The Complementary Roles of Memory, Morality, and Counterfactual Thinking in Constructing and Improving the Self

by

Matthew Lawrence Stanley

Department of Psychology & Neuroscience
Duke University

Date:_______________________
Approved:

___________________________
Elizabeth J. Marsh, Advisor

___________________________
Felipe De Brigard, Advisor

___________________________
Roberto Cabeza, Advisor

___________________________
Walter Sinnott-Armstrong

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology & Neuroscience in the Graduate School of Duke University

2021
ABSTRACT

The Complementary Roles of Memory, Morality, and Counterfactual Thinking in Constructing and Improving the Self

by

Matthew Lawrence Stanley

Department of Psychology & Neuroscience
Duke University

Date:________________________

Approved:

___________________________

Elizabeth J. Marsh, Advisor

___________________________

Felipe De Brigard, Advisor

___________________________

Roberto Cabeza, Advisor

___________________________

Walter Sinnott-Armstrong

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology & Neuroscience in the Graduate School of Duke University

2021
Abstract

Despite the ubiquity of wrongdoing in everyday life, the vast majority of people believe that they are truly morally good. Across 15 studies that employ a combination of correlational, experimental, and mediation designs, this dissertation first examines the role of memory in constructing, protecting, and maintaining a morally good self-concept, and then it investigates whether the ways in which our moral transgressions are remembered and mutated can play a role in learning from past mistakes. In two initial studies, autobiographical memories of people’s especially morally good past actions were particularly central to constructions of personal identity. In three subsequent studies, a “knew-it-all-along effect” after acting dishonestly offered a way for people to explain away those past improprieties that could have presented a threat to a morally good self-concept. Then, across seven additional studies, past wrongdoings were attributed to a distant, dissimilar past self who had changed considerably over time for the better. The results of three final studies indicate that remembering, reflecting on, and mutating past events can serve a directive function by strengthening intentions for future moral improvement.
Contents

Abstract .......................................................................................................................................................... iv

List of Tables ..................................................................................................................................................... xi

List of Figures .................................................................................................................................................. xii

Acknowledgements ......................................................................................................................................... xiv

1. General introduction ................................................................................................................................... 1

2. The centrality of remembered moral and immoral actions in constructing personal identity ....................... 6

   2.1 Introduction ............................................................................................................................................. 6

   2.2 Study 1 .................................................................................................................................................. 10

      2.2.1 Materials and method ..................................................................................................................... 10

         2.2.1.1 Participants ................................................................................................................................. 10

         2.2.1.2 Materials ................................................................................................................................... 11

         2.2.1.3 Procedure .................................................................................................................................. 11

         2.2.1.4 Statistical analyses ................................................................................................................... 12

      2.2.2 Results and discussion ................................................................................................................... 13

   2.3 Study 2 .................................................................................................................................................. 13

      2.3.1 Materials and method ..................................................................................................................... 14

         2.3.1.1 Participants ................................................................................................................................. 14

         2.3.1.2 Materials ................................................................................................................................... 14

         2.3.1.3 Procedure .................................................................................................................................. 15

