Journal of Experimental Social Psychology 70 (2017) 198–203

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Case Report

Moral traps: When self-serving attributions backfire in prosocial behavior

Stephanie C. Lin *, Julian J. Zlatev, Dale T. Miller

Stanford Graduate School of Business, Knight Management Center, 655 Knight Way, Stanford, CA 94305, USA

HIGHLIGHTS

- People make self-serving attributions when responding to prosocial requests.
- But, morally self-serving attributions can backfire when challenged.
- People are forced into prosocial action if their excuse for refusal is removed.
- This also occurs if an external incentive (e.g., payment) is removed.

ARTICLE INFO

Article history:
Received 18 August 2016
Revised 2 November 2016
Accepted 3 November 2016
Available online 1 December 2016

Keywords:
Morality
Attributions
Decision making
Prosocial behavior

ABSTRACT

Two assumptions guide the current research. First, people's desire to see themselves as moral disposes them to make attributions that enhance or protect their moral self-image: When approached with a prosocial request, people are inclined to attribute their own noncompliance to external factors, while attributing their own compliance to internal factors. Second, these attributions can backfire when put to a material test. Studies 1 and 2 demonstrate that people who attribute their refusal of a prosocial request to an external factor (e.g., having an appointment), but then have that excuse removed, are more likely to engage in prosocial behavior than those who were never given an excuse to begin with. Study 3 shows that people view it as more morally reprehensible to no longer honor the acceptance of a prosocial request if an accompanying external incentive is removed than to refuse a request unaccompanied by an external incentive. Study 4 extends this finding and suggests that people who attribute the decision to behave prosocially to an internal factor despite the presence of an external incentive are more likely to continue to behave prosocially once the external incentive is removed than are those for whom no external incentive was ever offered. This research contributes to an understanding of the dynamics underlying the perpetuation of moral self-regard and suggests interventions to increase prosocial behavior.

The pervasive desire for people to see themselves as good and virtuous (e.g., Aquino & Reed, 2002; Blasi, 2004; Monin & Jordan, 2009) drives them to make self-serving attributions—blaming self-interested behavior on outside circumstances (Shalvi, Gino, Barkan, & Ayal, 2015) and taking credit for particularly moral behaviors. An important but unanswered question, however, is what happens when circumstances put those attributions to a material test. For example, suppose John justifies his decision not to donate to a particular charity with the claim that the money will go to overhead. How might he react when he finds out that the organization actually has low overhead costs? Will he simply shift to another comforting external attribution (e.g., the cause is not worthy) or will he feel trapped by his original attribution and now donate? Or, consider Jane, who has a self-interested reason to engage in a prosocial task (e.g., being paid), but insists that she would have volunteered regardless of external incentives. Now imagine circumstances were to change such that the self-interested benefit was removed (e.g., no longer being paid). Will she decide not to volunteer after all, or will she act consistently with her original claim that she would volunteer even if she was not paid?

John and Jane's situations above raise the question of how flexible the capacity for psychological self-protection is. People are motivated to make attributions regarding the reasons behind behavior (Bem, 1972; Kelley, 1973). In the case of their own behavior, people are motivated to, and amazingly resourceful at, making attributions that maximize self-regard (Kunda, 1990; Miller & Ross, 1975). Still, people's capacity to avoid unpleasant self-relevant conclusions is not infinite (Kunda, 1990). Both their denials and boasts must be plausible. Consequently, we argue that people have to "walk the talk" of their attributions, even if doing so is costly. In the context of prosocial requests, we propose that self-serving attributions, both for compliance and

* Corresponding author.
E-mail addresses: sclin1@stanford.edu (S.C. Lin), jjzlatev@stanford.edu (J.J. Zlatev), dtmiller@stanford.edu (D.T. Miller).

http://dx.doi.org/10.1016/j.jesp.2016.11.004
0022-1031/

noncompliance, can ironically backfire; people who confront altered circumstances that challenge their earlier attributions feel compelled to act in accordance with them.

In four studies, we test our prediction that self-serving attributions can backfire when challenged. In part I, we hypothesize that, when attributing their refusal of prosocial requests to external factors (e.g., "I have an appointment at the time"), people will be psychologically forced into prosocial action when those external factors are removed (e.g., the appointment is canceled). In part II, we hypothesize that people over-attribute their compliance with prosocial requests to intrinsic motivation and under-attribute it to external incentives and will be similarly psychologically forced into prosocial action when those incentives are removed.

