COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e13007
© 2021 Cognitive Science Society LLC
ISSN: 1551-6709 online
DOI: 10.1111/cogs.13007

Memory and Counterfactual Simulations for Past Wrongdoings Foster Moral Learning and Improvement

Matthew L. Stanley,[a] Roberto Cabeza,[a] Rachel Smallman,[b]
Felipe De Brigard[c]

[a]*Department of Psychology and Neuroscience, Center for Cognitive Neuroscience, Duke University*
[b]*Department of Psychological and Brain Sciences, Texas A&M University*
[c]*Department of Psychology and Neuroscience, Center for Cognitive Neuroscience, Department of Philosophy, Duke Institute for Brain Sciences, Duke University*

Received 18 February 2021; received in revised form 17 April 2021; accepted 28 May 2021

Abstract

In four studies, we investigated the role of remembering, reflecting on, and mutating personal past moral transgressions to learn from those moral mistakes and to form intentions for moral improvement. Participants reported having ruminated on their past wrongdoings, particularly their more severe transgressions, and they reported having frequently thought about morally better ways in which they could have acted instead (i.e., morally upward counterfactuals; Studies 1–3). The more that participants reported having mentally simulated morally better ways in which they could have acted, the stronger their intentions were to improve in the future (Studies 2 and 3). Implementing an experimental manipulation, we then found that making accessible a morally upward counterfactual after committing a moral transgression strengthened reported intentions for moral improvement—relative to resimulating the remembered event and considering morally worse ways in which they could have acted instead (Study 4). We discuss the implications of these results for competing theoretical views on the relationship between memory and morality and for functional theories of counterfactual thinking.

Keywords: Moral psychology; Counterfactual; Autobiographical memory; Directive function; Cheating

None of these data and ideas have been presented at a conference, at a meeting posted on a listserv, or shared on a website.

Correspondence should be sent to Matthew Stanley, Department of Psychology and Neuroscience, Center for Cognitive Neuroscience, Duke University, Durham, NC 27708-0743, USA. Email: matthew.stanley@duke.edu

1. Introduction

Particularly egregious wrongdoings have frequently made national and international headlines in recent years. Wells Fargo employees fraudulently opened customer accounts to meet sales quotas; numerous allegations of sexual harassment and discrimination were made against prominent figures at Fox News; wealthy parents paid for their children's SAT test answers to be corrected (Operation Varsity Blues); employees at European Union banks were accused of laundering billions of dollars for kleptocrats; and Volkswagen employees developed and installed software in vehicles so that they could falsely pass emissions tests. Serious moral transgressions can have significant and deleterious impacts on individuals, organizations, institutions, and society at-large. As such, it is critical to understand the underlying psychological factors that influence people's capacity to learn from their more serious moral mistakes in service of fostering moral improvement over time. In contrast to recent work suggesting that people forget their past transgressions (Kouchaki & Gino, 2016, Shu, Gino, & Bazerman, 2011), we offer positive evidence that people remember and frequently ruminate upon their more severe moral transgressions. We then investigate whether the ways in which those transgressions are remembered and mutated play a role in learning from past mistakes to form intentions for future moral improvement.

1.1. Two perspectives on memory for moral transgressions

Recent psychological research might suggest that memories of our past moral transgressions cannot be used to help us learn from our mistakes and improve over time, because our past moral transgressions tend to be forgotten (i.e., an "unethical amnesia" effect; Kouchaki & Gino, 2016). On this view, people intentionally forget their past moral transgressions to avoid the experience of psychological distress and to eliminate evidence that could damage their otherwise favorable self-concepts (Kouchaki & Gino, 2016; Shu et al., 2011). Supporting this view, Reczek, Irwin, Zane, and Ehrich (2017) found that consumers exhibit "willfully ignorant memory" for items produced in an unethical way (e.g., produced in an overseas factory using child labor). Forgetting unethical product information is thought to alleviate negative affect and distress that the consumer might have otherwise experienced when purchasing the product. In complementary research, after reading an honor code meant to bring awareness to honesty standards, participants who then cheated on a task to earn more money "strategically forgot" certain content of the honor code (Shu et al., 2011). Relatedly, participants tended to remember their own moral transgressions involving cheating and dishonesty in a less vivid and less detailed way compared to other kinds of past events (Kouchaki & Gino, 2016; but see Stanley, Yang, & De Brigard, 2018). Ultimately, as Kouchaki and Gino (2016) suggest, people might frequently commit moral transgressions because they tend to be forgotten. If our past moral transgressions are forgotten, then it is unclear how those forgotten events could serve as explicit reference points for moral learning and improvement. This is especially problematic if our more severe moral transgressions are more likely to be forgotten, given that those particular memories are presumably more negative, distressing, and threatening to our favorable self-concepts.

In contrast, other research suggests that people can and do readily recall many of their own past moral transgressions instead of forgetting them (Huang, Stanley, & De Brigard, 2020; Stanley & De Brigard, 2019; Stanley, Bedrov, Cabeza, & De Brigard, 2020). Several recent studies have shown that, when prompted to do so, people can readily recall having committed many different moral transgressions from their recent and distant pasts—some of which they even judge to be extremely morally wrong (Escobedo & Adolphs, 2010; Hofmann, Wisneski, Brandt, & Skitka, 2014; Stanley, Henne, Iyengar, Sinnott-Armstrong, & De Brigard, 2017, 2019, 2020). The process of retrieving these moral transgressions is commonly accompanied by intense negative affect and distress (Escobedo & Adolphs, 2010; Huang et al., 2020; Stanley et al., 2017). Critically, Huang et al. (2020) found not only that people can readily recall their own moral transgressions they judge to be extremely morally wrong, but also that more severe moral transgressions are reportedly recalled more frequently (both voluntarily and involuntarily), with more detail, and with a stronger sense of reliving. These findings suggest that people might readily and frequently recall many of their moral transgressions (especially those judged to be more severe)—even though remembering them tends to be negative, distressing, and threatening to their favorable self-concepts.

In fact, more severe moral transgressions tend to be readily remembered and frequently retrieved, then those memories could serve as reference points for moral learning and improvement (i.e., they might serve a *directive function*). Outside of the moral domain, research has shown that memories of personal past events can serve a directive function (Pillemer, 2003). Pillemer (1998, 2001) and others (e.g., Pratt, Arnold, & Mackey, 2001) have identified different directive functions for remembered personal past events; these memories tend to be accessible in memory and frequently come to mind, providing guidance when encountering similar situations in the future. For example, Pillemer (1998) analyzed an individual's memory of being kidnapped at gunpoint. This memory was frequently recalled, and it guided that individual's beliefs about what kinds of situations should be avoided in the future. In theory, then, the frequent retrieval of our moral transgressions might serve as a reminder of how *not* to act in the future, should similar circumstances arise. The strong negative emotions that accompany remembering our particularly severe moral transgressions might even serve as a "functional signal" to change our thoughts and behavior to avoid similar negative emotions in the future.

1.2. Episodic counterfactual thinking and intention formation

When we recall our personal past events, we often consider alternative ways such past events could have transpired instead—i.e., we engage in *episodic counterfactual thinking* (De Brigard & Giovanello, 2012; De Brigard & Parikh, 2019; Özbek, Bohn, & Berntsen, 2018; Schacter, Benoit, De Brigard, & Szpunar, 2015). In other words, thinking about our past experiences often entails mentally simulating "what if" or "only if" possibilities (Byrne, 2016, 2017; De Brigard & Parikh, 2019; Kahneman & Miller, 1986). Such episodic counterfactual thoughts might be essential for how memory best serves a directive function. Episodic counterfactual thinking is pervasive after people experience negative events (Byrne, 2016; Roese, 1997), particularly when those events elicit negative emotions like regret, disappointment, guilt, and shame (Niedenthal, Tangney, & Gavanski, 1994; Tangney, 1995; van Dijk & Zeelenberg, 2005; Zeelenberg et al., 1998). According to the *functional theory of counterfactual thinking*, counterfactual thoughts after negative, adverse events can be functional for reasoning and goal pursuit (Epstude & Roese, 2008; Roese, 1997; Roese & Epstude, 2017). In particular, thoughts about ways in which a negative event could have turned out *better* (i.e., upward counterfactuals) help us to learn from past mistakes and improve future outcomes (Morris & Moore, 2000; Nasco & Marsh, 1999; Rim & Summerville, 2014; Roese, 1994, 1997), even though generating these upward counterfactuals tends to evoke unpleasant feelings (Davis, Lehman, Wortman, Silver, & Thompson, 1995; Markman et al., 1993).

A now substantial body of research indicates that upward counterfactual thinking about past negative events can be functional and adaptive. By considering ways in which a negative outcome could have turned out better, upward counterfactuals can strengthen intentions to act in particular ways (Roese, 1994; Smallman, 2013; Smallman & Roese, 2009), increase motivation (Dyczewski & Markman, 2012; Markman, McMullen, & Elizaga, 2008), facilitate behavior regulation (Epstude & Roese, 2008; Markman & McMullen, 2003; Roese, 1994; Roese & Epstude, 2017), and improve future performance (Morris & Moore, 2000; Nasco & Marsh, 1999). For instance, when participants thought about how they could have performed better on an anagram task, they persisted for longer in trying to solve subsequent anagrams and performed better in solving those anagrams, compared to participants who instead thought about how their performance could have been worse (i.e., downward counterfactuals; Markman et al., 2008). Complementary research has found that students who generated upward counterfactuals after receiving an exam grade performed better on a subsequent exam (Nasco & Marsh, 1999; see also, Roese, 1994), and that aviation pilots learned from "near misses" by generating upward counterfactuals (Morris & Moore, 2000). If people do frequently consider morally better alternative ways of acting when they recall their past transgressions (i.e., morally upward counterfactuals), then remembering those events and frequently simulating relevant counterfactuals could serve as reminders of how to act and as guides for forming intentions and goals to behave differently in the future.

1.3. Overview of studies and hypotheses

Across four studies, we tested several specific hypotheses regarding whether, how, and why people remember their own moral transgressions. The primary purpose of Study 1 is to lay the groundwork for the possibility that remembering past moral transgressions can serve a directive function, particularly for those more severe moral transgressions. To this end, participants were instructed to describe memories of their own wrongdoings, and then they reported the moral wrongness of their transgressions, their emotions while remembering the events, the frequency with which they have retrieved these memories, and the frequency with which they have considered morally better ways in which they could have acted instead. We hypothesized that participants would successfully recall their past moral transgressions and report having ruminated on them since the event occurred. We also hypothesized that, for their more severe moral transgressions relative to less severe moral transgressions, participants would report experiencing stronger negative emotions when recalling those events, having recalled them more often, and having thought about them counterfactually with greater frequency. We expected these results to hold even though remembering our more severe moral transgressions is more psychologically distressing and more threatening to our favorable self concept.

In Study 2, we first attempted to conceptually replicate the findings from Study 1 using a different memory cueing procedure in which participants were instructed to recall specific kinds of moral transgressions involving dishonesty, harm, and unfairness. This cueing procedure was developed to ensure that the effects from Study 1 generalize across different kinds of moral transgressions. In addition, we measured participants' intentions to behave differently in the future after recalling each event. The purpose of measuring intentions was to test the following additional hypothesis: That people would tend to report having more frequently generated morally upward counterfactuals after committing more severe moral transgressions, which, in turn, would be associated with stronger intentions to behave differently and better in the future.

