comprehension questions included in all of our outrage experiments (rather than the four in our punishment experiments). We note, however, that we found qualitatively identical results when re-analyzing our punishment experiments considering only these two questions. When investigating CRT performance, we found $n = 1576$ matches between the outrage conditions of Experiments 1-5 and 7 (which did not measure CRT) and our CRT dataset ($n = 785$ in Condemnation Only, $n = 791$ in Condemnation +Helping), $M_{age} = 37.85$ years, $SD_{age} = 11.97$ years, 46% male). Thus, when including Experiment 6 (which did measure CRT), we had CRT data for a total of $n = 4500$ subjects in the outrage conditions of our outrage experiments ($n = 2251$ in Condemnation Only, $n = 2249$ in Condemnation +Helping), $M_{age} = 36.72$ years, $SD_{age} = 11.92$ years, 45% male).

Our median split indicators of deliberativeness captured (i) whether both comprehension questions about the helping decision were correct (true for 82% of subjects) and (ii) whether at least 2 out of 3 CRT questions were correct (true of 41% subjects for whom we had CRT data). We again found a modest correlation between our continuous measures of comprehension and CRT performance, $r = .22$, $p < .001$.

Results. To investigate whether deliberativeness moderated the influence of helping opportunities on outrage, we used an analogous approach to Analyses 5a-b. Our results are reported in Table 7, and illustrated in Figure 6. For both of our median split indicators of

deliberativeness, we observed a significant effect of helping opportunities on outrage among both more and less deliberative subjects. Thus, in one-shot anonymous interactions, more deliberative subjects *did* report heightened outrage when helping was not possible. Unexpectedly, in fact, we observed that more deliberative subjects actually showed directionally *larger* effects of helping opportunities than less deliberative subjects, although we observed no significant interactions between our deliberativeness indicators and helping opportunities.

Next, we computed simple slopes for the effect of helping opportunities on outrage at one standard deviation above and below the mean for both of our continuous deliberativeness indicators. We found significant effects at one standard deviation below the mean on comprehension ($B = .08$, $t = 5.21$, $p < .001$) and CRT ($B = .08$, $t = 3.69$, $p < .001$) performance, and at one standard deviation above the mean on comprehension ($B = .09$, $t = 6.02$, $p < .001$) and CRT ($B = .11$, $t = 5.11$, $p < .001$) performance.

Overall, then, helping opportunities *did* influence outrage among more deliberative (as well as less deliberative) individuals. We also unexpectedly found that more deliberative individuals showed directionally *larger* effects of helping opportunities on outrage—the opposite pattern as we observed in the context of punishment—but did not observe significant interactions.

Statistic	**Comprehension** ***n*** **= 8440**		**CRT performance** ***n*** **= 4500**	
	Binary measure	Continuous measure	Binary measure	Continuous measure
Simple effect of Condemnation Only (CO) dummy among **less deliberative** subjects	$B = .07$ $t = 2.84$ $p = .005$		$B = .07$ $t = 3.63$ $p < .001$	
Simple effect of CO among **more deliberative** subjects	$B = .09$ $t = 7.40$ $p < .001$		$B = .12$ $t = 5.29$ $p < .001$	

Interaction between CO and measure of deliberativeness	$B = .01$ $t = 0.40$ $p = .690$	$B = .02$ $t = 0.57$ $p = .566$	$B = .04$ $t = 1.56$ $p = .118$	$B = .03$ $t = 1.01$ $p = .312$

Table 7. The effect of helping opportunities on outrage in one-shot anonymous interactions, as a function of deliberativeness.