1. Part I: backfiring of self-serving attributions during prosocial request refusal

Because people desire to maintain a moral self-image, they often feel uncomfortable engaging in self-other tradeoffs (Lin, Schaumberg, & Reich, 2016). Although individuals might prefer to take a self-interested action, they tend to want to do so without incurring moral self-reproach (Berman & Small, 2012). For this reason, people have developed justifications that allow themselves to engage in self-interested behavior while avoiding attributing that behavior to their own moral character (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996; Mazar, Amir, & Ariely, 2008; Pittarello, Leib, Gordon-Hecker, & Shalvi, 2015; Shalvi et al., 2015). For instance, people can take advantage of ambiguous moral wiggle room (Dana, Weber, & Kuang, 2007; Shalvi, Dana, Handgraaf, & De Dreu, 2011) or atone for their wrongdoings by engaging in prosocial or physically cleansing behaviors (Jordan, Mullen, & Murnighan, 2011; Shalvi et al., 2015; Zhong & Liljenquist, 2006). One more straightforward justification strategy relevant to the current research is attributing self-interested behavior to external forces (Snyder, Kleck, Strenta, & Mentzer, 1979).

The present research shows that offering external excuses for self-interested behavior is not cost-free and can backfire if circumstances remove the excuse. We argue that claims about why one did what one did in a particular situation can sow the seeds of a moral test (Miller & Monin, 2016). Consider the decision not to perform a particular prosocial behavior. The potential moral reproach from this decision can be mitigated by invoking a specific excuse (e.g., having an appointment at the time). Implicit in this excuse, however, is a moral claim (e.g., that one *would have* volunteered if possible) that heightens moral self-regard in the moment but sets a "trap" for the person if it is subsequently removed. To continue to refuse a prosocial request when the reason one initially gave for refusing it no longer applies (e.g., the appointment is canceled) would reveal not only that one has failed to live up to this heightened moral standard, but also that one is a hypocrite, which is a negative moral signal both to oneself and to others (Andrade & Ariely, 2009; Batson, Thompson, Seuferling, Whitney, & Strongman, 1999; Kreps & Monin, 2011). Accordingly, people cannot simply replace the former excuse with a new one, as it would undermine their earlier claim. Thus, we predict that when a person's excuse for not taking a prosocial action is removed, he or she will feel increased pressure to take that action.

We conducted Studies 1 and 2 to provide evidence that people who are given an excuse to refuse a prosocial request would be more likely to comply with the request when the excuse was removed than those who were never offered an excuse to begin with. We tested this using a scenario study (Study 1) followed by a study with a behavioral outcome (Study 2).

1.1. Study 1

Study 1 tested whether people who choose a self-interested action over a prosocial one when they have an external excuse for doing so would be more likely engage in the prosocial action when their original excuse was removed than those who never had such an excuse. We also sought to determine whether enhanced moral self-regard mediated this effect.

1.1.1. Methods

1.1.1.1. Participants. Given uncertainty about effect size, we adhered to the suggestion of using at least 50 participants per condition (Simmons, Nelson, & Simonsohn, 2013) and recruited 100 participants per condition, resulting in 200 participants from Amazon Mechanical Turk ($0.50 payment, $M_{age} = 32.12$, $SD_{age} = 10.37$, 114 men, 86 women). Data for this and subsequent studies were not analyzed until data collection was completed by Amazon Mechanical Turk (sometimes resulting in slightly larger samples than originally intended). We exclude no data for any study and report all the manipulations and measures used.

1.1.1.2. Procedure. Participants read a scenario that described how, in the spirit of the holidays, their town was offering food service at a local shelter, and that they were conflicted about whether to volunteer—they wanted to help the community, but it would also be a big hassle. In one condition, participants read that they realized they had an appointment at the same time. All participants then wrote their thoughts and feelings in free-response format. They then rated themselves on the nine traits (1 = *not at all characteristic of me*, 5 = *extremely characteristic of me*; $\alpha = 0.89$) previously shown to be viewed as indicative of moral character (Aquino & Reed, 2002), and filler items (creative, powerful, funny, athletic, disorganized, shy, neurotic) included to reduce demand effects. Afterwards, those who originally had an appointment were told that it was moved to a different day.