In Study 3, we investigated a potential moderator of the observed effects in the previous studies—punishment for wrongdoing. Theorists have recently argued that moral punishment can serve a "pedagogical function" (e.g., Cushman, 2015; Sarin, Ho, Martin, & Cushman, 2021). That is, moral punishment may offer a means to modify the future behaviors of possible social partners within a community. When punishment facilitates moral learning in a way that modifies community members' future behaviors, punishment can encourage cooperative and prosocial behaviors that are adaptively favorable (Boyd & Richerson, 1992; Cushman, 2015; Fehr & Gächter, 2002). If moral punishment serves this pedagogical function, then reflecting on our past moral transgressions that resulted in punishment (relative to no punishment) for transgressing might increase the reported frequency of recalling the event, increase the reported frequency of morally upward counterfactual thinking since the event occurred, and strengthen intentions for future moral improvement. Accordingly, we tested the hypotheses that reflecting on our past transgressions that were punished (relative to not punished) would increase the reported frequency of voluntary and involuntary recall, increase the reported frequency of morally upward counterfactual thinking, and strengthen intentions to behave differently and better in the future. We expected these results to hold even though remembering our punished transgressions may be more psychologically distressing and accompanied with more negative affect. In addition to addressing these possible moderation effects in Study 3, we also attempted to conceptually replicate the effects from Study 2.

Finally, in Study 4, we implemented an experimental manipulation to further investigate the function of recalling our past moral transgressions. We tested the hypothesis that making accessible a morally upward counterfactual when recalling a moral transgression strengthens intentions for moral improvement in the future. As comparison conditions, some participants resimulated the event as they remembered it, and other participants generated morally worse ways in which they could have acted instead (i.e., morally downward counterfactuals).

2. Study 1

The primary purpose of Study 1 is to test the hypothesis that, for their more severe moral transgressions relative to their less severe moral transgressions, participants would report experiencing stronger negative emotions when recalling those events, having recalled them more frequently, and having thought about them counterfactually with greater frequency.

2.1. Materials and method

2.1.1. Participants

One-hundred twenty individuals from the United States voluntarily participated in this study via Amazon's Mechanical Turk (AMT) for monetary compensation. Participant recruitment was restricted to individuals in the United States with a prior approval rating above 95%. Eighteen participants were excluded for failing to answer all questions about each memory, for providing clearly nonsensical responses to the memory cue (e.g., simply the word "GOOD" as the participant's entire response), for recalling an event that occurred more than 10 years ago, or for failing the attention check at the end (see below for details). As such, data were analyzed with the remaining 102 participants ($M_{\text{age}} = 35.18$ years, $SD = 10.67$, age range = [19, 70], 45 females, 57 males). The sample size was based on the sample sizes from Stanley et al. (2017, 2019), who used similar statistical techniques to address questions about remembered moral transgressions. In this first study and in all subsequent studies, we only analyzed the data after the sample size target was met. We report all the measures, manipulations, and exclusions in all studies. All studies reported herein were approved by the Duke Campus Institutional Review Board. Given the sensitive nature of participants' responses, data from all studies are available from the first author upon request and IRB approval.

2.1.2. Procedure

The study was self-paced. Participants were asked to recall a total of five distinct events, one at a time, from their personal pasts in which they did something they believed to be morally wrong. Participants were instructed that their remembered immoral behaviors could involve emotional harm, physical harm, unfairness, disloyalty, disrespect, cheating, or dishonesty. To encourage participants to remember specific autobiographical memories, participants were also instructed to only remember events that occurred on a particular day in a particular place.

For each memory, participants described the event in two to five sentences. They then typed in the month and year in which it occurred, and they selected one of the following options to best describe when it occurred: within the past day, within the past week, within the past 2 weeks, within the past month, within the past 2 months, within the past 6 months, within the past year, within the past 2 years, within the past 5 years, within the past 10 years. Then, participants answered the following question to assess the severity of the moral transgression: "how morally wrong was your behavior in this instance?" (1 = *slightly morally wrong*, 7 = *very morally wrong*). Two different questions about participants' emotional experience were then presented in a random order: "as you remember the event now, how positive or negative are your emotions?" (1 = *very negative*, 7 = *very positive*), and "as you remember the event now, how intense are your emotions?” (1 = *not at all intense*, 7 = *very intense*). Next, two questions about the frequency of having retrieved the events were presented in a random order: “since it happened, how often have you willfully thought about the event in your mind or talked about it?” (1 = *never*, 7 = *very often*), and “since it happened, has the memory of the event suddenly popped up in your thoughts by itself—that is, without your having attempted to remember it?” (1 = *never*, 7 = *very often*). The former question indexes the frequency of *voluntary* retrieval, while the latter indexes the frequency of *involuntary* retrieval (Berntsen, 2010; Johannessen & Berntsen, 2010; Marie Hall & Berntsen, 2008). Finally, participants were asked the following question about the frequency of morally upward counterfactual thinking for each remembered event: “since it happened, how often have you thought about or talked about morally better ways in which you could have acted?” (1 = *never*, 7 = *very often*).

After completing all ratings for all five memories, participants were asked the following: Do you feel that you paid attention, avoided distractions, and took the survey seriously? They responded by selecting one of the following: (1) no, I was distracted; (2) no, I had trouble paying attention; (3) no, I did not take the study seriously; (4) no, something else affected my participation negatively; or (5) yes. Participants were ensured that their responses would not affect their payment or their eligibility for future studies. Only those participants who selected “5” were included in the analyses (see exclusions above; for other studies employing similar attention checks, see Stanley, Marsh, & Kay, 2020; Stanley, Yin, & Sinnott-Armstrong, 2019). Participants then completed several demographics questions. Upon completion, participants were monetarily compensated for their time.

2.1.3. Data analyses

Data were analyzed using R (R Development Core Team, 2009) with the “lme4” software package (Bates, Maechler, Bolker, & Walker, 2015) and the “lmerTest” software package (Kuznetsova, Brockhoff, & Christensen, 2017). Data were fitted to linear mixed-effects models (LMEM), and subject was included as a random effect (random intercepts only in all models, as models that also included random slopes typically failed to converge). Significance for fixed effects was assessed using Satterthwaite approximations to degrees of freedom, and 95% confidence intervals around beta-values were computed using parametric bootstrapping (in our view, 95% CIs around beta-values offer the best available indication of effect size for LMEMs); see Boisgontier and Cheval (2016) for discussion of the movement toward mixed-effects modeling in the social and neural sciences. The alpha level for all statistical tests was set at .05.

Because emotions experienced while remembering past events and the judged morality of past events both differ as a function of when those events occurred in the past (Stanley et al., 2017), we ran additional models controlling for time in two complementary ways. One time variable (hereafter referred to as $time_A$) was coded as follows: 0 = *within the past day*; 1 = *within the past week*; 2 = *within the past 2 weeks*; 3 = *within the past month*; 4 = *within the past 2 months*; 5 = *within the past 6 months*; 6 = *within the past year*; 7 = *within the past 2 years*; 8 = *within the past 5 years*; 9 = *within the past 10 years*. Similar methods have been implemented to characterize the objective time that events occurred in the past (e.g., Escobedo & Adolphs, 2010; Stanley et al., 2017, 2019). The other time variable

Table 1
Means and standard deviations in Study 1

Variable	Mean	*SD*
Moral wrongness	4.59	1.75
Valence while remembering	2.80	1.29
Emotional intensity while remembering	3.41	1.77
Frequency of voluntary recall	3.08	3.27
Frequency of involuntary recall	3.27	1.73
Frequency of morally upward Counterfactual thinking	3.57	1.99

Note. $N = 102$. All variables measured on 7-point scales.

(hereafter referred to as $time_B$) indicates the number of months that have passed since the remembered event occurred, starting with remembered events that occurred in the same month as the experimental session coded as 0 (see Stanley, Henne, and De Brigard (2019) for a similar methodological approach).

2.2. Results

Table 1 depicts descriptive statistics for our variables of interest.

2.2.1. Emotional experience while remembering

We first tested the hypothesis that participants would report experiencing stronger negative emotions when recalling their more severe, relative to more minor, moral transgressions. A LMEM with the judged severity of the moral transgressions predicting emotional valence revealed a significant effect of the severity of the moral transgressions ($b = -.32$, $SE = 0.03$, $t = -10.17$, $p < .001$, 95% CI [−0.38, −0.26]) such that past transgressions judged to be more morally wrong tended to be experienced more negatively than transgressions judged to be less morally wrong. A second LMEM with the judged severity of the moral transgressions predicting emotional intensity revealed a significant effect of the severity of the moral transgressions ($b = .43$, $SE = 0.04$, $t = 11.13$, $p < .001$, 95% CI [0.36, 0.50]) such that past transgressions judged to be more morally wrong tended to be remembered with greater emotional intensity than transgressions judged to be less morally wrong.

Conceptually replicating effects obtained in Stanley et al. (2017), two additional LMEMs revealed that the judged severity of the remembered transgressions was significantly related to time_A ($b = .42$, $SE = 0.05$, $t = 8.39$, $p < .001$, 95% CI [0.32, 0.51]) and time_B ($b = 19.20$, $SE = 2.52$, $t = 7.63$, $p < .001$, 95% CI [14.28, 24.10]) in separate models; remembered behaviors judged to be more morally wrong occurred in the more distant past. Because of this, we sought to ensure that the effects of the judged severity of the moral transgression remained significantly related to valence and emotional intensity after statistically controlling for time_A and time_B. The effect of the severity of the moral transgression on valence did, in fact, remain statistically significant after controlling for time_A ($b = -.31$, $SE = 0.03$, $t = -9.31$, $p < .001$, 95% CI [−0.38, −0.24]) and time_B ($b = -.30$, $SE = 0.03$, $t = -9.10$, $p < .001$, 95% CI [−0.37, −0.24]) in separate models. And the effect of the severity of the moral transgression on emotional intensity did, in fact, remain significant after controlling for time$_{\mathrm{A}}$ (b = .45, SE = 0.04, t = 11.03, p < .001, 95% CI [0.37, 0.53]) and time$_{\mathrm{B}}$ (b = .44, SE = 0.04, t = 10.76, p < .001, 95% CI [0.35, 0.52]) in separate models.

2.2.2. *Frequency of voluntary and involuntary recall*

Next, we tested the hypothesis that participants would report having more frequently recalled their more severe, relative to more minor, moral transgressions. A LMEM with the judged severity of the moral transgressions predicting the frequency of voluntary recall revealed a significant effect of the severity of the moral transgressions (b = .25, SE = 0.04, t = 5.87, p < .001, 95% CI [0.16, 0.33]). Participants reported that past transgressions judged to be more morally wrong tended to be voluntarily recalled more frequently than transgressions judged to be less morally wrong. Because the effect of the judged severity of the moral transgressions on the frequency of voluntary recall could potentially be a by-product of when the events actually occurred in the past, we computed two additional LMEMs with the judged severity of the moral transgressions on the reported frequency of voluntary recall after statistically controlling for time$_{\mathrm{A}}$ and time$_{\mathrm{B}}$. This effect remained significant after statistically controlling for time$_{\mathrm{A}}$ (b = .28, SE = 0.04, t = 6.31, p < .001, 95% CI [0.19, 0.36]) and time$_{\mathrm{B}}$ (b = .27, SE = 0.04, t = 6.05, p < .001, 95% CI [0.17, 0.36]) in separate models. In addition, average ratings for voluntary recall were significantly above floor (p < .001; Table 1), indicating that participants voluntarily thought about their moral transgressions with some frequency.

Similarly, another LMEM with the judged severity of the moral transgressions predicting the reported frequency of involuntary recall revealed a significant effect of the severity of the moral transgressions (b = .31, SE = 0.04, t = 7.34, p < .001, 95% CI [0.22, 0.39]). As in the case of voluntary recall, participants reported that past transgressions judged to be more morally wrong tended to be involuntarily recalled more frequently than transgressions judged to be less morally wrong. Because the effect of the judged severity of the moral transgression on the frequency of involuntary recall could potentially have been driven by when the event actually occurred in the past, we computed two additional LMEMs with the judged severity of the moral transgression predicting the reported frequency of voluntary recall after statistically controlling for time$_{\mathrm{A}}$ and time$_{\mathrm{B}}$. This effect remained significant after statistically controlling for time$_{\mathrm{A}}$ (b = .35, SE = 0.04, t = 7.85, p < .001, 95% CI [0.26, 0.43]) and time$_{\mathrm{B}}$ (b = .35, SE = 0.04, t = 7.99, p < .001, 95% CI [0.26, 0.44]) in separate models. In addition, average ratings for involuntary recall were significantly above floor (p < .001; Table 1), indicating that participants involuntarily thought about their moral transgressions with some frequency.