         2.3.1.4 Statistical analyses ................................................................................................................... 16
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.1.2 Procedure</td>
<td>73</td>
</tr>
<tr>
<td>4.4.2 Results and discussion</td>
<td>75</td>
</tr>
<tr>
<td>4.5 General discussion</td>
<td>80</td>
</tr>
<tr>
<td>5. Remembering moral and immoral actions in constructing the self</td>
<td>87</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>87</td>
</tr>
<tr>
<td>5.2 Study 1</td>
<td>92</td>
</tr>
<tr>
<td>5.2.1 Materials and method</td>
<td>92</td>
</tr>
<tr>
<td>5.2.1.1 Participants</td>
<td>92</td>
</tr>
<tr>
<td>5.2.1.2 Procedure</td>
<td>93</td>
</tr>
<tr>
<td>5.2.1.3 Statistical analyses</td>
<td>94</td>
</tr>
<tr>
<td>5.2.2 Results and discussion</td>
<td>95</td>
</tr>
<tr>
<td>5.3 Study 2</td>
<td>98</td>
</tr>
<tr>
<td>5.3.1 Materials and method</td>
<td>99</td>
</tr>
<tr>
<td>5.3.1.1 Participants</td>
<td>99</td>
</tr>
<tr>
<td>5.3.1.2 Procedure</td>
<td>99</td>
</tr>
<tr>
<td>5.3.1.3 Statistical analyses</td>
<td>101</td>
</tr>
<tr>
<td>5.3.2 Results and discussion</td>
<td>101</td>
</tr>
<tr>
<td>5.4 Study 3</td>
<td>105</td>
</tr>
<tr>
<td>5.4.1 Materials and method</td>
<td>106</td>
</tr>
<tr>
<td>5.4.1.1 Participants</td>
<td>106</td>
</tr>
<tr>
<td>5.4.1.2 Procedure</td>
<td>107</td>
</tr>
<tr>
<td>5.4.1.3 Data analyses</td>
<td>108</td>
</tr>
</tbody>
</table>
5.4.2 Results and discussion........................................................................108

5.5 Study 4 .................................................................................................112

5.5.1 Materials and method........................................................................113
  5.5.1.1 Participants ..................................................................................113
  5.5.1.2 Procedure....................................................................................114
  5.5.1.3 Statistical analyses .....................................................................116

5.5.2 Results and discussion........................................................................116

5.6 General discussion..................................................................................120

6. The roles of memory and counterfactual simulation for past wrongdoings in fostering moral learning .................................................................126

6.1 Introduction............................................................................................126

6.2 Study 1 .................................................................................................133

  6.2.1 Materials and method......................................................................134
    6.2.1.1 Participants ..............................................................................134
    6.2.1.2 Procedure.................................................................................134
    6.2.1.3 Statistical analyses ..................................................................137

  6.2.2 Results and discussion.....................................................................138

6.3 Study 2 .................................................................................................143

  6.3.1 Materials and method......................................................................143
    6.3.1.1 Participants ..............................................................................143
    6.3.1.2 Procedure.................................................................................144
    6.3.1.3 Statistical analyses ..................................................................146
List of Tables

Table 1: CES means, standard deviations, and reliabilities are depicted.......................14

Table 2: Results are depicted from the linear mixed-effects model.........................18

Table 3: Examples of memories of lying in actor and recipient conditions.................58

Table 4: Overall means for morality, valence, and emotional intensity ratings split by actor and recipient conditions.................................................................62

Table 5: Examples of memories of emotional harm from both actor and recipient conditions...........................................................................................................66

Table 6: Summary of means (SDs) for each measured variable as a function of memory cue.........................................................................................................................102

Table 7: Results of three linear mixed-effects models with cued morality and the person doing the action as predictors of different, same, and change judgments, respectively.109

Table 8: Means and standard deviations in Study 1.................................................138

Table 9: Means (SDs) for intention judgments split by condition and intention-type....161
List of Figures

Figure 1: Boxplots for the proportion of questions correctly answered (across only the 12 difficult questions) in Study 2. .........................................................................................................................35

Figure 2: Boxplots for average knew-it-all-long ratings (across the 12 difficult questions) in Study 2. .........................................................................................................................37

Figure 3: Boxplots for the proportion of questions correctly answered (across only the 12 difficult questions) in Study 3. .........................................................................................................................43

Figure 4: Boxplots for average knew-it-all-long ratings (across the 12 difficult questions) in Study 3. .........................................................................................................................46

Figure 5: For lying (a-c) and emotional harm (d-f), frequency statistics are depicted for ratings of morality, valence, and emotional intensity. .................................................................61

Figure 6: Frequency statistics for morality, valence, and emotional intensity ratings for memories of lying (a) and emotional harm (b) in both actor and recipient conditions. ...77