Everyone was then asked how obligated they felt to volunteer, how guilty they would feel if they did not volunteer, how good they would feel about themselves if they volunteered, how enjoyable volunteering would be (on 5-point scales, 1 = *not at all*, 5 = *extremely*), and how likely they would be to volunteer (0–100% slider). Finally, they were asked why they made the decision that they did in free-response format, and indicated gender and age.

1.1.2. Results

When people had an excuse not to volunteer, they rated themselves as higher on moral characteristics than when they did not have an excuse $t(198) = 3.87$, $p < 0.001$, see Table 1 for full statistics. In addition, of the filler variables, people reported feeling marginally more powerful and funny, and significantly more creative in the excuse condition than the control condition, and thus we control for these variables in mediation analyses below.

After the excuse was removed, those in the excuse condition expressed more obligation to volunteer than those in the no excuse condition, $t(198) = 6.48$, $p < 0.001$. Those in the excuse condition also reported, once it was removed, that they would feel more guilty if they did not volunteer, $t(198) = 6.48$, $p < 0.001$, and would find volunteering more enjoyable if they did volunteer, $t(198) = 3.39$, $p < 0.001$. They did not, however, report they would feel better about themselves if they did volunteer, $t(198) = 1.33$, $p = 0.185$. Finally, participants in the excuse condition indicated that they would be more likely to volunteer than those in the control condition, $t(198) = 5.48$, $p < 0.001$.

A serial mediation analysis (PROCESS Model 6, Hayes, 2013, see Fig. 1) revealed that those in the excuse condition rated themselves as more moral, leading to higher anticipated guilt if they did not volunteer, ultimately leading them to be more likely to volunteer 95% CI: [0.95, 4.44] (controlling for power, creativity and humor, 95% CI: [0.54, 3.70]).

1.1.3. Discussion

In Study 1, the availability of an external justification (excuse) for self-interested behavior (not volunteering) elevated participants'

Table 1
Main effects of excuse condition on dependent variables in Study 1.

Variable	Condition means (SD)		B (95% CI)	$t(198)$	p	Cohen's d
	Excuse	Control				
Moral character	3.92 (0.67)	3.57 (0.57)	0.34 [0.17, 0.52]	3.87	<0.001	0.55
Obligated	3.28 (1.15)	2.27 (1.00)	1.00 [0.70, 1.30]	6.48	<0.001	0.92
Guilty	3.27 (1.25)	2.49 (1.24)	0.77 [0.42, 1.12]	4.36	<0.001	0.62
Feel good	4.00 (0.93)	3.82 (0.93)	0.18 [−0.08, 0.44]	1.33	0.185	0.19
Enjoyable	5.23 (1.41)	4.53 (1.52)	0.70 [0.29, 1.11]	3.39	<0.001	0.48
Likelihood of volunteering	67.12 (28.86)	44.52 (29.24)	22.60 [14.47, 30.73]	5.48	<0.001	0.78

moral self-regard and allowed those who did not intend to volunteer to construct the fiction that they were the type of person who would have volunteered if such a justification did not exist. As a consequence, those who would not have originally volunteered indicated they would volunteer when the justification was removed, presumably because of the guilt they anticipated feeling due to compromising their now-heightened self-regard and revealing their hypocrisy.

1.2. Study 2

In Study 2, we sought to conceptually replicate Study 1 with a real behavioral outcome. Specifically, we predicted that those who had a justification for not acting prosocially (donating to charity) that was subsequently removed would be more likely to donate to charity than those who originally had no such justification.

1.2.1. Methods

1.2.1.1. Participants. We aimed to again recruit 200 workers from Amazon Mechanical Turk, resulting in 203 workers from Amazon Mechanical Turk (113 men, 88 women, 2 declined to report gender) who participated in this study (M_{age} = 32.88, SD_{age} = 10.61, $0.50 payment).

1.2.1.2. Procedure. Participants were told that they would be deciding whether to allocate an extra $1 to the American Cancer Society or to themselves. Furthermore, they were told that there would be random probabilities assigned to each choice representing the likelihood that the money would actually be allocated to that choice. For instance, the choice could be between a 40% chance that the $1 would go to ACS and a 30% chance that the $1 would go to themselves. Participants viewed an example of this choice frame (30% and 40% were randomly assigned to the two choices), and responded to comprehension checks. Once they answered these questions correctly, they moved on to read about the actual choice.