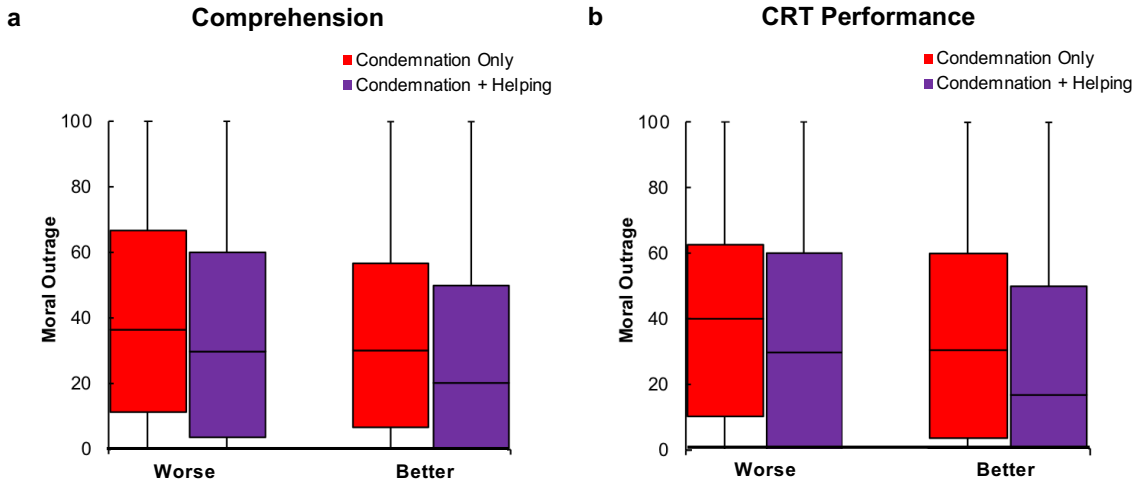

Figure 6. Deliberativeness does not significantly moderate the influence of helping opportunities on outrage in one-shot anonymous interactions. We show box plots (which draw lines at the 25th, 50th, and 75th percentiles, and illustrate the minimum and maximum values) for moral outrage as a function of helping opportunities and our binary indicators of deliberativeness.

Discussion

Together, Analysis 5a-c suggest that deliberativeness attenuates the influence of helping opportunities on punishment in one-shot anonymous interactions, but not on punishment when reputation is actually at stake, and not on reported outrage in one-shot anonymous interactions.

It is interesting that in one-shot anonymous interactions where it was not possible to help, more deliberative individuals reported heightened outrage but were not more likely to pay to punish. This pattern suggests that even when reputation is not at stake, deliberative individuals are not *always* insensitive to reputation cues. It is also consistent with a large body of evidence

that, depending on the individual and the situation, a particular emotional experience can give rise to many different behavioral expressions—or no expression at all (Roseman, 2011; Roseman, Wiest, & Swartz, 1994). In the context of our experiments, deliberativeness seems to be one individual difference that is relevant to whether a context that increases outrage (or the drive to report outrage) *also* increases costly punishment behavior.

Generally, one important reason for the limited correspondence between emotion feelings (like outrage) and emotion-related behaviors (like punishment) is that people can regulate their emotions (Gross, 1998b), and there are substantial individual differences in when, how, and whether emotion regulation occurs (Gross & John, 2003). In our experiments, more deliberative individuals may have engaged in emotion regulation that hampered their punishment behavior, but not their experience of or drive to report outrage.

Given that punishing was costly and outrage was costless to report, this pattern may reflect that deliberative individuals specifically regulate their sensitivity to reputation cues when such a sensitivity will be *costly*. This explanation would be consistent with the proposal that emotions constitute adaptive response *tendencies*, but that these tendencies are not always optimal for a situation and thus need to be regulated (Gross, 1998b). In line with this proposal, it is hypothesized that deliberation has the function of preventing typically-advantageous behaviors in atypical contexts where they are costly (Kahneman, 2011; D. G. Rand et al., 2014; D. G. Rand et al., 2017; Shenhav et al., 2017; Stanovich, 2005). An interesting question concerns the process through which deliberative individuals regulate their sensitivity to reputation cues when making one-shot anonymous punishment decisions. Like non-deliberative subjects, deliberative subjects reported heightened outrage when helping was not possible, so what process prevented them from enacting more costly punishment?