Figure 7: For memories of lying, frequency statistics are displayed for ratings of morality, valence, and emotional intensity. ........................................................................................................78

Figure 8: For memories of emotional harm, we display frequency statistics for ratings of morality, valence, and emotional intensity. .................................................................................................80

Figure 9: Means and standard error bars are depicted for the judged morality of the remembered actions, different judgments, same judgments, and change judgments......96

Figure 10: For the self (a) and other (b) conditions, means and SE bars are depicted for the four outcome variables. ..........................................................................................................................111

Figure 11: Means and standard error bars are depicted for valence (a) and emotional intensity ratings (b) for remembered transgressions. .........................................................................................148

Figure 12: Means and standard error bars are depicted for the frequency of voluntary (a) and involuntary (b) recall for remembered transgressions.................................................................150

Figure 13: Means and standard error bars are depicted for the frequency of morally upward counterfactual thinking (a) and intention strength (b). .................................................................151
Figure 14: Separate mediation models are depicted for remembered transgressions involving dishonesty (a), harm (b), and unfairness (c).
Acknowledgements

I have been fortunate enough to work with so many brilliant and good people across several different departments and at all levels of university life: undergraduate students, lab managers, staff, fellow graduate students, post-docs, and faculty. I have learned so much from these relationships and collaborations, and I have undoubtedly become a better thinker, scientist, and person because of them. Outside of the more immediate Duke University community, I am beyond grateful to my parents, Mike and Ann, and my wife, Ashley, for their unwavering support—especially when things looked bleak and uncertain. Also, my dog is pretty good.
1. General introduction

Almost everyone cares deeply about possessing positive moral traits and qualities, and most people believe that positive moral traits and qualities (e.g., honesty, fairness, compassion) play an integral role in defining who they are and who they wish to be (Aquino & Reed, 2002; Stanley & De Brigard, 2019; Strohminger et al., 2017). Prentice and colleagues (2019) have even suggested people *need* the feeling or experience of being morally good, and that satisfying this moral need leads to positive psychological outcomes (e.g., well-being, flourishing). In some cases, committing and remembering our moral transgressions can lead to serious negative health outcomes, including anxiety, depression, self-harm, and posttraumatic stress disorder (Griffin et al., 2019). And yet, people lie, cheat, steal, harm, and treat others unfairly with surprising frequency in everyday life (Escobedo & Adolphs, 2010; Hofmann et al., 2014; Stanley et al., 2017; Stanley & De Brigard, 2019). How is it, then, that most people care deeply about being morally good and have convinced themselves that they truly are morally good, all while committing moral transgressions so frequently?

I argue that several distinct but complementary mechanisms involving memory help people to protect and maintain their beliefs that they are morally good and that they possess positive moral traits and qualities. The process of searching memory is rarely (if ever) impartial or objective. Although our memories are not completely disconnected from reality, they are still malleable, error-prone, and influenced by our
biases and motivations (Conway, 2005; De Brigard, 2014; Kunda, 1990; Schacter, 1999). Memory search, reconstruction, and interpretation can be used in ways to mitigate the damage caused by our moral transgressions in order to enable a morally good self-concept in the present.

Chapters 2-5 in this dissertation focus on the role of memory in constructing, protecting, and maintaining a morally good self-concept. Utilizing a well-validated measure—the Centrality of Event Scale (Berntsen & Rubin, 2006)—Chapter 2 presents evidence that autobiographical memories of our especially morally good past actions, relative to memories of our moral transgressions, are particularly central to constructions of personal identity (Stanley, Bedrov, Cabeza, & De Brigard, 2020). Autobiographical memory refers to a complex set of mental processes that involve recollecting events and experiences from our personal pasts (Rubin, 1988), and the “critical defining feature for autobiographical memory is the importance of the information to one’s sense of self and one’s life history” (Roediger & Marsh, 2003: 485). By coming to believe that our past morally good actions are particularly central to our life stories and personal identities, we can come to view ourselves as morally good and as the possessors of positive moral traits.