Participants read that if they chose to allocate the $1 to ACS there was, depending on the condition, 70% or 100% chance that the money would go to ACS. Previous research using this paradigm showed that people use uncertainty about the prosocial choice as an excuse to make the selfish choice (Exley, 2015). Accordingly, we assumed that those for whom the choice to donate had a 70% chance of resulting in donation would feel more justified in not making this choice than those for whom the chance was 100%. All participants read that if they allocated the $1 to themselves, there was a 100% chance that they would receive the money. Participants then made their Time 1 (T1) choice.

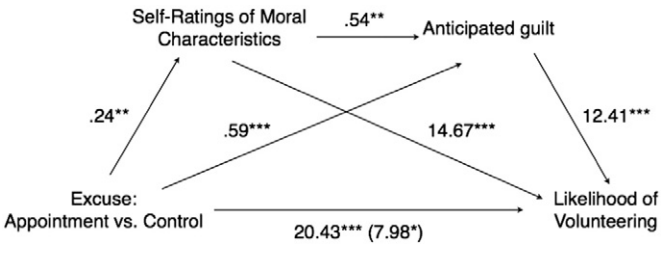

Fig. 1. Serial mediation analysis from Study 1. Values in figure are unstandardized regression coefficients controlling for power, humor, and creativity. Value in parentheses indicates remaining direct effect when controlling for self-ratings of moral characteristics and anticipated guilt. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

If participants indicated at T1 that they would allocate the money to themselves, they were asked at what probability, if any, they would change their minds and allocate the money to ACS. If they indicated at T1 that they would allocate the money to ACS, they were asked at what probability, if any, they would change their minds and allocate the money to themselves. Next, all participants were told that, for protocol reasons, the choice was actually between 100% to ACS and 100% to themselves (see Fig. 2). They were then asked to make a choice between those options (T2 choice) using both a binary and a continuous measure (1 = *strongly prefer to ACS*, 7 = *strongly prefer to myself*). They were then asked how unappealing or appealing it was to give to ACS and themselves given the 100% ACS vs. 100% self-choice (1 = *extremely unappealing*, 7 = *extremely appealing*).[1] Lastly, they indicated any comments or points of confusion they had about the study, their gender, and their age. Following the study, participants who chose to receive a bonus were paid, and donations were made on behalf of those who chose to donate.

1.2.2. Results

At T1, few participants opted to donate to charity whether they were told there was a 70% or 100% probability that the money would go to charity (6.00% vs. 8.74%). A binomial logistic regression revealed this difference was not significant, $B = -0.41$, SE = 0.55, $p = 0.46$, odds ratio = 0.67. The main dependent variable was likelihood of donating to ACS at T2, when all participants were told there was a 100% chance that the money would go to ACS. We found that those who originally had an excuse (70% condition) were over three times more likely to donate (26.00% [17.4, 34.6]) than those who originally had no excuse (7.84% [2.62, 13.06]), $B = 1.42$ [0.61, 2.33], SE = 0.43, $p = 0.001$, odds ratio = 4.13, see Fig. 3. See supplemental materials for analyses of additional variables.

1.2.3. Discussion

In Study 2, although all participants ultimately faced the same choice between prosocial and self-interested behavior, those who were given an excuse for not donating were more likely to donate when the excuse was removed than those who were not originally given the excuse. Participants who had a justification for choosing not to donate—the uncertainty that the money would actually be donated—were more likely to donate when they later discovered the donation would certainly go to charity than were those who assumed that the donation would certainly go to charity all along. Notably, participants in this study took advantage of a relatively flimsy excuse, as there was a high probability (70%) that the money would be donated; previous research has demonstrated

[1] Results of these final three variables are summarized in the supplemental materials.

	Excuse		No excuse	
Time 1 choice	To ACS	To You	To ACS	To You
	70%	100%	100%	100%
Time 2 choice	To ACS	To You	To ACS	To You
	100%	100%	100%	100%

Fig. 2. Experimental design of Study 2: choices made at T1 and T2 for the excuse and no excuse conditions.

that people will use uncertainty as an excuse even at much higher probabilities (e.g., 95%; Exley, 2015).

2. Part II: backfiring of self-serving attributions during prosocial request compliance

Because prosocial behavior prototypically involves putting others' interests above one's own, engaging in prosocial behavior for a self-interested gain, otherwise known as "tainted altruism," is often viewed as less morally pure than engaging in prosocial behavior free from self-interested motives (Newman & Cain, 2014; Zlatev & Miller, 2016). One can attempt to mitigate how tainted the act seems by asserting that one was internally motivated (Barasch, Levine, Berman, & Small, 2014), but doing so will psychologically constrain one's future behavior if circumstances change. Consider the earlier example of Jane, who was offered an incentive to engage in a prosocial behavior that was subsequently removed. However less appealing the removal of the incentive makes the prosocial action, failure to take the action would undermine her moral status even more so than an initial refusal would have done. As the psychological stakes of not complying have increased, so too, we predict, does the likelihood of compliance.