One possibility is that they were driven to enact more punishment but inhibited that drive. This mechanism is consistent with evidence that people often engage in the "response-focused" emotion regulation strategy of *suppressing* emotion-related behaviors, despite being driven to engage in them (Gross, 1998a, 1998b; Gross & John, 2003). For example, deliberative subjects who lacked the opportunity to help might have chosen not to punish—despite a relatively strong drive to do so—because they reasoned that punishing would be costly and would not materially benefit them. Or, they might have suppressed their drive to punish by constructing self-serving moral justifications (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009) (e.g., by reasoning that punishing would actually be morally wrong because the non-helper likely really needed the money, or because punishing is a destructive action that only serves to harm others). Such processes could suppress the drive to engage in a typically-advantageous behavior in an atypical context where it is costly.

Alternatively, it is possible that when reputation was not at stake, deliberative subjects who did not have the opportunity to help were *not* driven to enact heightened punishment, despite reporting heightened outrage. This would imply that for these subjects, while our manipulation altered the experience of or drive to report outrage, it did not alter the drive to punish. This mechanism is consistent with evidence that people often engage in the "antecedent-focused" emotion regulation strategy of *cognitive reappraisal*, which involves thinking about a situation differently so as to change one's emotional experience—in this case, their affective drive to punish (Gross, 1998a, 1998b; Gross & John, 2003). Under this scenario, insofar as deliberative subjects reasoned that punishing was personally costly or morally wrong, these processes would have served to prevent them from ever feeling driven to punish (rather than to

help them suppress that drive). Future research should attempt to discriminate between these potential mechanisms.

A related question for future research pertains to the psychological process through which helping opportunities *did* influence punishment when reputation *was* at stake (as observed in Analysis 5b). In such contexts, did subjects explicitly reason about the reputation value of punishment, motivating their sensitivity to helping opportunities? Or did they rely on emotional feelings like outrage, or the affective drive to punish, in the absence of strategic reasoning? And did the answer vary with deliberativeness?

Exp.	Key design features	Conditions	Analyzed variables measured (in order)	Sample size	Methodological notes	In which analyses?
1	-Reputation not at stake -Key DV: outrage	-Condemnation Only -Condemnation+Helping (counterbalanced) -Helping Only	-Comprehension -Outrage	1195 (~400/ condition)		1, 5c
2		-Condemnation Only -Condemnation+Helping (helping first)	-Comprehension -Outrage -PRBP	819 (~400/ condition)		1,2,5c
3			-Comprehension -Outrage -GRC	817 (~400/ condition)		
4			-Comprehension -Outrage -PRBP & GRC in random order	811 (~400/ condition)		
5				804 (~400/ condition)		
6	-Reputation not at stake -Key DVs: affective outrage and punishment	-Condemnation / Punishment Only -Condemnation / Punishment+Helping (helping first)	-CRT -Comprehension (for outrage task) -Affective outrage -Comprehension (for punishment task) -Punishment -Beliefs re: can other players affect payoff? -Other-, self-, and experimenter-signaling concerns -Follow-up experiment helping (if participated)	2924 (~1500/ condition)	-Outrage-rating task framed more neutrally than in Exps 1-5 -Transgression had already occurred (i.e., was not hypothetical) -Filler memory task between measurement of affective outrage and punishment	1,3,4,5a,5c
7	-Reputation not at stake -Key DV: outrage	-Condemnation Only -Condemnation+Helping (helping first)	-Comprehension -Outrage -Beliefs re: can other players affect payoff?	1447 (~750/ condition)	-Transgression had already occurred (i.e., was not hypothetical) -Wording for beliefs question modified from Exp. 6	1, 5c
8	-Reputation not at stake -Key DV: punishment	-Punishment Only -Punishment+Helping (counterbalanced) -Helping Only	-Comprehension -Punishment	1160 (~400/ condition)		3,5a