Chapter 3 focuses on a metacognitive effect for a particular kind of moral transgression—cheating. For these kinds of transgressions, people who cheat tend to believe that they just knew the answers all along (Stanley, Stone, & Marsh, 2021). A
“knew-it-all-along effect” after cheating offers one way for people to explain away their past improprieties that might pose a threat to a morally good self-concept. Cheating might not be as distasteful (or might be justified) if people believe they already knew the answers. In this sense, cheating might be considered accidental or unnecessary, and thus, not a real threat to a morally good self-concept.

Chapters 4 and 5 then offer complementary evidence for a more general way in which people can construct, protect, and maintain a morally good self-concept in the present. That is, people advantageously utilize their memories in social comparison and psychological distancing to foster a sense of positive moral change within a life story; in doing so, they can facilitate a belief in their own moral superiority and a belief in their own personal moral improvement over time. Memories of our more serious past improprieties can be attributed to a distant, dissimilar past self who has changed considerably over time for the better (Stanley, Henne, & De Brigard, 2019; Stanley, Henne, Iyengar, Sinnott-Armstrong, & De Brigard, 2017). People can dismiss or explain away their more serious past moral transgressions (e.g., “My most serious transgressions were committed by a dissimilar self, and I do not act like that anymore”) and portray themselves favorably in the present. Whereas recalling our past moral transgressions can be met perceptions of meaningful and positive transformation in the self over time, our remembered morally praiseworthy actions can be attributed to a stable, unchanging self.
Whereas Chapters 2-5 address ways in which a morally good self-concept is constructed, protected, and maintained—despite the human propensity for frequent wrongdoing—Chapter 6 addresses a memory-facilitated means by which people can nevertheless learn from their moral mistakes. That is, Chapter 5 offers evidence for a directive function of memory and counterfactual thinking in service of strengthening future intentions to act in morally better ways (Stanley, Cabeza, Smallman, & De Brigard, under review). Even if people psychologically distance themselves from their particularly serious past transgressions, they can still remember those transgressions and frequently recall them. Psychological distancing does not preclude moral learning, intention formation, or moral improvement facilitated via memory processes. People ruminate upon their past moral transgressions, particularly those transgressions they judge to be more serious, and they frequently simulate morally better ways in which they could have acted instead (i.e., morally upward counterfactuals). The more people report having mentally simulated morally better ways in which they could have acted, the stronger their intentions are to improve in the future. Making accessible a morally upward counterfactual after committing a moral transgression also strengthens intentions for moral improvement. So, the motivation to construct, protect, and maintain morally good self-concepts does not entail that our past moral transgressions will not be frequently recalled, ruminated upon, and mutated to facilitate learning and intention
formation for moral improvement. People are not necessarily destined to repeatedly commit the same moral transgressions throughout their lives.
2. The centrality of remembered moral and immoral actions in constructing personal identity

2.1 Introduction

There is a venerable philosophical view according to which memory underwrites personal identity through time (Locke, 1690/2009). More recently, this view has influenced psychological research where perceived personal identity has been linked to autobiographical memory (Conway, 2005; Conway & Pleydell-Pearce, 2000; Schacter, 2008), a complex set of mental processes that involve recollecting information, events, and experiences from our personal pasts that are of relevance to the self (Rubin, 1988).2 Remembering our own unique past experiences helps us to distinguish ourselves from others and to construct a sense of self over time (Conway 2005; Conway & Pleydell-Pearce, 2000; Conway, Singer, & Tagini, 2004; McAdams, 2013; Rubin, 1988). But, presumably, some memories of past events are more important than others in constructing personal identity. Here, we argue for, and offer evidence in support of, the claim that our memories of our most morally right actions—relative to less morally right actions and immoral actions—play a more central role in constructing personal identity.