Thus, whereas Studies 1 and 2 examined the unintended consequences of over-attributing refusals of prosocial requests to external justifications, Studies 3 and 4 examined the unintended consequences of under-attributing compliance with prosocial requests to incentives. We hypothesized that people would find it to be morally compromising to first accept a prosocial request when it is accompanied by an external incentive, and then renege on this decision if the external incentive was removed. Such a refusal, we contend, would give the lie to their original claim that their initial compliance reflected intrinsic not extrinsic motivation (Study 3). Because of people's reluctance to contradict an earlier moral claim, we hypothesized that people would be more likely to comply with a prosocial request when it was initially accompanied by an external incentive that was then removed than with one that never included an external incentive (Study 4).

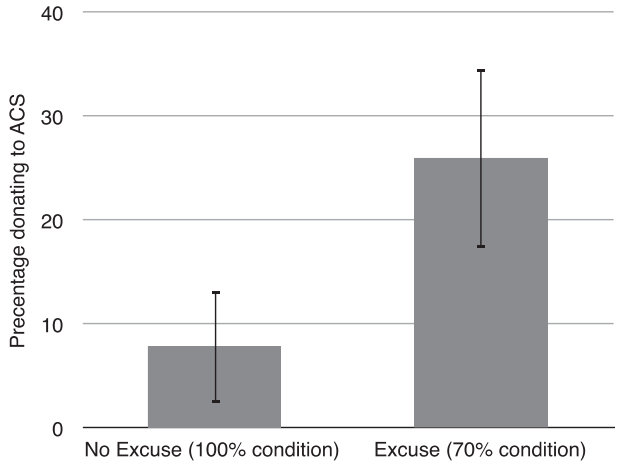

Fig. 3. Percentage of donation by excuse condition. Error bars are 95% confidence intervals.

2.1. Study 3

Study 3 sought to establish that people who first are presented with a request for help accompanied by the expectation of compensation are judged morally more harshly if they refuse the request when the offer of compensation is revoked than those who refuse a prosocial request for which no external incentive was ever offered.

2.1.1. Methods

2.1.1.1. Participants. We again aimed to recruit 200 Amazon Mechanical Turk workers, resulting in 210 Amazon Mechanical Turk workers (122 men, 82 women, two identified as other, four declined to report gender) who participated in this study (M_{age} = 32.15, SD_{age} = 9.95, $0.25 payment).

2.1.1.2. Procedure. Participants were assigned to either the paid or non-paid request condition. In the paid condition, participants read, "Imagine that John was asked whether he would be willing to help serve food at a local homeless shelter. He would be paid $10/hour for helping. He is not sure whether he has time to help. He is leaning towards helping out, but he hasn't decided yet." In the non-paid condition, participants read the same statement, only they were told John would not be paid for his work. On the next page, participants in the paid condition read, "The next day, he finds that, although he thought he would be paid to help, he was wrong, and the job is volunteer-only. He decides he is too busy to volunteer after all." In the non-paid condition, participants simply read, "The next day, he decides he is too busy to volunteer after all."

Participants were then asked to think about what kind of person John is, and rated John on the moral traits used in Study 1 (1 = *not at all characteristic of John*, 5 = *extremely characteristic of John*; α = 0.94), along with the same filler items. They were also asked why they judged him the way they did, in free-response format. Lastly, they reported gender and age.

2.1.2. Results

As expected, John was viewed as less moral when he refused the request to volunteer after initially believing he would be paid to volunteer (M = 2.33, SD = 0.79) than if he never believed he would have been paid (M = 3.09, SD = 0.88), B = −0.76 [−0.98, −0.53], $t(206)$ = 6.51, $p < 0.001$, d = 0.90 (see Fig. 4). There were no differences by condition for any other traits except shyness (M_{paid} = 2.11, SD_{paid} = 0.95; $M_{non\text{-}paid}$ = 2.39, $SD_{non\text{-}paid}$ = 1.05; $t(203)$ = 1.94, p = 0.054, d = 0.27).