9	-Manipulated if reputation at stake -Key DV: punishment	-Punishment Only **vs** Punishment+Helping (counterbalanced) **vs** Helping Only **X** -TG **vs** No TG		2331 (~400/ condition)		3,5a,5b
10		-Punishment Only **vs** Punishment+Helping (counterbalanced) **X** -TG **vs** No TG		3104 (~775/ condition)	-Reputation condition emphasized how subjects' decisions would look to TG senders	
11	-Reputation at stake -Key DV: punishment	-Punishment Only -Punishment+Helping (counterbalanced) -Helping Only		1199 (~400/ condition)	-Previously published and re-analyzed here (Jordan et al., 2016)	5b
12				563 (~200/ condition)		

Table 8. Overview of experiments. For each experiment, we report the key design features (specifically, whether reputation was at stake and the key DV(s)), experimental conditions (along with counterbalancing information), measured variables that were analyzed, final sample size (as well as the approximate number per condition, which was the target number recruited), methodological notes, and analyses the experiment was included in.

General Discussion

Across five analyses of twelve different experiments, we have provided evidence that (i) moral outrage is influenced by cues of the potential signaling value of punishment, and (ii) these cues also influence one-shot anonymous punishment among less deliberative individuals. Together, our results suggest that a reputation framework—and specifically the hypothesis that punishment serves to signal trustworthiness—can shed light on when and why people express outrage and incur personal costs to punish wrongdoing, even when reputation is not actually at stake. They thus contribute to our understanding of key features of human morality, and have numerous theoretical implications.

A reputation heuristics account of one-shot anonymous punishment

First, our results support a reputation heuristics account of one-shot anonymous punishment. We found that helping opportunities influenced one-shot anonymous punishment, but not among more deliberative individuals. This pattern provides insight into why, from an ultimate perspective, less deliberative individuals were sensitive to helping opportunities in a context where reputation was not at stake. Our results suggest that these individuals relied on the heuristic that reputation is typically at stake in order to avoid the cognitive (Bear et al., 2017; Bear & Rand, 2016) and/or social (Critcher et al., 2013; M. Hoffman et al., 2015; Jordan et al., 2016b) costs of constantly calculating who is currently watching. If less deliberative individuals had instead been sensitive to helping opportunities because it is actually optimal to attend to reputation cues even when reputation appears not to be at stake (e.g, as an error management strategy; Delton et al., 2011), we would expect more deliberative individuals to have shown the same sensitivity.

Thus, our results suggest that one-shot anonymous punishment reflects a reputation *heuristic*. They therefore contribute to and extend evidence that social heuristics shape moral decision-making (Kiyonari, Tanida, & Yamagishi, 2000). In particular, previous research has provided evidence that one-shot anonymous *cooperation* can reflect the heuristic that interactions are typically repeated or observed (Bear & Rand, 2016; Everett, Ingbretsen, Cushman, & Cikara, 2017; D. Rand, Greene, & Nowak, 2012; D. G. Rand, 2016), and our results extend this evidence to the domain of punishment.

An interesting open question is how, from a proximate psychological perspective, reputation heuristics are implemented in contexts where reputation is not at stake. What kinds of reputation concerns do people have, and what makes them sensitive to cues of the potential reputation value of their possible actions? Experiment 6 provided some preliminary evidence that

our subjects may have been concerned about looking good in the eyes of generically described "others" (possibly reflecting their imagination about how potential observers would evaluate their behavior), as well as in the eyes of the experimenter and in their own eyes. Future research should investigate the relative contributions of these different reputation motivations.

Critically, however, our reputation heuristics hypothesis makes a unique prediction that is independent of the particular reputational concern(s) that are at play: in one-shot anonymous interactions, deliberative individuals should be relatively unwilling to pay costs to act on those concerns. By supporting this prediction, our results provide interesting context for the possibility that helping opportunities influenced outrage and punishment because people were concerned about being viewed positively by others, the experimenter, or themselves. Specifically, our results suggest that among more deliberative individuals, the reputation concerns underlying their sensitivity to helping opportunities did not persist in contexts where reputation was not actually at stake, and acting on them would be costly.