2.2.3. *Frequency of morally upward counterfactual thinking*

Finally, we tested whether the judged severity of the moral transgression predicted the reported frequency of generating morally upward counterfactuals about the event. A LMEM with the judged severity predicting the moral transgression on the reported frequency morally upward counterfactual thinking since the event occurred revealed a significant effect of the severity of the moral transgression (b = .41, SE = 0.05, t = 8.42, p < .001, 95% CI [0.31, 0.50]). Participants reported that past transgressions judged to be more morally wrong tended to elicit more frequent counterfactual thoughts. More specifically, participants reported having generated morally upward counterfactuals more often after having committed more severe moral transgressions relative to less severe moral transgressions. Because the frequency of morally upward counterfactual thinking since the event occurred could potentially be the byproduct of when the event actually occurred in the past, we computed two additional LMEMs with the judged severity of the moral transgression predicting the reported frequency of morally upward counterfactual generation after statistically controlling for time$_{\mathrm{A}}$ and time$_{\mathrm{B}}$. This effect remained significant after controlling for time$_{\mathrm{A}}$ ($b = .47$, $SE = 0.05$, $t = 9.11$, $p <$ .001, 95% CI [0.36, 0.56]) and time$_{\mathrm{B}}$ ($b = .46$, $SE = 0.05$, $t = 8.94$, $p < .001$, 95% CI [0.36, 0.56]) in separate models.

2.3. Discussion

Overall, the results from Study 1 offer support for our initial hypotheses. Participants reported experiencing stronger negative emotions when recalling their more severe moral transgressions than their more minor transgressions; participants reported having more frequently retrieved—both voluntarily and involuntarily—their more severe moral transgressions relative to their more minor transgressions; and they reported having considered morally upward counterfactuals more frequently for their more severe moral transgressions than for their more minor transgressions. All these effects remained statistically significant even after controlling for when the events occurred in the past.

Our findings suggest that even though people report experiencing more extreme negative emotions when recalling their more severe moral transgressions, they nevertheless report having more frequently recalled and thought about them relative to their less severe moral transgressions. But to protect a favorable self-concept, we might expect our more serious moral transgressions to be recalled and ruminated upon very rarely, if at all. Why might people tend to frequently and repeatedly retrieve their particularly severe moral transgressions, if the act of retrieving them and ruminating upon them induces negative affect and threatens a morally good self-concept? Our suggestion is that these memories of particularly severe past transgressions might serve a different function. That is, memories of particularly severe moral transgressions might often serve a *directive* function. Providing some initial support for this explanation, participants reported that, when they recalled and ruminated upon their more severe moral transgressions, they frequently considered morally better ways in which they could have acted instead.

3. Study 2

In Study 2, we more directly investigated whether the function of frequently simulating morally upward counterfactuals might be directive. That is, we investigated whether frequently simulating better possible ways in which they could have acted after committing a moral transgression might serve to chart specific courses of action for the future and to form intentions to perform those alternative actions (if relevant circumstances were to arise in the future). In doing so, we also obtained more generalizable support for our hypotheses in Study 1 by investigating transgressions involving dishonesty, harm, and unfairness, respectively.

3.1. Materials and method

3.1.1. Participants

Two-hundred individuals from the United States voluntarily participated in this study via AMT for monetary compensation. Participant recruitment was restricted to individuals in the United States with a prior approval rating above 95%. Thirty participants were excluded for failing to answer all questions about each memory, for providing clearly nonsensical responses to memory cue, for recalling an event that occurred more than 10 years ago, or for failing the attention check at the end (see attention check below). As such, data were analyzed with the remaining 170 participants ($M_{\text{age}} = 32.96$ years, $SD = 8.61$, age range = [20, 69], 68 females, 101 males).

3.1.2. Procedure

The study was self-paced. Participants were asked to recall a total of six distinct events, one at a time, from their personal pasts that occurred within the past 10 years. Participants were provided with a unique cue for each of the six memories: (1) recall a specific past experience in which you were dishonest with another person, and you believe your action was *very* morally wrong (or your most morally wrong dishonest behavior from the past 10 years); (2) recall a specific past experience in which you were dishonest with another person, and you believe your action was just *slightly* morally wrong; (3) recall a specific past experience in which you harmed another person, and you believe your action was *very* morally wrong (or your most morally wrong harmful behavior from the past 10 years); (4) recall a specific past experience in which you harmed another person, and you believe your action was just *slightly* morally wrong; (5) recall a specific past experience in which you were unfair to another person, and you believe your action was *very* morally wrong (or your most morally wrong unfair behavior from the past 10 years); (6) recall a specific past experience in which you were unfair to another person, and you believe your action was just *slightly* morally wrong; We randomized the order in which these cues were presented across participants. This cueing procedure ultimately produced three distinct matched pairs of remembered actions—with one judged to be more morally wrong than the other—for each kind of behavior: (1) dishonesty; (2) harming; (3) and unfairness.

For each remembered behavior, participants described the event in two to five sentences. They then typed in the month and year in which it occurred, and they selected one of the following options to best describe when it occurred: within the past day, within the past week, within the past 2 weeks, within the past month, within the past 2 months, within the past 6 months, within the past year, within the past 2 years, within the past 5 years, within the past 10 years. As a manipulation check, participants answered the following question to assess the severity of the moral transgression: "how morally wrong was your behavior in this instance?" (1 = *slightly morally wrong*, 7 = *very morally wrong*). Two different questions about participants' emotional experience were then asked in a random order: "as you remember the event now, how positive or negative are your emotions?" (1 = *very negative*, 7 = *very positive*), and "as you remember the event now, how intense are your emotions?" (1 = *not at all intense*, 7 = *very intense*). Next, two questions about the frequency of having retrieved the events were asked in a random order: "since it happened, how often have you willfully thought about the event in your mind or talked about it?" (1 = *never*, 7 = *very often*), and "since it happened, has the memory of the event suddenly popped up in your thoughts by itself—that is, without your having attempted to remember it?" (1 = *never*, 7 = *very often*). Participants were then asked the following question about the frequency of morally upward counterfactual thinking for each remembered event: "since it happened, how often have you thought about or talked about morally better ways in which you could have acted?" (1 = *never*, 7 = *very often*). Finally, participants were asked about their current intentions to behave differently in the future: "if you were to find yourself in a similar situation in the future, would you act in a morally better way?" (1 = *definitely no*, 7 = *definitely yes*).

After completing all ratings for all six memories, participants responded to the same attention check question as in Study 1. As in Study 1, we excluded participants who reported being distracted, having trouble paying attention, failing to avoid distractions, and not taking the survey seriously. Participants then completed several demographics questions. Upon completion, participants were monetarily compensated for their time.

3.1.3. Data analyses

The statistical approach and software packages used in Study 1 were also used in Study 2.

3.2. Results

3.2.1. Manipulation check

An initial LMEM was computed to ensure that remembered actions generated from the *very* morally wrong cue were, in fact, judged to be more morally wrong on the 7-point scale than remembered actions generated from the *slightly* morally wrong cue. This expectation was corroborated for all kinds of remembered transgressions: for those involving dishonesty ($b = 3.34$, $SE = 0.15$, $t = 22.19$, $p < .001$, 95% CI [3.04, 3.64]), for those involving harm ($b = 3.00$, $SE =0\ .15$, $t = 19.53$, $p < .001$, 95% CI [2.68, 3.31]), and for those involving unfairness ($b = 2.98$, $SE = 0.15$, $t = 19.46$, $p < .001$, 95% CI [2.69, 3.27]). So, this binary variable indexing the judged severity of remembered moral transgressions will be used in subsequent analyses.

3.2.2. Emotional experience while remembering

We tested the hypothesis that participants would report experiencing stronger negative emotions when recalling their more severe, relative to more minor, moral transgressions. LMEMs with the judged severity of the moral transgression predicting valence revealed a significant effect of the severity of the moral transgression for dishonesty violations ($b = -1.03$, $SE = 0.12$, $t = -8.55$, $p < .001$, 95% CI [−1.26, −0.79]), harm violations ($b = -.80$, $SE = 0.12$, $t = -6.86$, $p < .001$, 95% CI [−1.02, −0.57]), and unfairness violations ($b = -.70$, $SE = 0.11$, $t = -6.40$, $p < .001$, 95% CI [−0.92, −0.48]) (see Fig. 1a). In all

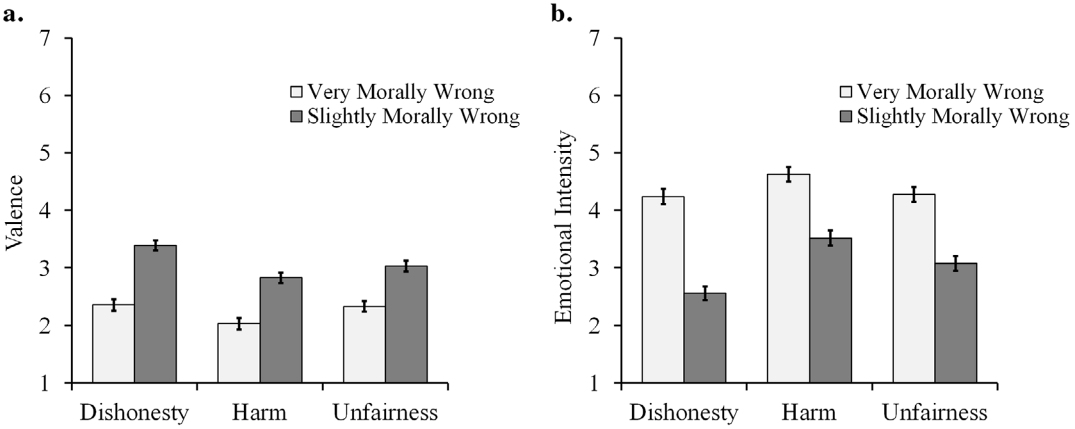


Fig. 1. Means and standard error bars are depicted for valence (a) and emotional intensity ratings (b) for remembered transgressions involving dishonesty, harm, and unfairness, all as a function of the judged severity of the moral transgression (slightly morally wrong *versus* very morally wrong).

cases, past transgressions judged to be more morally wrong tended to be experienced more negatively relative to transgressions judged to be less morally wrong. Moreover, LMEMs with the judged severity of the moral transgression predicting emotional intensity revealed a significant effect of the severity of the moral transgression for dishonesty violations (b = 1.68, SE = 0.16, t = 10.77, p < .001, 95% CI [1.36, 2.00]), harm violations (b = 1.11, SE = 0.16, t = 6.88, p < .001, 95% CI [0.81, 1.43]), and unfairness violations (b = 1.20, SE = 0.14, t = 8.53, p < .001, 95% CI [0.93, 1.47]) (see Fig. 1b). In all cases, past transgressions judged to be more morally wrong tended to be experienced with greater emotional intensity relative to transgressions judged to be less morally wrong.

The cued morality (slightly morally wrong *vs*. very morally wrong) of the remembered behaviors was significantly related to time$_A$ and time$_B$ for dishonesty, harm, and unfairness violations (all ps < .001; Supplemental Table 1). In all cases, remembered behaviors judged to be more morally wrong occurred in the more distant past. So, we sought to ensure that the effect of the cued severity of the remembered transgression on valence and emotional intensity persisted after statistically controlling for time$_A$ and time$_B$. The effect of the cued severity of the moral transgression on valence and emotional intensity did, in fact, remain significant after controlling for time$_A$ and time$_B$ for violations involving dishonesty, harm, and unfairness (all ps < .001). Supplemental Table 2 depicts full results from these models with valence as the outcome variable, and Supplemental Table 3 depicts full results from these models with emotional intensity as the outcome variable.