2 For our purposes, “personal identity” refers to a first-person subjective judgment about what makes a person think that they are the same or different over time. We do not use “personal identity” to refer to a metaphysical claim about what makes a person the same or different over time, or an epistemic claim about whether or not a person is justified in claiming to be the same person over time.
Most people believe they are morally good, care deeply about possessing morally good traits and qualities, and believe that positive moral traits and qualities play an integral role in defining who they are and who they wish to be (Aquino & Reed, 2002; Goodwin, 2015; Monin & Jordan, 2009; Strohminger et al., 2017). In fact, recent evidence suggests that people tend to believe that moral traits and qualities are more central, or essential, to personal identity than other features of our mental lives (Strohminger et al., 2017; Strohminger & Nichols, 2015). For instance, people are most likely to report that a particular individual was no longer the same person after exhibiting changes in moral qualities relative to changes in perception, preferences, emotions, and even memories (Molouki & Bartels, 2017; Strohminger & Nichols, 2014, 2015). In particular, morally good qualities like honesty and compassion are perceived to be more central to the self than morally bad qualities (Strohminger et al., 2017). Removing a particular person’s morally good qualities produces greater impressions of identity change relative to morally bad qualities (Strohminger et al., 2017).

We suggest that memories of specific past events in which we acted in a morally good way could still play a particularly important role in how we construct our personal identities over time. Accordingly, memories would still play an essential role in constructing and fostering personal identity, even if moral traits and qualities are most central to conceptions of personal identity. Autobiographical memories serve many different functions (Bluck & Alea, 2011; Bluck, Alea, Habermas, & Rubin, 2005), and one
of those functions is to define who we believe ourselves to be (Hyman & Faries, 1992; McLean & Pasupathi, 2006; McLean & Thorne, 2003). To serve this self-definitional function, memories are often mentally linked together and organized into a life narrative (Bluck & Habermas, 2000). When recalling and interpreting the significance of past events, our especially morally right actions may be particularly important for defining who we are and who we wish to be, and for constructing these life narratives. By coming to believe that our past morally good behaviors are particularly central to our life stories and personal identities, we may come to view ourselves as morally good and as the possessors of positive moral traits (e.g., honesty, fairness, compassion, kindness).

Despite the common belief that they are fundamentally and essentially morally good, people still behave immorally with surprising frequency (Hofmann et al., 2014), and they can readily recall committing many different moral transgressions (Stanley & De Brigard, 2019). Some of these transgressions they even judge to be severely morally wrong (Stanley et al., 2017, 2019). By coming to believe that our past morally good actions are more central to and representative of who we truly are relative to our past immoral actions, we may come to believe that we are indeed morally good—even though we frequently transgress.

To investigate the centrality of remembered moral and immoral actions in constructing personal identity, we utilized the Centrality of Event Scale (CES), originally developed to measure the extent to which a traumatic memory represents a reference
point for personal identity (Berntsen & Rubin, 2006). The CES includes items that ask participants about the extent to which specific past events have become anchors for their personal identities, reference points for the ways they understand themselves, and central parts of their life stories. The CES is not specific to negative or traumatic memories, however, and the measure has also been used to assess the centrality of remembered emotionally positive events (Zaragoza, Scherman, Salgado, Shao, & Berntsen, 2015), imagined future events (Rubin, Berntsen, Deffler, & Brodar, 2019), and vicarious memories (Pillemer, Steiner, Kuwabara, Thomsen, & Svob, 2015). The more central an event is, the more integrated it is into a person’s identity and life story, guiding both past and future interpretations of personal experiences (Berntsen & Rubin, 2006, 2007; Rubin, Boals, & Berntsen, 2008). Nevertheless, the CES has yet to be used to investigate the centrality of events that involve morality.

Our overarching goal is to investigate the centrality of remembered events involving morally right and wrong actions in constructing personal identity. Study 1 investigates whether, for morally right actions only, remembered actions judged to be more morally right relative to less morally right are more central to personal identity. Study 2 investigates whether remembered morally right actions are more central to personal identity than remembered morally wrong actions.
2.2 Study 1

In Study 1, participants were asked to recall five of their own past morally right actions. For each memory, participants indicated just how morally right they believed the action was, and they completed the CES. We hypothesized that remembered past actions judged to be more morally right (relative to less morally right) will be more central to personal identity and to the life story.