Our results also have implications for *when* social heuristics are most likely to motivate typically-advantageous behaviors in atypical contexts. We found that less deliberative individuals engaged in more one-shot anonymous punishment than more deliberative individuals, supporting a reputation heuristics account of one-shot anonymous punishment. But this pattern was much stronger when helping was not possible (and thus punishment, if observed, would have been an effective signal of trustworthiness). This may suggest that in atypical contexts more generally, social heuristics are most likely to motivate typically-advantageous behaviors when they *would* be advantageous in typical contexts.

Related, our results suggest that despite relying on a reputation heuristic, less deliberative individuals were nonetheless sensitive to whether helping is possible. This may imply that it is

fairly cognitively demanding or socially costly to determine that reputation is not at stake—but *less* demanding or costly to determine that if reputation *were* at stake, punishment would have limited reputation value because helping would also be observable. It also raises the important future question of which reputation cues people who rely on reputation heuristics are and are not sensitive to in contexts where nobody is watching.

The nature and functions of moral outrage

In addition to supporting a social heuristics hypothesis for one-shot anonymous punishment, our results have theoretical implications for the nature and function of moral outrage. Introspection suggests that we experience outrage as a private and genuine response to wrongdoing that simply indexes the magnitude of immorality that has occurred. But our results suggest that the experience of (or drive to report) outrage also tracks the reputation benefits we may gain from punishing. This proposal is not mutually exclusive with the idea that outrage is experienced genuinely, but it supports theories of emotions as adaptive motivators of action (Cosmides & Tooby, 2000; Fredrickson, 2001; Frijda, 1986; Lazarus, 1991), and moral outrage specifically as a motivator of punishment (Carlsmith et al., 2002; Darley & Pittman, 2003; Fessler & Haley, 2003; Fiske & Tetlock, 1997; Goldberg et al., 1999; Jordan et al., 2015).

Additionally, because moral outrage appears to track the potential reputation value of punishment even when reputation is not at stake, our results are consistent with theories that moral emotions and judgements are usually not caused by reasoning (Haidt, 2001), and can "misfire" in contexts where they are not adaptive (Greene, 2014; Gross, 1998b; Haidt, 2001; Inbar, Pizarro, Knobe, & Bloom, 2009; Kahneman, 2011). Moreover, we find that when reputation is not at stake, deliberative individuals are sensitive to reputation cues when reporting outrage but not when enacting punishment. This result that is consistent with evidence that

emotion feelings do not always translate to emotion-related behaviors (Roseman, 2011; Roseman et al., 1994), and that these gaps can reflect that some individuals respond to "misfiring" by adaptively engaging in emotion regulation (Gross, 1998b; Gross & John, 2003).

Implications for moral licensing

Our outrage and punishment results also connect to a large body of work on moral licensing (Monin & Miller, 2001). Moral licensing refers to a phenomenon in which engaging in one moral behavior makes an individual feel free to subsequently behave less morally. Licensing effects have been documented in the context of political correctness, prosocial behavior, and consumer choice (Merritt et al., 2010), and are often discussed as reflecting self-concept maintenance motives.

In our experiments, the effects of helping opportunities on punishment and outrage among subjects who chose *to* help may be thought of as extending licensing effects to the domains of punishment, as well as emotions and judgements. As discussed in Analysis 2, subjects in our one-shot anonymous experiments may have been concerned with their self-concepts. And if choosing to help (an act of morality that is straightforward and "positive") reduces the probability of punishing wrongdoing (another act of morality, albeit one that is less straightforward and more "negative"), it may plausibly reflect that helping makes people feel licensed not to punish. Moreover, helping opportunities also reduced *outrage*, suggesting that licensing effects may extend to the domains of emotions and judgements.

Importantly, however, we also found evidence that helping opportunities reduced punishment and outrage among subjects who *declined* to help. Declining to help should *not* affirm an individual's positive moral self-concept, and thus should *not* make an individual feel licensed to not punish. Thus, the observed effects among non-helpers are unlikely to have

reflected a licensing psychology, and are also inconsistent with the broader theory of *moral balancing* (Mullen & Monin, 2016). Moral balancing proposes that while moral behavior should license subsequent immorality, immoral behavior should induce compensation efforts—increasing subsequent morality. Thus, balancing predicts that having the opportunity to help should *increase* punishment among non-helpers, which is the opposite of what we found. Instead, our results are consistent with our signaling theory, which proposes that declining to help sends a strong signal of untrustworthiness—reducing the reputation value of punishment and thus the probability that it will occur. In other words, our signaling theory makes a prediction that contrasts with balancing theory, and which was born out in our data.