3.2.3. Voluntary and involuntary recall

We hypothesized that participants would report having more frequently recalled their more severe, relative to more minor, moral transgressions. LMEMs with the judged severity of the moral transgressions predicting the reported frequency of voluntary recall revealed a

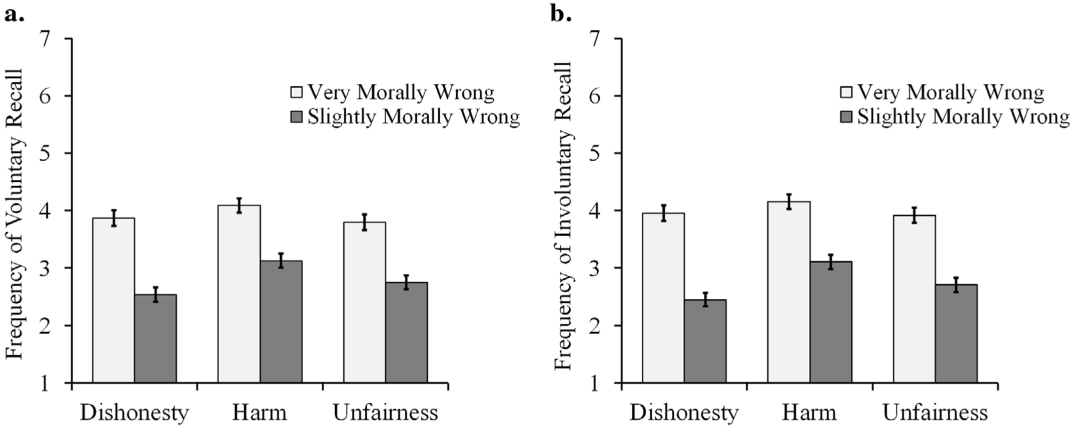


Fig. 2. Means and standard error bars are depicted for the frequency of voluntary (a) and involuntary (b) recall for remembered transgressions involving dishonesty, harm, and unfairness, all as a function of the judged severity of the moral transgression (slightly morally wrong *versus* very morally wrong).

significant effect of the severity of the moral transgressions for dishonesty violations (b = 1.34, SE = 0.17, t = 7.82, p < .001, 95% CI [0.99, 1.65]), harm violations (b = .96, SE = 0.15, t = 6.35, p < .001, 95% CI [0.66, 1.25]), and unfairness violations (b = 1.05, SE = 0.14, t = 7.33, p < .001, 95% CI [0.78, 1.34]) (see Fig. 2a). In all cases, participants reported that voluntary recall was more frequent after transgressions judged to be more morally wrong. Moreover, LMEMs with the judged severity of the moral transgressions predicting the reported frequency of involuntary recall revealed a significant effect of the severity of the moral transgressions for dishonesty violations (b = 1.51, SE = 0.15, t = 9.87, p < .001, 95% CI [1.23, 1.81]), harm violations (b = 1.05, SE = 0.15, t = 7.07, p < .001, 95% CI [0.74, 1.37]), and unfairness violations (b = 1.22, SE = 0.14, t = 8.86, p < .001, 95% CI [0.95, 1.47]) (see Fig. 2b). In all cases, participants reported that involuntary recall was more frequent after transgressions judged to be more morally wrong.

As before, we computed additional LMEMs to ensure that the effects of the judged severity of the moral transgressions on the reported frequency of voluntary and involuntary recall were not merely a by-product of differences in when the events actually occurred in the past. These additional LMEMs revealed that the effects remained statistically significant after controlling for time$_A$ and time$_B$ in separate models (all ps < .001). Supplemental Table 4 depicts full results from these models with the frequency of voluntary recall as the outcome variable, and Supplemental Table 5 depicts full results from these models with the frequency of involuntary recall as the outcome variable.

3.2.4. *Frequency of morally upward counterfactual thinking*

We hypothesized that participants would report having generated morally upward counterfactuals more frequently for their more severe, relative to more minor, moral transgressions. LMEMs with the judged severity of the moral transgressions predicting the frequency of

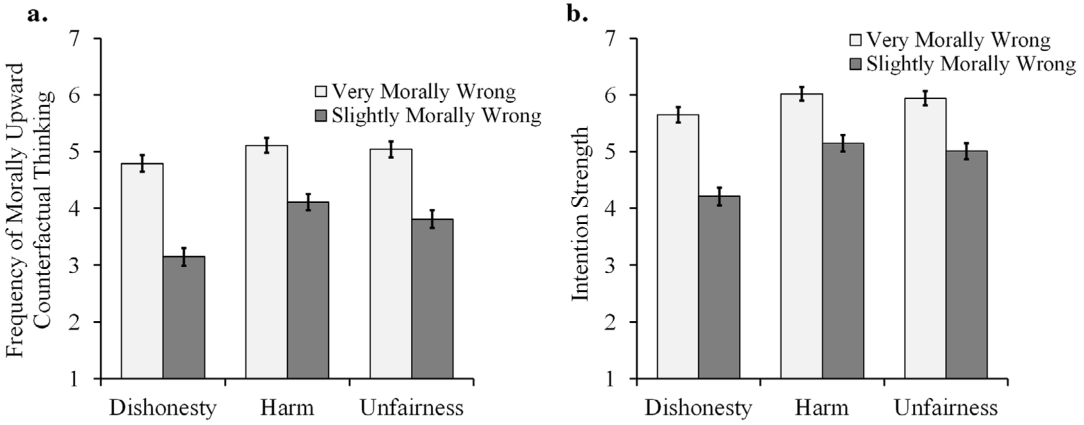


Fig. 3. Means and standard error bars are depicted for the frequency of morally upward counterfactual thinking (a) and strength of intentions to show moral improvement in the future (b) for transgressions involving dishonesty, harm, and unfairness, all as a function of the judged severity of the moral transgression (slightly morally wrong *versus* very morally wrong).

morally upward counterfactual thinking revealed a significant effect of the severity of the moral transgression for dishonesty violations ($b = 1.64$, $SE = 0.18$, $t = 9.03$, $p < .001$, 95% CI [1.29, 1.98]), harm violations ($b = .99$, $SE = 0.18$, $t = 5.41$, $p < .001$, 95% CI [0.63, 1.34]), and unfairness violations ($b = 1.22$, $SE = 0.17$, $t = 7.06$, $p < .001$, 95% CI [0.88, 1.55]) (see Fig, 3a). In all cases, for past transgressions judged to be more morally wrong, participants reported that they tended to more frequently consider morally upward counterfactuals about those events.

To ensure that the effects of the judged severity of the moral transgressions on the frequency of morally upward counterfactual thinking was not merely a by-product of differences in objective temporal distance, we computed two additional LMEMs. The results indicate that the judged severity of the moral transgressions remain significantly related to the reported frequency of counterfactual thinking even after controlling for time_A and time_B in separate models (all *p*s < .001). (See Supplemental Table 6 for full results.)

3.2.5. *Intentions for moral improvement*

We hypothesized that, when reflecting on their more severe past moral transgressions, participants would report stronger intentions to behave differently and better in the future. LMEMs with the judged severity of the moral transgressions predicting the strength of intentions to behave differently in the future revealed a significant effect of the severity of the moral transgression for dishonesty violations ($b = 1.44$, $SE = 0.19$, $t = 7.53$, $p < .001$, 95% CI [1.07, 1.81]), harm violations ($b = .87$, $SE = 0.18$, $t = 4.97$, $p < .001$, 95% CI [0.51, 1.19]), and unfairness violations ($b = .93$, $SE = 0.16$, $t = 5.71$, $p < .001$, 95% CI [0.62, 1.26]) (Fig. 3b). In all cases, for past transgressions judged to be more morally wrong, relative to less morally wrong, participants tended to report stronger intentions to behave differently in the future.

3.2.6. Mediation models

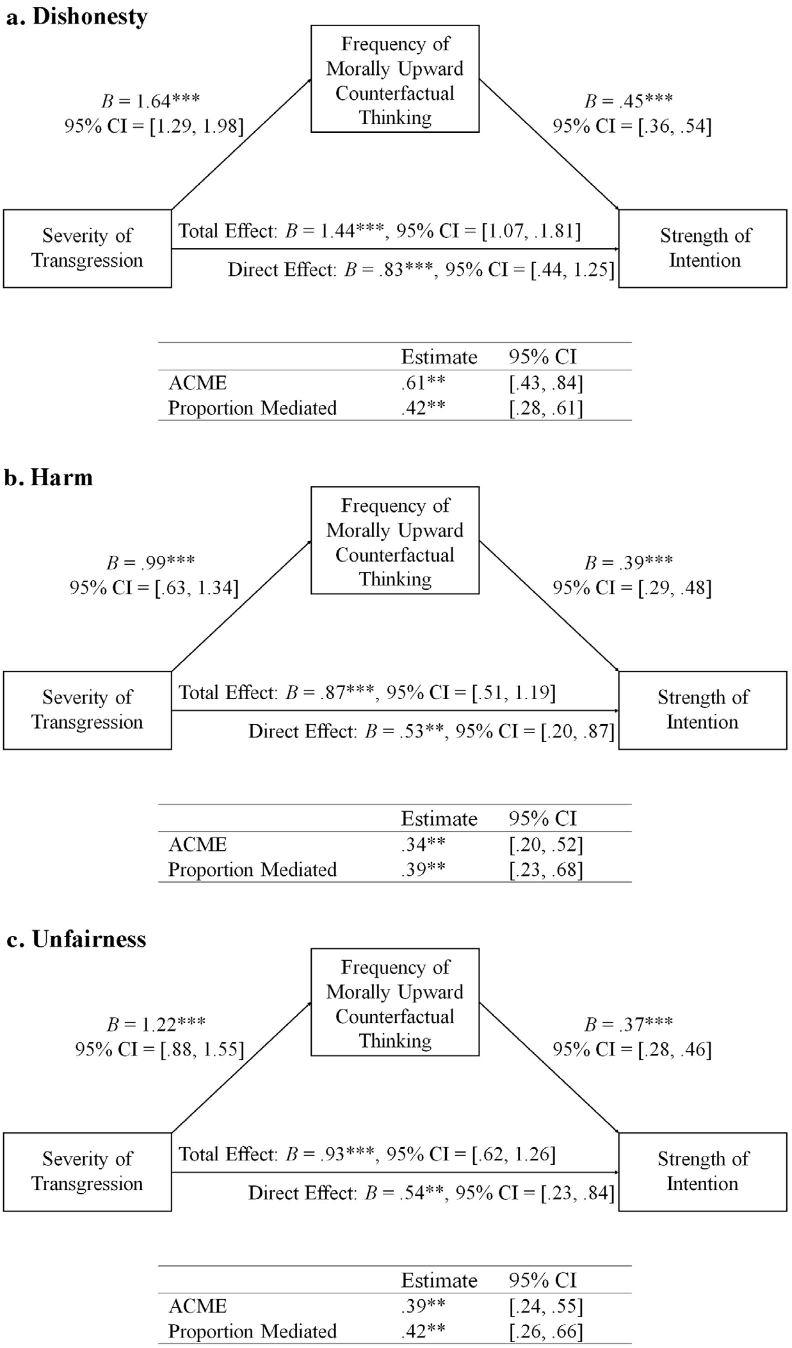
Since we hypothesized that people tend to more frequently generate morally upward counterfactuals after committing more severe moral transgressions to form stronger intentions to behave differently and better in the future, we conducted three separate mediation analyses for dishonesty, harm, and unfairness violations. The average causal mediation effect (ACME), or the indirect effect, and the proportion mediated were both computed using the 'mediation' package in R (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014). These analyses revealed that the frequency of counterfactual thinking does, in fact, mediate the relationship between the judged severity of the moral transgressions and intentions to behave differently in the future for all three kinds of violations. People reported having generated morally upward counterfactuals more frequently for remembered actions judged to be more morally wrong relative to less morally wrong, and this predicted stronger intentions to behave differently in the future. Fig. 4 depicts full results from all three mediation analyses.