2.2.1 Materials and method

2.2.1.1 Participants

Sixty-five American residents voluntarily participated in this study via Amazon’s Mechanical Turk (AMT) for monetary compensation. Participant recruitment was restricted to those who had at least 50 previously accepted HITs and a prior approval rating above 90%. Four participants were excluded for not answering all questions or for failing the attention check at the end. As such, data were analyzed with the remaining 61 participants ($M_{\text{age}} = 35 \text{ years}, SD = 9$, age range = [22, 65], 25 females, 36 males). All participants reported being fluent English speakers. For all 15 studies reported in this dissertation, informed consent was obtained from each participant in accordance with protocols approved by the Duke University Campus Institutional Review Board.
2.2.1.2 Materials

The Centrality of Event Scale (CES; Berntsen & Rubin, 2006, 2007) was devised to measure the extent to which a memory forms a central component of personal identity, a turning point in the life story, and a reference point for everyday inferences. The CES consists of 20 items rated on 5-point scales (1 = totally disagree; 5 = totally agree). The scale only has a single underlying factor (Berntsen & Rubin, 2006) and was reliable (α = .98).3

2.2.1.3 Procedure

The study was a single, self-paced session. Participants were asked to recall five distinct past events, one at a time, in which they did something morally right. To help participants recall these events, they were told that the remembered morally right actions might involve, but are not limited to, helping, honesty, loyalty, and fairness. Participants were also instructed that each remembered event must be specific in time and place. For each remembered event, participants briefly described the event in 2-5 sentences and indicated how morally right their action was on a scale from 1 (slightly morally right) to 5 (very morally right); after making the moral judgment, participants completed the CES.

3 Because each participant generated five memories, we also randomly selected one memory from each participant to compute an α value. Then, of the four remaining memories from each participant, we randomly selected one more memory from each of them to compute another α value. We then repeated this for the remaining memories generated by each participant. These five different α values ranged from .96 to .98.
After making 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 assured that their responses would not affect their payment or eligibility for future studies. Only those participants who selected ‘5’ were included in the analyses (see exclusions above). This same attention check has been used in recent published research (e.g., Stanley, Marsh, & Kay, 2020; Stanley, Yin, & Sinnott-Armstrong, 2019). Upon completion, participants were monetarily compensated for their time.

2.2.1.4 Statistical analyses

Data were analyzed using R (R Development Core Team, 2009) with the ‘lme4’ software package (Bates, Maechler, Bolker, & Walker, 2015) and with the ‘lmerTest’ software package (Kuznetsova, Brockhoff, & Christensen, 2017). Data were fit to linear mixed-effects models (LMEM), and subject was included as a random effect (both random intercepts and slopes were included) in all models. 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.
2.2.2 Results and discussion

The described events covered diverse actions, including: returning lost possessions (e.g., wallets, rings) to their owners, volunteering at animal shelters, giving food and money to the homeless, giving extra money back to cashiers, donating to charity, caring for sick friends and family, etc. The mean value of the CES was 2.56 ($SD = .98$), and the mean value of moral rightness ratings was 4.44 ($SD = .45$). A linear mixed-effects model revealed that the more morally right the remembered past action was judged to be, the more central the event was perceived to be ($b = .29$, $SE = .07$, $t = 3.97$, $p < .001$, 95% CI = [.15, .43]). This result corroborates our hypothesis that remembered past actions judged to be more morally right (relative to less morally right) are also perceived as more central to personal identity and to the life story.

2.3 Study 2

The primary purpose of Study 2 is to directly compare the centrality of the most morally right and the most morally wrong past actions participants can remember. We hypothesized that the most morally right remembered actions will be more central than the most morally wrong remembered actions. We also attempt to replicate the basic effect obtained from Study 1: that more morally right remembered actions tend to be more central than less morally right remembered actions. Finally, we also statistically account for possible differences in emotional arousal associated with the memories. To
this end, participants provide ratings of emotional arousal, and we use these ratings as statistical controls in LMER models.