Moreover, our signaling theory and results may help shed light on the factors that moderate licensing effects after people behave morally. We have proposed a reputation-based explanation for why choosing to help reduces outrage and punishment. And we have supported this proposal by showing that deliberative individuals cease to be sensitive to helping opportunities when reputation is not at stake, and reacting to wrongdoing is costly (i.e., in our punishment but not our outrage experiments). Might this moderation pattern extend to licensing effects more generally? Our reputation theory predicts that (i) licensing may occur whenever engaging in an initial moral act reduces the reputation value of a subsequent moral act, and (ii) deliberative individuals may not show these licensing effects whenever reputation is not at stake and the subsequent moral act is costly.

Future Directions

Our experiments investigated moralistic punishment and outrage in the context of one canonical, but relatively minor, act of selfishness. Specifically, we measured reactions to an AMT worker who declined to share money with another AMT worker. On the one hand, the fact that this straightforward transgression is not embedded in rich contextual details suggests that our results may be likely to generalize to other transgressions. On the other hand, its relatively minor nature raises the question of whether our results would generalize to more severe moral violations. We used the term "moral outrage" in this paper to refer to the set of affective, cognitive, and behavioral responses people have to wrongdoing. However, subjects' reactions would probably not be colloquially described as "outraged", given their relatively low absolute ratings on our scale. Future research should investigate the effect of reputation cues on moralistic outrage and punishment in response to a more diverse set of transgressions, including those that are more extreme, and that are more concrete and realistic.

Another important direction for future research is investigating the influence of reputation heuristics on outrage and punishment across cultures. Our experiments are all conducted via AMT and only investigate American subjects, raising questions about generalizability (Henrich, Heine, & Norenzayan, 2010b). Research investigating moralistic punishment across cultures has demonstrated that it is widespread, and that punishment of selfishness seems to universally increase with the severity of selfishness (Henrich et al., 2006). Nonetheless, the prevalence of moralistic punishment varies considerably, and the different mechanisms (both proximate and ultimate) driving punishment across cultures remain unclear. Is there substantial cross-cultural variation in the extent to which punishment serves to signal trustworthiness, and in the extent to which signaling cues influence punishment even when

reputation is not actually at stake? And might such variance correlate with the prevalence of punishment? Future research should address these important questions.

Conclusion

Third-party punishment is central to human morality, and plays a key role in promoting cooperation. But from an ultimate perspective, it is also puzzling, especially in the context of one-shot anonymous interactions: why should we make personal sacrifices to punish wrongdoing towards others? Our results support the theory that even in such contexts, some people reply on the heuristic that reputation is *typically* at stake. As a result, even when reputation is not actually at stake, reputation cues can shape moral outrage—and, among less deliberative individuals, costly punishment. Our results thus demonstrate how a reputation framework can shed light on these key features of human morality.

References

Aquino, K., & Reed, I. (2002). The self-importance of moral identity. *Journal of personality and social psychology, 83*(6), 1423.

Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters, 122*(2), 308-310.

Balafoutas, L., & Nikiforakis, N. (2012). Norm enforcement in the city: a natural field experiment. *European Economic Review, 56*(8), 1773-1785.

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological bulletin, 137*(4), 594.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325-344.

Barden, J., Rucker, D. D., & Petty, R. E. (2005). "Saying one thing and doing another": Examining the impact of event order on hypocrisy judgments of others. *Personality and Social Psychology Bulletin, 31*(11), 1463-1474.

Baron, J., & Ritov, I. (1993). Intuitions about penalties and compensation in the context of tort law. *Journal of Risk and Uncertainty, 7*(1), 17-33.