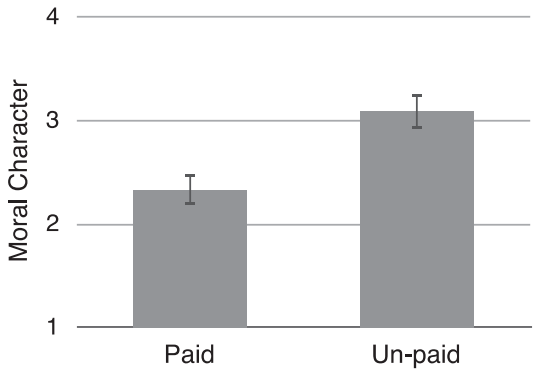

Fig. 4. Judgment of moral character by incentive condition in Study 3. Error bars are 95% confidence intervals.

2.1.3. Discussion

Study 3 supported our hypothesis that people are judged harshly if it is assumed that they would only engage in a prosocial action if they were materially compensated for doing so. That is, a target was viewed as less moral when he refused a prosocial request after payment was removed than when he simply refused the request with no mention of an external incentive. For instance, one participant in the paid condition stated, "I think John is greedy and also a hypocrite towards other people and himself, hiding his lack of empathy behind fake motivations (little time to help the poor)." We hypothesized that the belief that complying with a prosocial request when accompanied by external incentives but refusing to do so without them is morally reprehensible should have behavioral consequences similar to those found in Study 2. Specifically, we predicted that people would be more willing to comply with a prosocial request when the extrinsic incentive accompanying it was removed than when no extrinsic incentive was ever offered.

2.2. Study 4

Study 4 examined whether people's desire to downplay the role that an accompanying extrinsic incentive played in their decision to comply with a prosocial request could lay a future moral trap for them. This trap would be sprung in the event that the originally offered incentive was removed. They would then be trapped into complying because to refuse to do so would reflect negatively on their moral character, more so even than if they had refused the request in the first instance, as shown in Study 3. Specifically, we predicted that people who initially were paid to engage in a prosocial behavior would be more likely to engage in it after the payment was removed than those who had never been offered payment.

2.2.1. Methods

2.2.1.1. Participants. We anticipated a smaller effect size for this study than the previous studies and thus aimed to recruit 300 workers from Amazon Mechanical Turk, resulting in 310 Amazon Mechanical Turk workers (127 men, 137 women, two identified as other) who participated in this study ($M_{age} = 32.10$, $SD_{age} = 10.67$).

2.2.1.2. Procedure. Participants first completed a 2-min unrelated study regarding exercise habits in exchange for $0.25 (at the end of which gender and age were collected). After the initial study, they learned of the following prosocial opportunity: "We are collaborating with a non-profit organization whose goal is to motivate at-risk youth to stay in school. This organization is collecting essays for the students. Anyone is qualified to write these essays! If you agree to write the essay, there will be more instructions, including essay prompts." At T1, in the payment removal condition, they were asked whether they would be willing to write a short essay for this organization in exchange for a $0.75 bonus payment. In the control condition, they were asked if they would be willing to write a short essay for this organization, but told that they would not be given a bonus payment.

Following this, participants were asked how helpful they believed it would be to write the essay for the non-profit, how much good writing the essay would do, how motivated they were to help at-risk youth, and how important it was to them to help the non-profit (1 = *not at all*, 5 = *extremely*), all of which were averaged to form one composite of motivation to help ($\alpha = 0.83$).

Finally, at T2, all participants were told that, for protocol reasons, they would not be offered bonus payment for helping the organization. They then once again indicated whether they would be willing to write an essay for the organization. Those who chose to write the essay completed the task as described, for the sake of consistency.

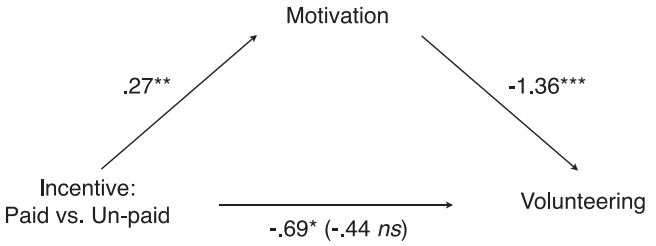

Fig. 5. Mediation analysis from Study 4. Values in figure are logistic regression coefficients. Value in parentheses indicates remaining direct effect when controlling for motivation. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

2.2.2. Results

As expected, a binomial logistic regression revealed that, at T1, participants in the payment removal condition (36.71%) were more likely to help than those in the control condition (11.84%), $B = -1.46$ [−2.07, −0.89], SE = 0.30, $p < 0.001$, odds ratio = 0.23. Furthermore, participants in the payment removal condition indicated being more intrinsically motivated to help ($M = 3.04$, SD = 0.91) than those in the control condition ($M = 2.77$, SD = 0.85), $t(308) = 2.69$, $p = 0.008$, $d = 0.31$.