3.3. Discussion

Overall, the results from Study 2 replicate the key results from Study 1 using a different cueing procedure to obtain memories of specific kinds of moral transgressions, namely those involving dishonesty, harm, and unfairness. For all three kinds of moral transgressions, participants reported experiencing stronger negative emotions when recalling their more severe moral transgressions relative to their less severe moral transgressions; participants reported having more frequently recalled—both voluntarily and involuntarily—their more severe moral transgressions relative to their less severe ones; and when they recalled their transgressions, they reported having generated morally upward counterfactuals more frequently for their more severe moral transgressions than for their less severe moral transgressions. All these effects remained significant even after statistically controlling for when the events occurred in the past.

Our results from Study 2 also extend those from Study 1. Specifically, for all three kinds of moral transgressions, the judged severity of the transgressions predicted intentions to behave differently and better in the future. When the remembered transgression was judged to be more severe (relative to less severe), participants formed stronger intentions to behave in a morally better way in the future. Critically, for all three kinds of moral transgressions, the reported frequency of simulating morally better ways of acting also predicted intentions to behave differently and better in the future. The more frequently participants reported having simulated morally better ways in which they could have acted instead, the stronger their intentions were to behave differently in the future.

The findings from Study 2 provide further support for a directive function of remembering past moral transgressions. There may be functional utility in remembering and reflecting upon our severe moral transgressions, even if remembering them induces negative affect and threatens a morally good self-concept. The reported frequency of considering morally better ways in which they could have acted predicts the strength of intentions to behave in a morally better way in the future.


a. Dishonesty
Frequency of Morally Upward Counterfactual Thinking
B = 1.64***
95% CI = [1.29, 1.98]
B = .45***
95% CI = [.36, .54]
Severity of Transgression
Total Effect: B = 1.44***, 95% CI = [1.07, .1.81]
Direct Effect: B = .83***, 95% CI = [.44, 1.25]
Strength of Intention

	Estimate	95% CI
ACME	.61**	[.43, .84]
Proportion Mediated	.42**	[.28, .61]

b. Harm
Frequency of Morally Upward Counterfactual Thinking
B = .99***
95% CI = [.63, 1.34]
B = .39***
95% CI = [.29, .48]
Severity of Transgression
Total Effect: B = .87***, 95% CI = [.51, 1.19]
Direct Effect: B = .53**, 95% CI = [.20, .87]
Strength of Intention

	Estimate	95% CI
ACME	.34**	[.20, .52]
Proportion Mediated	.39**	[.23, .68]

c. Unfairness
Frequency of Morally Upward Counterfactual Thinking
B = 1.22***
95% CI = [.88, 1.55]
B = .37***
95% CI = [.28, .46]
Severity of Transgression
Total Effect: B = .93***, 95% CI = [.62, 1.26]
Direct Effect: B = .54**, 95% CI = [.23, .84]
Strength of Intention

	Estimate	95% CI
ACME	.39**	[.24, .55]
Proportion Mediated	.42**	[.26, .66]

Fig. 4. Separate mediation models are depicted for remembered transgressions involving dishonesty (a), harm (b), and unfairness (c). For all three kinds of transgressions, the frequency of counterfactual thinking mediates the relationship between the judged severity of the moral transgressions and intentions to behave differently in the future.

4. Study 3

The primary purpose of Study 3 was to investigate whether punishment for wrongdoing moderates the effects obtained in the previous studies. Moral punishment may play a critical role in moral learning by discouraging future behaviors that violate moral norms widely held by members of a community (Cushman, 2015). Adopting this pedagogical account, being punished for wrongdoing may boost the frequency with which we recall our transgressions, the frequency with which we think about morally better ways in which we could have acted instead and, ultimately, the strength of our intentions to behave differently and better in the future. In this way, recalling and counterfactually mutating personal past events in which we transgressed may help to facilitate this pedagogical function of moral punishment.

4.1. Materials and method

4.1.1. Participants

Five-hundred fifty-one individuals from the United States voluntarily participated in this study via AMT for monetary compensation. Participant recruitment was restricted to individuals in the United States who had completed at least 500 HITs with a prior approval rating above 95%. Eighty-seven participants were excluded for failing to answer all questions about each memory, for providing clearly nonsensical responses to memory cue, for recalling an event that occurred before 2010, for failing the punishment manipulation check, or for failing the attention check at the end (see attention check below). As such, data were analyzed with the remaining 464 participants ($M_{\text{age}} = 40.09$ years, $SD = 11.84$, age range = [19, 78], 205 females, 256 males). We aimed to recruit at least 550 participants to ensure that we would have at least 100 participants per cell of the 2 × 2 design, after expected exclusions.

4.1.2. Procedure

The study was self-paced. Participants were asked to recall one of four events from their personal pasts that occurred since 2010 (this study was conducted in April 2021). Specifically, participants were randomly assigned to one these memory cues: (1) recall a specific past experience in which you did something *very* morally wrong (or your most morally wrong behavior between 2010 and now), and you were punished for your action; (2) recall a specific past experience in which you did something just *slightly* morally wrong, and you were punished for your action; (3) recall a specific past experience in which you did something *very* morally wrong (or your most morally wrong behavior between 2010 and now), and you were not punished for your action; and (4) recall a specific past experience in which you did something just *slightly* morally wrong, and you were not punished for your action. This cueing procedure ultimately yielded a 2 (punishment vs. no punishment) × 2 (slightly morally wrong cue vs. very morally wrong cue) between-subjects design. Note that because the two measures of calendar time from the previous studies (time_{A} and time_{B}) were so closely related to each another, we simplified Study 3 by only measuring time_{B} (i.e., the number of months passed since the event occurred).

For each remembered behavior, participants described the event in two to five sentences. They then typed in the month and year in which it occurred. Participants were then presented with two manipulation checks. First, participants answered the following question to assess the severity of the moral transgression: “how morally wrong was your behavior in this instance?” (1 = *slightly morally wrong*, 7 = *very morally wrong*). Second, participants indicated whether they were punished or not for their transgression (binary response: yes vs. no). After these manipulation checks, two different questions about participants’ emotional experience were asked in a random order: “as you remember the event now, how positive or negative are your emotions?” (1 = *very negative*, 7 = *very positive*), and “as you remember the event now, how intense are your emotions?” (1 = *not at all intense*, 7 = *very intense*). Next, two questions about the frequency of having retrieved the events were asked in a random order: “since it happened, how often have you willfully thought about the event in your mind or talked about it?” (1 = *never*, 7 = *very often*), and “since it happened, has the memory of the event suddenly popped up in your thoughts by itself—that is, without your having attempted to remember it?” (1 = *never*, 7 = *very often*). Participants were then asked the following question about the frequency of morally upward counterfactual thinking for each remembered event: “since it happened, how often have you thought about or talked about morally better ways in which you could have acted?” (1 = *never*, 7 = *very often*). Finally, participants were asked about their current intentions to behave differently in the future: “if you were to find yourself in a similar situation in the future, would you act in a morally better way?” (1 = *definitely no*, 7 = *definitely yes*).

After completing all ratings for the memory, participants responded to the same attention check question as in the previous studies. As in the previous studies, we excluded participants who reported being distracted, having trouble paying attention, failing to avoid distractions, and not taking the survey seriously. Participants also completed several demographics questions. Upon completion, participants were monetarily compensated for their time.

4.2. Results

4.2.1. Manipulation check

An initial 2 × 2 analysis of variance (ANOVA) was computed to ensure that remembered actions generated from the *very* morally wrong cue were, in fact, judged to be more morally wrong on the 7-point scale than remembered actions generated from the *slightly* morally wrong cue, and that moral wrongness judgments did not significantly differ as a function of the punishment condition (punishment vs. no punishment). This 2 (slightly morally wrong vs. very morally wrong) × 2 (punishment vs. no punishment) revealed a significant main effect of moral severity ($F(1, 460) = 156.35, p < .001, \eta_p{}^2 = .25$), but no significant main effect of punishment ($F(1, 460) = 0.61, p = .44, \eta_p{}^2 = .001$) and no significant interaction between moral severity and punishment ($F(1, 460) = 1.61, p = .21, \eta_p{}^2 = .003$). So, this binary variable indexing the judged severity of remembered moral transgressions will be used in subsequent analyses.

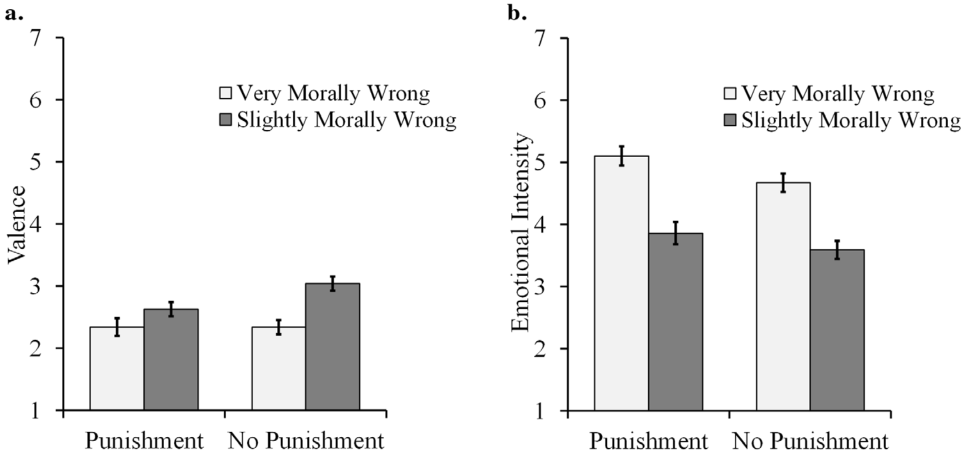


Fig. 5. Means and standard error bars are depicted for valence (a) and emotional intensity ratings (b) for remembered transgressions as a function of the judged severity of the moral transgression (slightly morally wrong *versus* very morally wrong) and whether the participant was punished for transgressing (punishment *versus* no punishment).

4.2.2. Emotional experience while remembering

We tested the hypothesis that participants would report experiencing stronger negative emotions when recalling their more severe, relative to more minor, moral transgressions. We also tested whether having been punished for transgressing moderates the effect of moral severity on emotion. To these ends, we first computed a 2 (slightly morally wrong vs. very morally wrong) × 2 (punishment vs. no punishment) ANOVA with valence as the outcome variable. There was a main effect of moral severity ($F(1, 460) = 16.73, p < .001, \eta_p{}^2 = .04$), but no main effect of punishment ($F(1, 460) = 2.78, p = .096, \eta_p{}^2 = .01$) and no interaction between moral severity and punishment ($F(1, 460) = 2.79, p = .095, \eta_p{}^2 = .01$). Past transgressions judged to be more morally wrong tended to be experienced more negatively than transgressions judged to be less morally wrong (see Fig. 5a). We then computed a 2 (slightly morally wrong vs. very morally wrong) × 2 (punishment vs. no punishment) ANOVA with emotional intensity as the outcome variable. There was a significant main effect of moral severity ($F(1, 460) = 55.66, p < .001, \eta_p{}^2 = .11$) and a significant main effect of punishment ($F(1, 460) = 4.99, p = .026, \eta_p{}^2 = .01$), but no significant interaction between moral severity and punishment ($F(1, 460) = 0.24, p = .62, \eta_p{}^2 = .001$). Past transgressions judged to be more morally wrong tended to be experienced with greater emotional intensity than transgressions judged to be less morally wrong. Past transgressions that were punished also tended to be experienced with greater emotional intensity than transgressions that were not punished (see Fig. 5b).