Batson, C. D., Kennedy, C. L., Nord, L. A., Stocks, E., Fleming, D. Y. A., Marzette, C. M., et al. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology, 37*(6), 1272-1285.

Bear, A., Kagan, A., & Rand, D. G. (2017). Co-evolution of cooperation and cognition: the impact of imperfect deliberation and context-sensitive intuition. *Proc. R. Soc. B, 284*(1851), 20162326.

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531-3535.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizeable groups. *Ethology and Sociobiology, 13*(3), 171-195.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313-7318.

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*(2), 284.

Charness, G., Cobo-Reyes, R., & Jimenez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior & Organization, 68*(1), 18-28.

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. *Handbook of emotions, 2*(2), 91-115.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*(3), 308-315.

Crockett, M. (2013). Models of morality. *Trends in cognitive sciences*.

Crockett, M., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General, 143*(6), 2279.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review, 17*(3), 273-292.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS one, 4*(8), e6699.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review, 7*(4), 324-336.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*.

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences, 108*(32), 13335-13340.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*(7185), 348-351.

Effron, D. A., Lucas, B. J., & O'Connor, K. (2015). Hypocrisy by association: When organizational membership increases condemnation for wrongdoing. *Organizational Behavior and Human Decision Processes, 130*, 147-159.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of personality and social psychology, 71*(2), 390.

Everett, J. A., 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.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63-87.

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological science, 25*(3), 656-664.

FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J., & Phelps, E. A. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature communications, 5*.

Fessler, D. M., & Haley, K. J. (2003). The Strategy of Affect: Emotions in Human Cooperation 12.

Fiske, A. P., & Tetlock, P. E. (1997). Taboo trade-offs: reactions to transactions that transgress the spheres of justice. *Political psychology, 18*(2), 255-297.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives, 19*(4), 25-42.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American psychologist, 56*(3), 218.

Frijda, N. H. (1986). *The emotions*: Cambridge University Press.

Goette, L., Huffman, D., & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *The American economic review*, 212-216.

Goldberg, J. H., Lerner, J. S., & Tetlock, P. E. (1999). Rage and reason: The psychology of the intuitive prosecutor. *European Journal of Social Psychology, 29*(56), 781-795.

Greene, J. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*: Penguin.

Gromet, D. M., Okimoto, T. G., Wenzel, M., & Darley, J. M. (2012). A victim-centered approach to justice? Victim satisfaction effects on third-party punishments. *Law and Human Behavior, 36*(5), 375.

Gross, J. J. (1998a). Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *Journal of personality and social psychology, 74*(1), 224.

Gross, J. J. (1998b). The emerging field of emotion regulation: an integrative review. *Review of general psychology, 2*(3), 271.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology, 85*(2), 348.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review, 108*(4), 814.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences, 11*(2003), 852-870.

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, 108*(50), 19931-19936.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of theoretical biology, 208*(1), 79-89.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., et al. (2010a). Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. *Science, 327*(5972), 1480-1484.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and brain sciences, 33*(2-3), 61-83.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767-1770.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362-1367.

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences, 112*(6), 1727-1732.

Hoffman, M. L. (2001). *Empathy and moral development: Implications for caring and justice*: Cambridge University Press.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of personality and social psychology, 97*(6), 963.

Horita, Y. (2010). Punishers may be chosen as providers but not as recipients. *Letters on Evolutionary Behavioral Science, 1*(1), 6-9.

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of personality and social psychology, 100*(4), 719.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion, 9*(3), 435.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1553), 2635-2650.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. (2016a). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473-476.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016b). Uncalculating Cooperation is used to Signal of Trustworthiness. *PNAS, 113*(31), 8658-8663.

Jordan, J. J., McAuliffe, K., & Rand, D. (2015). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*.

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, 111*(35), 12710-12715.

Jordan, J. J., & Rand, D. (2017). Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology, 421*, 189-202.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling. *Psychological science, 28*(3), 356-368.

Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.

Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: confusion or a heuristic? *Evolution and human behavior, 21*(6), 411-427.