The main dependent variable was the actual completion of the prosocial task at T2. A binomial logistic regression revealed that participants in the payment removal condition were more likely to agree to complete the task (27.22% [20.28, 34.16]) than those in the control condition (15.79% [9.99, 21.59]), $B = -0.69$ [−1.26, −0.138], SE = 0.29, $p = 0.016$, odds ratio = 0.50. The effect of payment removal on the willingness to help at T2 was mediated by their attribution to internal motivations, 95% CI: [−0.089, −0.011], see Fig. 5.

2.2.3. Discussion

In Study 4, participants who were offered an external incentive to engage in a prosocial behavior that was later removed were more likely to engage in that prosocial behavior than those who were never promised an external reward. Furthermore, mediation analysis suggested those with removed incentives were more likely to comply because they attributed their original compliance to intrinsic motivation, rather than the external reward. The results support our reasoning that the higher volunteer rate in the former case was due to their desire to protect the original self-serving attributions they made for helping.

3. General discussion

It is well known that people are skilled at generating attributions that enhance and protect their moral self-image. We show, however, that these self-serving attributions carry risk. They can constrain people's choices when future circumstances put those attributions to the test. A person who wants to claim that his refusal of a prosocial request was due to external constraints is trapped when those constraints no longer obtain. He cannot simply shift excuses but rather feels compelled to honor his original claim. To not do so imperils his moral self- and public-image[2] more than does refusing in the first place. Similarly trapped is the person who claims that her compliance with a request for help was not due to the presence of an extrinsic incentive. Her claim that her prosocial act reflects virtuousness sets up a bigger moral test (Miller & Monin, 2016) when that incentive is no longer available. Specifically, to opt for the self-interested action in these

[2] While the effect may be stronger in public contexts, when one faces judgment from others, we contend that the process is internalized, as participants made anonymous choices in private. However, the line between self and other judgment is blurred, as one's own internal moral standards tend to be the internalized standards of close others or society. Thus, concerns about judgment from the self and from others may be difficult to fully disentangle.

circumstances would endanger her moral image by invalidating her previous claim.

Our findings contribute to the literature examining the strategies people use to reduce cognitive dissonance (Festinger, 1962) that stems from inconsistencies between moral self-image and self-interested behavior. That is, people use justifications to protect and enhance their moral self-image (Bandura et al., 1996; Mazar et al., 2008; Monin & Jordan, 2009; Shalvi et al., 2015). We show that these justifications can be constrained, leaving people in a more morally precarious place than they started.

Our results and analysis also provide a potential alternative account for the well-known door-in-the-face phenomenon (Cialdini et al., 1975), whereby asking for an extreme favor before a smaller favor increases compliance with the smaller favor. This effect has previously been attributed to "reciprocal concessions," such that the person of whom the request is made feels obligated to accept the more reasonable request because he had denied the requester's previous request. However, our findings suggest that this effect may depend critically on people privately or publically generating excuses for their initial refusal (e.g., "it is unreasonable") that then constrains them when a more reasonable request is made.

The findings in Part II also suggest limitations on the crowding-out and overjustification effects whereby external incentives overwhelm intrinsic motivation. That is, a monetary fine can overwhelm motivation to engage in prosocial behavior by changing the perception of the behavior to an economic exchange (Gneezy & Rustichini, 2000) and a monetary reward can lead people to attribute their behavior to extrinsic rather than intrinsic motivation (Lepper, Greene, & Nisbett, 1973). Our findings demonstrate that if an activity is incentivized positively (in contrast to negatively, Gneezy & Rustichini, 2000) and has prosocial benefits (in contrast to non-prosocial behaviors, Lepper et al., 1973), people will resist attributing their behavior to the presence of an external force and, consequently, will show *more* rather than *less* engagement in the activity when the external motivator is removed.