As in the previous studies, transgressions judged to be more morally wrong, relative to less morally wrong, occurred in the more distant past ($M_{\text{diff}} = 8.77$ months, $SE_{\text{diff}} = 3.39$, $t(462) = 2.59$, $p = .010$, 95% CI [2.10, 15.43], Cohen's $d = 0.24$). Because of this, we sought to ensure that the effects of the judged severity of the moral transgression remained significantly related to valence and emotional intensity after statistically controlling for the

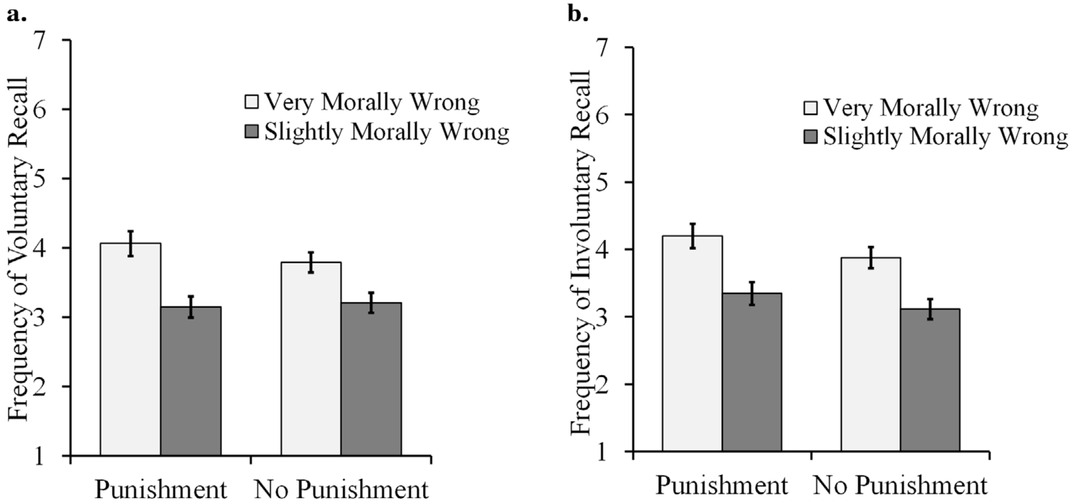


Fig. 6. Means and standard error bars are depicted for the frequency of voluntary (a) and involuntary (b) recall as a function of the judged severity of the moral transgression (slightly morally wrong *versus* very morally wrong) and whether the participant was punished for transgressing (punishment *versus* no punishment).

number of months passed since the event occurred. The same patterns of results were, in fact, obtained when statistically controlling for the number of months passed since the event occurred (see Supplemental Tables 7 and 8 for full results).

4.2.3. Voluntary and involuntary recall

We tested the hypothesis that participants would report having more frequently recalled their more severe, relative to more minor, moral transgressions. We also tested whether having been punished for transgressing moderates the effect of moral severity on the reported frequency of recall. To these ends, we first computed a 2 (slightly morally wrong vs. very morally wrong) × 2 (punishment vs. no punishment) ANOVA with the reported frequency of voluntary recall as the outcome variable. There was a main effect of moral severity ($F(1, 460) = 23.07$, $p < .001$, $\eta_p^2 = .05$), but no main effect of punishment ($F(1, 460) = 0.47$, $p = .49$, $\eta_p^2 = .001$) and no interaction between moral severity and punishment ($F(1, 460) = 1.09$, $p = .30$, $\eta_p^2 = .002$). Participants reported that past transgressions judged to be more morally wrong tended to be voluntary recalled more frequently than transgressions judged to be less morally wrong (see Fig. 6a). We then computed a 2 (slightly morally wrong vs. very morally wrong) × 2 (punishment vs. no punishment) ANOVA with the reported frequency of involuntary recall as the outcome variable. There was a main effect of moral severity ($F(1, 460) = 24.83$, $p < .001$, $\eta_p^2 = .05$), but no main effect of punishment ($F(1, 460) = 2.87$, $p = .091$, $\eta_p^2 = .01$) and no interaction between moral severity and punishment ($F(1, 460) = 0.08$, $p = .78$, $\eta_p^2 = .000$). Participants reported that past transgressions judged to be more morally wrong tended to be involuntary recalled more frequently than transgressions judged to be less morally wrong (see Fig. 6b). The same patterns of results were obtained

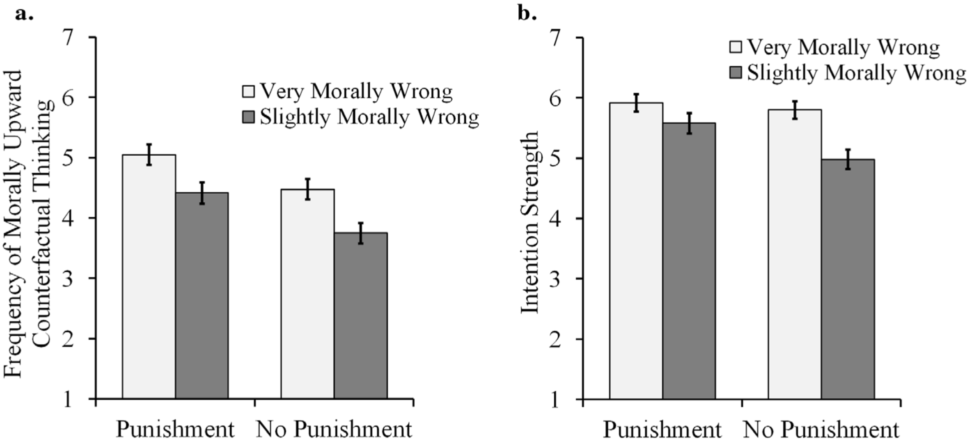


Fig. 7. Means and standard error bars are depicted for the frequency of morally upward counterfactual thinking (a) and strength of intentions to show moral improvement in the future (b) for transgressions as a function of the judged severity of the moral transgression (slightly morally wrong *vs*. very morally wrong) and whether the participant was punished for transgressing (punishment *vs*. no punishment).

when statistically controlling for the number of months passed since the event occurred (see Supplemental Tables 9 and 10 for full results).

4.2.4. *Frequency of morally upward counterfactual thinking*

We tested the hypothesis that participants would report having generated morally upward counterfactuals more frequently for their more severe, relative to more minor, moral transgressions. We also tested whether having been punished for transgressing moderates the effect of moral severity on the reported frequency of morally upward counterfactual thinking. To these ends, we computed a 2 (slightly morally wrong vs. very morally wrong) × 2 (punishment vs. no punishment) ANOVA with the reported frequency of morally upward counterfactual thinking as the outcome variable. There was a main effect of moral severity ($F(1, 460) = 15.63$, $p < .001$, $\eta_p{}^2 = .03$) and a main effect of punishment ($F(1, 460) = 12.99$, $p < .001$, $\eta_p{}^2 = .03$), but no interaction between moral severity and punishment ($F(1, 460) = 0.10$, $p = .76$, $\eta_p{}^2 = .000$). Participants reported having generated morally upward counterfactuals more frequently for their more severe moral transgressions than for their less severe transgressions. Participants also reported having generated morally upward counterfactuals more frequently when they were punished for their transgressions relative to when they were not punished (see Fig. 7a). The same pattern of results was obtained when statistically controlling for the number of months passed since the event occurred (see Supplemental Table 11 for full results).

4.2.5. *Intentions for moral improvement*

We tested the hypothesis that, when reflecting on their more severe past moral transgressions, participants would report stronger intentions to behave differently and better in the future. We also tested whether having been punished for transgressing moderates the effect of moral severity on the reported strength of future intentions. To these ends, we computed a 2 (slightly morally wrong vs. very morally wrong) × 2 (punishment vs. no punishment) ANOVA with reported intention strength as the outcome variable. There was a main effect of moral severity ($F(1, 460) = 14.19, p < .001, {\eta_\mathrm{p}}^2 = .03$) and a main effect of punishment ($F(1, 460) = 5.29, p = .022, {\eta_\mathrm{p}}^2 = .01$), but no interaction between moral severity and punishment ($F(1, 460) = 2.37, p = .12, {\eta_\mathrm{p}}^2 = .005$). Participants reported stronger intentions for moral improvement for their more severe moral transgressions than for their less severe transgressions. Participants also reported stronger intentions for moral improvement when they were punished for their transgressions relative to when they were not punished (see Fig. 7b). The same pattern of results was obtained when statistically controlling for the number of months passed since the event occurred (see Supplemental Table 12 for full results).

4.2.6. Mediation model

Since we hypothesized that people tend to more frequently generate morally upward counterfactuals after committing more severe moral transgressions to form stronger intentions to behave differently and better in the future, we conducted a mediation analysis. The average causal mediation effect (ACME), or the indirect effect, and the proportion mediated were both computed using the "mediation" package in R (Tingley et al., 2014). Collapsing across punishment conditions, we found that the reported frequency of counterfactual thinking mediates the relationship between the judged severity of the moral transgressions and reported intentions to behave differently (ACME = 0.24, $p < .001$, 95% CI [0.11, 0.37]; Prop. Mediated = 0.39, $p < .001$, 95% CI [0.19, 0.76]). Participants reported having generated morally upward counterfactuals more frequently for remembered actions judged to be more morally wrong relative to less morally wrong ($M_\mathrm{diff} = 0.68$, $SE_\mathrm{diff} = 0.17$, $t(462) = 3.92$, $p < .001$, 95% CI [0.34, 1.02], Cohen's $d = 0.37$), and the reported frequency of morally upward counterfactual thinking was positively related to the strength of intentions to behave differently in the future ($r(462) = .40$, $p < .001$, 95% CI [0.32, 0.47]).

4.3. Discussion

The findings from Study 3 suggest that being punished for wrongdoing may boost the reported frequency with which people consider morally better ways in which they could have acted instead and the reported strength of their intentions for future moral improvement. Interestingly, we found no positive evidence that punishment moderates the reported frequency of recalling the transgressions (voluntarily or involuntarily); the effect of punishment was specific to how often participants reported having considered counterfactually *mutated* versions of the remembered event as opposed to just having remembered the event itself. Overall, these findings are consistent with a pedagogical function of moral punishment, and mutating personal past events in which we committed moral transgressions may help to facilitate this pedagogical function. In addition, Study 3 successfully replicated (conceptually) the effects from the previous studies with a different experimental design.

5. Study 4

Study 4 further examined the possible directive function of remembering past moral transgressions with an experimental manipulation to make accessible different kinds of simulations. More specifically, the primary purpose of Study 4 was to test the hypothesis that making accessible a morally upward counterfactual strengthens intentions for moral improvement—relative to resimulating the remembered event or making accessible a morally downward counterfactual. To this end, participants recalled a past cheating transgression, and then they simulated a morally upward counterfactual, simulated a morally downward counterfactual, or resimulated the remembered event. Participants then reported their intentions to cheat in the future.

Our central hypothesis in Study 4 posits a *content-specific pathway* (Epstude & Roese, 2008; Roese & Epstude, 2017; Smallman & Roese, 2009) by which counterfactual simulations in one domain of morality (cheating) influence intentions for future behavior within that same narrow domain. That is, both the intentions and simulated counterfactuals address cheating, a specific and narrow kind of moral transgression. As a secondary, exploratory objective, we also investigated whether recalling a cheating transgression and simulating an upward counterfactual about that cheating transgression influences other intentions through *content-neutral pathways*. A content-neutral pathway involves counterfactuals that influence intentions in domains that are *independent* of the counterfactual content and context (Epstude & Roese, 2008; Roese & Epstude, 2017; Smallman & Roese, 2009). That is, the simulating counterfactuals in one domain (e.g., cheating) may elicit the behavioral intentions in a different domain (e.g., intentions involving loyalty, theft, or showing disrespect). Outside of the moral domain, there is some evidence supporting this content-neutral pathway for behavior: negative affect and counterfactual generation evoked by failure induces greater effort and striving on certain unrelated, subsequent tasks (Markman et al., 2008; McMullen & Markman, 2000). Other research, however, has not found positive evidence for upward counterfactual simulation strengthening unrelated intentions through this content-neutral pathway (e.g., Smallman & Roese, 2009). Because past research has produced mixed support for this content-neutral pathway, no specific hypothesis was generated.