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological science*, 0956797615624469.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*(2), 75-84.

Lazarus, R. S. (1991). *Emotion and adaptation*: Oxford University Press on Demand.

Leary, M. R. (1983). A brief version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin, 9*(3), 371-375.

Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology, 47*(2), 477-480.

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences, 108*(28), 11375-11380.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research, 45*(6), 633-644.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition, 134*, 1-10.

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and personality psychology compass, 4*(5), 344-357.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of personality and social psychology, 81*(1), 33.

Montada, L., & Schneider, A. (1989). Justice and emotional reactions to the disadvantaged. *Social Justice Research, 3*(4), 313-344.

Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual review of psychology, 67*.

Nelissen, R. (2008). The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior, 29*(4), 242-248.

Nelissen, R., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making, 4*(7), 543-553.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics, 92*(1), 91-112.

Ohtsuki, H., Iwasa, Y., & Nowak, M. A. (2009). Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature, 457*(7225), 79-82.

Perugini, M., & Leone, L. (2009). Implicit self-concept and moral action. *Journal of Research in Personality, 43*(5), 747-754.

Raihani, N. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in ecology & evolution*.

Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded–but third-party helpers even more so. *Evolution*.

Rand, D., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*(7416), 427-430.

Rand, D. G. (2016). Cooperation, Fast and Slow Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation. *Psychological Science*, 0956797616654455.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., et al. (2014). Social heuristics shape intuitive cooperation. *Nature communications, 5*.

Rand, D. G., Tomlin, D., Bear, A., Ludvig, E. A., & Cohen, J. D. (2017). Cyclical population dynamics of automatic versus controlled processing: An evolutionary pendulum. *Psychological review, 124*(5), 626.

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences, 109*(37), 14824-14829.

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: the factor structure and item properties of the original and brief fear of negative evaluation scale. *Psychological assessment, 16*(2), 169.

Roseman, I. J. (2011). Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. *Emotion Review, 3*(4), 434-443.

Roseman, I. J., Wiest, C., & Swartz, T. S. (1994). Phenomenology, behaviors, and goals differentiate discrete emotions. *Journal of personality and social psychology, 67*(2), 206.

Ruttan, R. L., McDonnell, M.-H., & Nordgren, L. F. (2015). Having "been there" doesn't mean I care: When prior experience reduces compassion for emotional distress. *Journal of personality and social psychology, 108*(4), 610.

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological science, 20*(4), 523-528.

Salerno, J. M., & Peter-Hagene, L. C. (2013). The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science, 24*(10), 2069-2078.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., et al. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience, 40*, 99-124.

Skitka, L. J., Bauman, C. W., & Mullen, E. (2004). Political tolerance and coming to psychological closure following the September 11, 2001, terrorist attacks: An integrative approach. *Personality and Social Psychology Bulletin, 30*(6), 743-756.

Stagnaro, M., Pennycook, G., & Rand, D. G. (2018). Cognitive reflection is a stable trait. *Available at SSRN*.

Stanovich, K. E. (2005). *The robot's rebellion: Finding meaning in the age of Darwin*: University of Chicago Press.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annu. Rev. Psychol., 58*, 345-372.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of personality and social psychology, 78*(5), 853.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision making, 4*(6), 479.

Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of consulting and clinical psychology, 33*(4), 448.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and social Psychology, 51*(1), 110.

Young, L., Chakroff, A., & Tom, J. (2012). Doing good leads to more good: The reinforcing power of a moral self-concept. *Review of Philosophy and Psychology, 3*(3), 325-334.

Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of theoretical Biology, 53*(1), 205-214.

Zefferman, M. R. (2014). Direct reciprocity under uncertainty does not explain one-shot cooperation, but demonstrates the benefits of a norm psychology. *Evolution and Human Behavior, 35*(5), 358-367.

Zimmermann, J., & Efferson, C. (2017). One-shot reciprocity under error management is unbiased and fragile. *Evolution and Human Behavior, 38*(1), 39-47.