2.3.1 Materials and method

2.3.1.1 Participants

Two-hundred twenty American residents voluntarily participated in this study via AMT for monetary compensation. Participant recruitment was restricted to those with least 50 previously accepted HITs and a prior approval rating above 90%. Twenty participants were excluded for not answering all questions or for failing the attention check. As such, data were analyzed with the remaining 200 participants ($M_{\text{age}} = 35$ years, $SD = 11$, age range = [21, 72], 72 females, 128 males). All participants reported being fluent English speakers.

2.3.1.2 Materials

The 20-item CES used in Study 1 was also used in Study 2 (see Table 1 for reliabilities).

Table 1: CES means, standard deviations, and reliabilities are depicted.

<table>
<thead>
<tr>
<th>CES, most morally right</th>
<th>Mean</th>
<th>SD</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CES, slightly morally right</td>
<td>1.93</td>
<td>.88</td>
<td>.97</td>
</tr>
<tr>
<td>CES, slightly morally wrong</td>
<td>1.80</td>
<td>.84</td>
<td>.97</td>
</tr>
<tr>
<td>CES, most morally wrong</td>
<td>2.64</td>
<td>1.05</td>
<td>.97</td>
</tr>
</tbody>
</table>
2.3.1.3 Procedure

The study was a single, self-paced session. Participants were asked to recall four distinct events, one at a time, from their personal pasts. They were instructed that each remembered event must be specific in time and place. Participants were provided with a unique cue for each of the four memories: (1) recall the most morally right thing you have done in your past, (2) recall the most morally wrong thing you have done in your past, (3) recall a slightly morally wrong thing you have done in your past, and (4) recall a slightly morally right thing you have done in your past. This yielded a 2 (moral valence: morally right versus morally wrong) x 2 (moral intensity: most morally right/wrong versus slightly morally right/wrong) within-subjects design. The order in which the four different memory cues were presented was randomized across participants. In cases in which participants were instructed to describe their morally right remembered behaviors, we explicitly instructed them to recall morally right behaviors—regardless of the cued moral intensity on the particular trial (i.e., regardless of whether participants were asked to provide a slightly or very morally right remembered behavior). And in cases in which participants were instructed to describe their morally wrong remembered behaviors, we explicitly instructed them to recall morally wrong behaviors—regardless of the cued moral intensity on the particular trial (i.e., regardless of whether participants were asked to provide a slightly or very morally wrong remembered behavior). To help participants recall events, they were told that
remembered morally right actions might involve, but are not limited to, helping, honesty, loyalty, and fairness, and that morally wrong actions might involve, but are not limited to, harm, dishonesty, disloyalty, and unfairness.

For each remembered event, participants briefly described the event in 2-5 sentences and made a rating for emotional arousal associated with each memory (1 = *not at all emotionally intense*, 5 = *very emotionally intense*). As a manipulation check, for the two morally right remembered actions, participants indicated how morally right the actions were on a scale from 1 (*slightly morally right*) to 5 (*very morally right*); for the two morally wrong remembered actions, participants indicated how morally wrong the actions were on a scale from 1 (*slightly morally wrong*) to 5 (*very morally wrong*). Then, participants completed the CES.

Participants then answered the same attention check question that was in Study 1. We excluded participants who reported being distracted, having trouble paying attention, failing to avoid distractions, and not taking the survey seriously (see exclusions above). Upon completion, participants were monetarily compensated for their time.

2.3.1.4 Statistical analyses

The same statistical software used in Study 1 was also used in Study 2.
2.3.2 Results and discussion

Means, standard deviations, and reliabilities are presented in Table 1. The described morally right behaviors were similar in kind to those in Study 1, and they included: returning lost possessions (e.g., wallets, rings) to their owners, rescuing injured animals, giving food, clothing, and money to the homeless, giving extra money back to cashiers, donating to charity, giving money to friend in financial trouble, caring for sick friends and family, etc. The described morally wrong behaviors were also diverse, including: physically harming other people, cheating on significant others, breaking promises to friends and family, shoplifting, stealing money, cheating on college exams, etc.