5.1. *Materials and method*

5.1.1. *Participants*

Five-hundred one individuals from the United States voluntarily participated in this study via AMT for monetary compensation. Participant recruitment was restricted to individuals in the United States with a prior approval rating above 95%. Eighty-five participants were excluded for failing to answer all questions about each memory, for providing clearly nonsensical responses to memory cue, or for failing either of the two attention checks (see below for details). As such, data were analyzed with the remaining 416 participants ($M_{\text{age}} = 37.20$, $SD = 10.39$, age range = [20, 75], 176 females, 237 males). The sample size was determined to ensure that we would have more participants per condition (at least 100 participants per condition after exclusions) than other recent investigations into relationships between counterfactual thinking and intention formation (Smallman, 2013; Smallman & McCulloch, 2012; Smallman & Roese, 2009).

5.1.2. Procedure

The study was self-paced. Participants were asked to recall and describe, in two to five sentences, an event in which they cheated and believe their act of cheating was morally wrong. We cued participants to recall a case of cheating, because cheating transgressions are specific, commonplace, and widely studied in the literature. Participants were then randomly assigned to one of three possible simulation conditions in a between-subjects fashion. In the *morally upward counterfactual* condition, participants described, in two to five sentences, an alternative way in which they could have acted in the remembered event that would have been morally better. In the *morally downward counterfactual* condition, participants described, in two to five sentences, an alternative way in which they could have acted in the remembered event that would have been morally worse. In the *recall* condition, participants described, in two to five sentences, the same memory again using different language. These three conditions were developed to be comparable in cognitive demand and consequential thinking (see Kray et al. (2010) for a similar design and a similar point). To ensure that participants followed instructions, we then asked participants whether they described (1) an alternative way in which they could have acted that would have been morally better, (2) an alternative way in which they could have acted that would have been morally worse, or (3) the memory again as they believe it actually happened.

Participants then completed the intention judgment phase of the study (adapted from Smallman & Roese, 2009), in which participants indicated whether they would perform specific behaviors in the future (1 = *definitely no*, 7 = *definitely yes*). We included one critical item embedded in a set of 14 total items to test the content-specific pathway. For this critical item, participants indicated whether they would cheat in the future. To test the content-neutral pathway, the remaining items described other possible violations of moral (e.g., disloyalty to a friend) and social (e.g., wearing clothes backwards) norms. Examples of the content-neutral moral items include “In the future I will be disloyal to a friend” and “In the future I will steal something that does not belong to me.” Examples of content-neutral social norm items include “In the future I will talk to myself in public” and “In the future I will eat soup with a spoon.” Note that all social norm items were adapted from Clifford, Iyengar, Cabeza, and Sinnott-Armstrong (2015). The inclusion of these additional content-neutral items was meant to help conceal the aims of the study, to reduce demand characteristics, and to conduct exploratory analyses for possible content-neutral effects (see Supplemental material for all items).

After rating all items in the intention judgment phase of the study, participants were asked whether they paid attention, avoided distractions, and took the survey seriously. As in the previous studies, we excluded participants who reported being distracted, having trouble paying attention, failing to avoid distractions, and not taking the survey seriously. Participants then completed several demographics questions. Upon completion, participants were monetarily compensated for their time.

Table 2
Means (SDs) for intention judgments ratings split by condition and intention-type

Condition	Cheating Intention	Average Moral Intention	Average Social Norm Intention
Morally upward counterfactual	2.52 (1.39)	2.20 (0.94)	2.30 (0.92)
Morally downward counterfactual	3.00 (1.54)	2.34 (1.01)	2.42 (1.05)
Resimulation	2.93 (1.64)	2.26 (1.06)	2.32 (1.06)

Note. Total $N = 416$.

5.2. Results

We hypothesized that making accessible a morally upward counterfactual after recalling a cheating transgression would strengthen intentions not to cheat in the future—relative to recalling the event as they believe it occurred and making accessible a morally downward counterfactual. We computed a one-way between-subjects ANOVA with simulation condition (morally upward counterfactual, morally downward counterfactual, or resimulation) on intentions to cheat in the future (i.e., the critical content-specific item). There was a significant effect of condition on cheating intention judgments ($F(2, 413) = 4.01$, $p = .019$, $\eta_p{}^2 = .02$). Subsequent post hoc comparisons revealed that participants in the morally upward counterfactual condition reported stronger intentions to *not* cheat in the future than participants in the resimulation condition ($M_{\text{diff}} = 0.41$, $SE_{\text{diff}} = 0.18$, $p = .025$, 95% CI [0.05, 0.76], Cohen's $d = 0.27$) and participants in the morally downward counterfactuals conditions ($M_{\text{diff}} = 0.48$, $SE_{\text{diff}} = 0.18$, $p = .009$, 95% CI [0.12, 0.84], Cohen's $d = 0.33$). There was no difference in cheating intentions between resimulation and morally downward counterfactual conditions ($M_{\text{diff}} = 0.07$, $SE_{\text{diff}} = 0.18$, $p = .70$, 95% CI [–0.29, 0.43], Cohen's $d = 0.04$). Table 2 presents means and standard deviations for each condition.

5.2.1. Exploratory analyses

To address possible content-neutral effects, we investigated whether making accessible a morally upward counterfactual after recalling a cheating transgression would strengthen intentions for moral improvement more generally. To this end, we computed an average intention judgment for the six content-neutral moral items ($M = 2.27$, $SD = 1.00$, $\alpha = .80$). A one-way between-subjects ANOVA was then computed with simulation condition (morally upward counterfactual, morally downward counterfactual, or resimulation) on intentions for general moral improvement. There was no significant effect of condition on average moral intention judgments ($F(2, 413) = 0.73$, $p = .48$, $\eta_p{}^2 = .004$). (See Table 2 for means and standard deviations for each condition.)

We also investigated whether making accessible a morally upward counterfactual after recalling a cheating transgression would strengthen intentions to obey social norms. To this end, we computed an average intention judgment for the seven social norm items ($M = 2.35$, $SD = 1.01$, $\alpha = .75$). A one-way between-subjects ANOVA was then computed with simulation condition (morally upward counterfactual, morally downward counterfactual, or resimulation) on intentions for obeying social norms. There was no significant effect of condition on average social norm intention judgments ($F(2, 413) = 0.57$, $p = .56$, $\eta_p^2 = .003$). (See Table 2 for means and standard deviations for each condition.)

5.3. Discussion

Overall, the results from Study 4 support our primary hypothesis: that making accessible a morally upward counterfactual strengthens intentions for moral improvement—relative to resimulating the remembered event as it occurred or making accessible a morally downward counterfactual. This result was only obtained via a content-specific pathway. That is, recalling a cheating transgression and simulating a morally upward counterfactual about that cheating transgression strengthened intentions not to cheat in the future, but it did not influence intentions for other kinds of moral behaviors (e.g., loyalty, theft) or for obeying certain social norms.

6. General discussion

In four studies, we investigated the role of remembering and reflecting on our past moral transgressions in service of learning from those past mistakes to facilitate moral improvement. Even though participants reported experiencing strong negative emotions when recalling their severe moral transgressions, they nevertheless tended to frequently recall and think about those transgressions, both voluntarily and involuntarily. To begin to explain this pattern of results, we found evidence that remembering and thinking about our own moral transgressions may serve a *directive function*. When participants recalled their moral transgressions, particularly those judged to be seriously morally wrong, they reported having frequently considered morally better ways in which they could have acted instead. The more that participants reported having simulated morally better ways in which they could have acted, the stronger their intentions were to behave differently and better in the future. An experimental manipulation then revealed that making accessible a morally upward counterfactual when recalling a moral transgression strengthens intentions for moral improvement in the future.

Recent research suggests that people tend to forget their own wrongdoings to reduce psychological distress and discomfort, while concomitantly protecting a favorable self-concept (Kouchaki & Gino, 2016; Reczek et al., 2017; Shu et al., 2011). In contrast, our findings suggest that people do remember their past moral transgressions, especially those they judge to be severe, and that they frequently retrieve and ruminate on them. This was the case even though frequently retrieving and ruminating on our particularly severe past transgressions is quite threatening to our favorable self-concepts. Frequently retrieving and thinking about our past transgressions seems to serve a directive function that influences intentions for moral improvement. Consequently, two distinct, seemingly conflicting psychological functions have been identified in the literature: maintaining a morally good self-concept may require forgetting some of our own past moral transgressions, but learning from our past mistakes and forming intentions for moral improvement is aided by the ability to remember and think about our own past moral transgressions. Importantly, however, our findings do not entail that people cannot forget at least some of their moral transgressions to maintain a morally good self-concept, just as prior work does not entail that all our past moral transgressions are forgotten such that those events could not serve a directive function. Future work will investigate the particular circumstances under which people forget their moral transgressions to maintain a morally good self-concept, and the particular circumstances under which people remember their moral transgressions to facilitate forming future intentions and engaging in particular behaviors.

Counterfactual thinking about past events provides us with the opportunity to imagine better or worse alternatives to reality. The *functional theory of counterfactual thinking* posits that simulating upward counterfactuals—especially after negative experiences—serves a preparatory function, helping people to learn from past mistakes, to solve problems, to form intentions for specific future behaviors, and to guide goal pursuit (Epstude & Roese, 2008; Roese, 1994, 1997; Roese & Epstude, 2017). Such counterfactual thoughts often come to mind effortlessly and involuntarily in our daily lives. Our results are broadly consistent with the functional theory of counterfactual thinking. We, however, provide several novel contributions to this theoretical framework and literature. First, we characterized and investigated a novel kind of episodic counterfactual thinking—morally upward counterfactual thinking—that occurs after committing moral transgressions. Second, we found that the reported *frequency* of simulating morally upward counterfactuals predicts the strength of intentions to behave differently in the future. The more that people reported having simulated upward counterfactuals after having negative experiences, the stronger their intentions are to behave differently in the future. Third, we found that having been punished for wrongdoing is associated with increased reported frequency of morally upward counterfactual thinking and stronger behavioral intentions.

Behavioral intentions, such as those formed after reflecting on specific past events, are effectively self-instructions for performing future actions in service of accomplishing particular goals (Sheeran & Webb, 2016). The concept of an intention has proven valuable for researchers interested in predicting actual future behavior and outcomes. Several prominent theoretical frameworks—e.g., the theory of reasoned action (Fishbein & Ajzen, 1975) and the theory of planned behavior (Ajzen, 1991)—posit that the most important and immediate predictor of behavior is the intention to perform it. Numerous studies have found that intentions do indeed predict actual behavior across diverse circumstances: Sheeran (2002) performed a meta-analysis on 10 existing meta-analyses investigating the relationship between intentions and behavior, finding that the sample-weighted average correlation between intentions and subsequent behavior was indicative of a large effect (see also, Sheeran & Webb, 2016). Consequently, there is reason to suspect that the strength of intentions to behave differently in the future after recalling our moral transgressions predicts the likelihood of then behaving in a more morally upstanding way in the future. With that being said, future work will more directly investigate the intention-behavior link in the moral domain to ensure that the intentions formed for moral improvement after reflecting on past transgressions actually predict behavior.

6.1. Limitations and future directions

Our studies do have some limitations worth noting. First, the relationship between the frequency of morally upward counterfactual thinking and the strength of intentions to behave differently in the future is correlational; this entails that we cannot draw strong conclusions about the frequency of morally upward counterfactual thinking *causing* the strength of intentions to behave differently. We do, however, believe that it was reasonable to include the frequency of morally upward counterfactual thinking as a mediator between the judged morality of the remembered action and the strength of intentions to behave differently in the future, given prior empirical findings that support established theory. The functional theory of counterfactual thinking posits a unidirectional relationship between upward counterfactual thinking and behavioral intentions (Roese & Epstude, 2017). That is, upward counterfactual simulations are thought to influence behavioral intentions (and not vice versa), and converging lines of empirical research support this contention (McCulloch & Smallman, 2014; Roese & Epstude, 2017; Roese, Park, Smallman, & Gibson, 2008; Smallman, 2013; Smallman & McCulloch, 2012; Smallman & Roese, 2009). From this theoretical perspective, an experience activates counterfactual thinking, counterfactual thinking activates intentions for future behavior, and behavioral intentions bring about the corresponding behavior (Epstude & Roese, 2008; Roese & Epstude, 2017; Smallman, 2013).