An open question is whether this effect would occur for more subtle or non-conscious justification tactics, such as moral licensing, whereby a good deed bolsters a sense of moral self-worth, allowing people to subsequently engage in morally questionable behavior (Merritt, Effron, & Monin, 2010). If such a good deed was negated (e.g., one thinks one hired a black job candidate (Monin & Miller, 2001) but the files had been switched such that the hired candidate was actually white), perhaps people would not only lose their moral license, but would feel morally indebted and subsequently perform more moral deeds.

Finally, the psychology underlying our results can be leveraged to increase prosocial behavior. For example, organizations might emphasize that ordinary excuses for not engaging in prosocial behavior (i.e., too much money used for overhead) are not legitimate. Organizations might also ask people to consider whether they would be willing to help for self-interested gain before asking whether they would like to volunteer to help others.

4. Conclusion

When people adopt cognitive strategies to elevate their moral self-view they restrict their future options if changing circumstances nullify those strategies. After strategies are nullified, people are not free to merely shift to another self-serving cognitive strategy. Instead, they must "walk the talk" of their original strategies or expose their moral image to more threat. Self-serving strategies, thus, are double-edged swords: they raise people's moral self-view but increase its distance to fall.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jesp.2016.11.004.

References

Andrade, E. B., & Ariely, D. (2009). The enduring impact of transient emotions on decision making. *Organizational Behavior and Human Decision Processes*, *109*, 1–8.

Aquino, K., & Reed, A., II (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*, 1423–1440.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, *71*, 364–374.

Barasch, A., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or selfless? On the signal value of emotion in altruistic behavior. *Journal of Personality and Social Psychology*, *107*, 393–413.

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, *77*, 525–537.

Bem, D. J. (1972). Self-perception theory. *Advances in Experimental Social Psychology*, *6*, 1–62.

Berman, J. Z., & Small, D. A. (2012). Self-interest without selfishness: The hedonic benefit of imposed self-interest. *Psychological Science*, *23*, 1193–1199.

Blasi, A. (2004). Moral functioning: Moral understanding and personality. In D. K. Lapsley, & D. Narvaez (Eds.), *Moral development, self, and identity*. Psychology Press.

Cialdini, R. B., Vincent, J. E., Lewis, S. K., Catalan, J., Wheeler, D., & Darby, B. L. (1975). Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique. *Journal of Personality and Social Psychology*, *31*, 206–215.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*, 67–80.

Exley, C. L. (2015). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, *83*, 587–628.

Festinger, L. (1962). *A theory of cognitive dissonance*. Stanford University Press.

Gneezy, U., & Rustichini, A. (2000). Fine is a price. *Journal of Legal Studies*, *29*, 1–17.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin*, *37*, 701–713.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*, 107–128.

Kreps, T., & Monin, B. (2011). "'Doing well by doing good'?" Ambivalent moral framing in organizations. *Research in Organizational Behavior*, *31*, 101–125.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the 'overjustification' hypothesis. *Journal of Personality and Social Psychology*, *28*, 129–137.

Lin, S. C., Schaumberg, R. L., & Reich, T. (2016). Sidestepping the rock and the hard place: The private avoidance of prosocial requests. *Journal of Experimental Social Psychology*, *64*, 35–40.

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

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*, 344–357.

Miller, D. T., & Monin, B. (2016). Moral opportunities versus moral tests. In J. Forgas, L. Jussim, & P. van Lange (Eds.), *The social psychology of morality*. New York, New York: Psychology Press.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*, 213–225.

Monin, B., & Jordan, A. H. (2009). The dynamic moral self: A social psychological perspective. In D. Narvaez, & D. K. Lapsley (Eds.), *Personality, identity, and character: Explorations in moral psychology* (pp. 341–354). Cambridge University Press.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*. http://dx.doi.org/10.1037//0022-3514.8l.l.33.

Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological Science*, *25*, 648–655.

Pittarello, A., Leib, M., Gordon-Hecker, T., & Shalvi, S. (2015). Justifications shape ethical blind spots. *Psychological Science*, *26*, 794–804.

Shalvi, S., Dana, J., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, *115*, 181–190.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science*, *24*, 125–130.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after p-hacking. *Advances in Consumer Research*, *41*.

Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology*, *37*, 2297–2306.

Zhong, C. -B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*, 1451–1452.

Zlatev, J. J., & Miller, D. T. (2016). Selfishly benevolent or benevolently selfish: When self-interest undermines versus promotes prosocial behavior. *Organizational Behavior and Human Decision Processes*, *137*, 112–122.