Second, we relied on participants' self-reports of their frequency of recall, frequency of upward counterfactual thinking, and intention strength. Our self-report measurement strategy is limited for two reasons. First, participants may not be able to accurately report how often they have reflected on and mutated these kinds of events, so these frequency judgments are likely noisy. Second, for the intention judgments in particular, participants' ratings may be influenced by social expectations as well as their own sense of morality. Most people believe they are morally good and want others to see them as morally good (Aquino & Reed, 2002; Stanley & De Brigard, 2019), so the intention judgments might have been inflated. Future research may be able to alleviate these concerns by employing non-retrospective methods (e.g., ecological momentary assessment, diaries, etc.) to obtain more accurate measures of frequency of recall and frequency of counterfactual simulation.

Third, in assessing the relationship between the frequency of morally upward counterfactual thinking and the strength of intentions to behave differently in the future, we did not examine the actual *content* of those counterfactual simulations in shaping those intentions. Recent work has suggested that the relationship between counterfactual thinking and behavioral intentions is influenced by the content of the simulated counterfactuals (Smallman, 2013). For example, focusing on highly specific counterfactuals as opposed to more abstract counterfactuals is more likely to strengthen behavioral intentions (Smallman, 2013). Future work will explore the content of morally upward counterfactual simulations (in addition to the frequency of simulation over time) in moderating the relationship between counterfactual thinking and behavioral intentions.

Despite these potential limitations, the studies reported herein have several strengths worth highlighting. For example, we consistently found support for our hypotheses across multiple different kinds of remembered transgressions, allowing for stronger generalizations. In addition, insights from contemporary moral psychology research have predominantly been acquired through the use of vignettes, questionnaire data, and hypothetical thought experiments (e.g., trolley problems). Although this research has produced valuable insights into moral judgment and decision making, the use of these materials and methods is rather limited by the artificial nature of the stimuli and situations that are far removed from the kinds of morally relevant situations we encounter in everyday life. In contrast, we made use of people's memories of moral transgressions that they had personally committed in the real world.

One other possible avenue for future research is exploring the possible moderating effect of public versus private transgressions on the frequency of recall, the frequency of counterfactual thinking, and intention strength. Public transgressions, especially those that involve close others, might be recalled more frequency, be mutated more frequently, and be associated with stronger intentions for future moral improvement than private transgressions. People might be more motivated to learn from their public transgressions to facilitate reputation management and the management of social relationships.

6.2. Conclusions

We have taken a functional theoretical approach to remembering past moral transgressions, finding that people do remember at least some their past transgressions and frequently consider morally better alternative ways in which they could have acted. Simulating these morally upward counterfactuals predicts intentions to behave in morally better ways in the future. Several lines of research have documented pronounced biases that help us to protect our favorable self-concepts against the unpleasant reality that we do act unethically (Stanley & De Brigard, 2019). Despite this motivation to protect our otherwise favorable self-concepts, we offer evidence that recalling and ruminating upon our past transgressions can actually serve a useful function, allowing us to learn from our more serious blunders to form intentions for future moral improvement. People are not necessarily destined to repeatedly commit the same moral transgressions throughout their lives.

Acknowledgments

This research was partly funded by a Duke Bass Connections grant awarded to MLS, RC, and FDB.

Conflict of Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Aquino, K., & Reed, A. II (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effect models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Berntsen, D. (2010). The unbidden past: Involuntary autobiographical memories as a basic mode of remembering. *Current Directions in Psychological Science*, *19*(3), 138–142.

Boisgontier, M. P., & Cheval, B. (2016). The ANOVA to mixed model transition. *Neuroscience & Biobehavioral Reviews*, *68*, 1004–1005.

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

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology*, *67*, 135–157.

Byrne, R. M. (2017). Counterfactual thinking: From logic to morality. *Current Directions in Psychological Science*, *26*(4), 314–322.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, *47*(4), 1178–1198.

Cushman, F. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass*, *10*(2), 117–133.

Davis, C. G., Lehman, D. R., Wortman, C. B., Silver, R. C., & Thompson, S. C. (1995). The undoing of traumatic life events. *Personality and Social Psychology Bulletin*, *21*(2), 109–124.

De Brigard, F., & Giovanello, K. S. (2012). Influence of outcome valence in the subjective experience of episodic past, future, and counterfactual thinking. *Consciousness and Cognition*, *21*(3), 1085–1096.

De Brigard, F., & Parikh, N. (2019). Episodic counterfactual thinking. *Current Directions in Psychological Science*, *28*(1), 59–66.

Dyczewski, E. A., & Markman, K. D. (2012). General attainability beliefs moderate the motivational effects of counterfactual thinking. *Journal of Experimental Social Psychology*, *48*(5), 1217–1220.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, *12*(2), 168–192.

Escobedo, J. R., & Adolphs, R. (2010). Becoming a better person: Temporal remoteness biases autobiographical memories for moral events. *Emotion*, *10*, 511–518.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*, 1340–1343.

Huang, S., Stanley, M. L., & De Brigard, F. (2020). The phenomenology of remembering our moral transgressions. *Memory & Cognition*, *48*, 277–286.

Johannessen, K. B., & Berntsen, D. (2010). Current concerns in involuntary and voluntary autobiographical memories. *Consciousness and Cognition*, *19*(4), 847–860.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136.

Kouchaki, M., & Gino, F. (2016). Memories of unethical actions become obfuscated over time. *Proceedings of the National Academy of Sciences*, *113*, 6166–6171.

Kray, L. J., George, L. G., Liljenquist, K. A., Galinsky, A. D., Tetlock, P. E., & Roese, N. J. (2010). From what might have been to what must have been: Counterfactual thinking creates meaning. *Journal of Personality and Social Psychology*, *98*(1), 106–118.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Marie Hall, N., & Berntsen, D. (2008). The effect of emotional stress on involuntary and voluntary conscious memories. *Memory*, *16*(1), 48–57.

Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1993). The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology*, *29*(1), 87–109.

Markman, K. D., & McMullen, M. N. (2003). A reflection and evaluation model of comparative thinking. *Personality and Social Psychology Review*, *7*(3), 244–267.

Markman, K. D., McMullen, M. N., & Elizaga, R. A. (2008). Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology*, *44*(2), 421–428.

McCulloch, K. C., & Smallman, R. (2014). The implications of counterfactual mind-sets for the functioning of implementation intentions. *Motivation and Emotion*, *38*(5), 635–644.

Morris, M. W., & Moore, P. C. (2000). The lessons we (don't) learn: Counterfactual thinking and organizational accountability after a close call. *Administrative Science Quarterly*, *45*(4), 737–765.

Nasco, S. A., & Marsh, K. L. (1999). Gaining control through counterfactual thinking. *Personality and Social Psychology Bulletin*, *25*(5), 557–569.

Niedenthal, P. M., Tangney, J. P., & Gavanski, I. (1994). " If only I weren't" versus" If only I hadn't": Distinguishing shame and guilt in counterfactual thinking. *Journal of Personality and Social Psychology*, *67*(4), 585–595.

Özbek, M., Bohn, A., & Berntsen, D. (2018). Why do I think and talk about it? Perceived functions and phenomenology of episodic counterfactual thinking compared with remembering and future thinking. *Quarterly Journal of Experimental Psychology*, *71*(10), 2101–2114.

Pillemer, D. B. (1998). *Momentous events, vivid memories*. Cambridge, MA: Harvard University Press.

Pillemer, D. B. (2001). Momentous events and the life story. *Review of General Psychology*, *5*, 123–134.

Pillemer, D. B. (2003). Directive functions of autobiographical memory: The guiding power of the specific episode. *Memory*, *11*(2), 193–202.

Pratt, M. W., Arnold, M. L., & Mackey, K. (2001). Adolescents' representations of the parent voice in stories of personal turning points. In D. P. McAdams, R. Josselson, & A. Lieblich (Eds.), (Eds), *Turns in the road: Narrative studies of lives in transition* (pp. 227–252). Washington, DC: American Psychological Association.

Reczek, R. W., Irwin, J. R., Zane, D. M., & Ehrich, K. R. (2017). That's not how I remember it: Willfully ignorant memory for ethical product attribute information. *Journal of Consumer Research*, *45*(1), 185–207.

Rim, S., & Summerville, A. (2014). How far to the road not taken? The effect of psychological distance on counterfactual direction. *Personality and Social Psychology Bulletin*, *40*(3), 391–401.

Roese, N. J. (1994). The functional basis of counterfactual thinking. *Journal of Personality and Social Psychology*, *66*(5), 805–818.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133.

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In J. M. Olson (Ed.), *Advances in experimental social psychology* (Vol. *56*, pp. 1–79). New York, NY: Academic Press.

Roese, N. J., Park, S., Smallman, R., & Gibson, C. (2008). Schizophrenia involves impairment in the activation of intentions by counterfactual thinking. *Schizophrenia Research*, *103*(1-3), 343–344.

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544.

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory*, *117*, 14–21.

Sheeran, P. (2002). Intention–behaviour relations: A conceptual and empirical review. *European Review of Social Psychology*, *12*, 1–36.

Sheeran, P., & Webb, T. L. (2016). The intention–behavior gap. *Social and Personality Psychology Compass*, *10*(9), 503–518.

Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin*, *37*(3), 330–349.

Smallman, R. (2013). It's what's inside that counts: The role of counterfactual content in intention formation. *Journal of Experimental Social Psychology*, *49*(5), 842–851.

Smallman, R., & McCulloch, K. C. (2012). Learning from yesterday's mistakes to fix tomorrow's problems: When functional counterfactual thinking and psychological distance collide. *European Journal of Social Psychology*, *42*(3), 383–390.

Smallman, R., & Roese, N. J. (2009). Counterfactual thinking facilitates behavioral intentions. *Journal of Experimental Social Psychology*, *45*(4), 845–852.

Stanley, M. L., Bedrov, A., Cabeza, R., & De Brigard, F. (2020). The centrality of remembered moral and immoral actions in constructing personal identity. *Memory*, *28*(2), 278–284.

Stanley, M. L., & De Brigard, F. (2019). Moral memories and the belief in the good self. *Current Directions in Psychological Science*, *28*, 387–391.

Stanley, M. L., Henne, P., & De Brigard, F. (2019). Remembering moral and immoral actions in constructing the self. *Memory & Cognition*, *47*(3), 441–454.

Stanley, M. L., Henne, P., Iyengar, V., Sinnott-Armstrong, W., & De Brigard, F. (2017). I'm not the person I used to be: The self and autobiographical memories of immoral actions. *Journal of Experimental Psychology: General*, *146*(6), 884–895.

Stanley, M. L., Marsh, E. J., & Kay, A. C. (2020). Structure-seeking as a psychological antecedent of beliefs about morality. *Journal of Experimental Psychology: General*, *149*, 1908–1918.

Stanley, M. L., Yang, B. W., & De Brigard, F. (2018). No evidence for unethical amnesia for imagined actions: A failed replication and extension. *Memory & Cognition*, *46*, 787–795.

Stanley, M. L., Yin, S., & Sinnott-Armstrong, W. (2019). A reason-based explanation for moral dumbfounding. *Judgment and Decision Making*, *14*(2), 120–129.

Tangney, J. P. (1995). Recent advances in the empirical study of shame and guilt. *American Behavioral Scientist*, *38*(8), 1132–1145.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*, 5.

Van Dijk, E., & Zeelenberg, M. (2005). On the psychology of 'if only': Regret and the comparison between factual and counterfactual outcomes. *Organizational Behavior and Human Decision Processes*, *97*(2), 152–160.

Zeelenberg, M., van Dijk, W. W., Van der Pligt, J., Manstead, A. S., Van Empelen, P., & Reinderman, D. (1998). Emotional reactions to the outcomes of decisions: The role of counterfactual thought in the experience of regret and disappointment. *Organizational Behavior and Human Decision Processes*, *75*(2), 117–141.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information