An initial LMEM was computed to verify that remembered actions generated from the most morally right cue were, in fact, rated as more morally right than remembered actions generated from the slightly morally right cue. This expectation was confirmed ($b = 1.42, SE = .10, t = 14.43, p < .001, 95\% CI = [1.22, 1.61]$). Similarly, a second LMEM was computed to verify that remembered actions generated from the most morally wrong cue were, in fact, rated as more morally wrong than remembered actions generated from the slightly morally wrong cue. This expectation was also confirmed ($b = 1.94, SE = .10, t = 18.51, p < .001, 95\% CI = [1.72, 2.14]$).

We next computed a LMEM that included moral valence (morally right coded as 0 versus morally wrong coded as 1) and moral intensity (slightly morally right/wrong
coded as 0 versus most morally right/wrong coded as 1) as fixed factors, and event centrality served as the outcome variable. There was no significant effect of moral valence ($b = -.13$, $SE = .07$, $t = -1.90$, $p = .058$, 95% CI = [-.27, .00]), but there was a significant effect of moral intensity ($b = 1.07$, $SE = .07$, $t = 15.01$, $p < .001$, 95% CI = [.93, 1.21]), suggesting that the most morally right/wrong remembered actions are associated with higher CES scores than slightly morally right/wrong remembered actions. However, these effects were qualified by a significant interaction between moral valence and moral intensity ($b = -.22$, $SE = .10$, $t = -2.24$, $p = .026$, 95% CI = [-.43, -.03]). A similar pattern of effects was obtained after statistically controlling for emotional arousal ratings associated with each memory, although the main effect of moral valence reached statistical significance after including emotional arousal in the model (see Table 2).

**Table 2: Results are depicted from the linear mixed-effects model.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>$b$</th>
<th>$SE$</th>
<th>$t$</th>
<th>$p$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moral Valence</td>
<td>-.16</td>
<td>.06</td>
<td>-2.52</td>
<td>.012</td>
<td>[-.28, -.04]</td>
</tr>
<tr>
<td>Moral Intensity</td>
<td>.41</td>
<td>.08</td>
<td>5.41</td>
<td>&lt; .001</td>
<td>[.26, .57]</td>
</tr>
<tr>
<td>Emotional Arousal</td>
<td>.37</td>
<td>.02</td>
<td>16.22</td>
<td>&lt; .001</td>
<td>[.32, .41]</td>
</tr>
<tr>
<td>Moral Valence x Moral Intensity</td>
<td>-.19</td>
<td>.09</td>
<td>-2.08</td>
<td>.037</td>
<td>[-.38, -.01]</td>
</tr>
</tbody>
</table>

*Note. The outcome variable is event centrality. All 95% CIs are for the beta-estimates.*

To further interrogate the interaction effect, we first computed separate follow-up LMEMs for remembered morally right and morally wrong actions, separately. The primary purpose of these follow-up tests was to replicate the effect in Study 1: that the more morally right a remembered behavior is judged to be (only among past behaviors judged to be morally right at least to some degree), the more central it is to personal
identity and the life story. So, we first compared CES scores for slightly morally right versus the most morally right remembered actions. Then, we compared CES scores for slightly morally wrong versus the most morally wrong remembered actions. For remembered morally right actions, the most morally right remembered actions were more central than slightly morally right remembered actions ($b = 1.07, SE = .07, t = 14.29, p < .001, 95\% \text{ CI} = [.93, 1.21]$). This effect held after statistically controlling for the emotional arousal associated with each remembered event ($b = .29, SE = .09, t = 3.18, p = .002, 95\% \text{ CI} = [.10, .48]$). These results corroborate our hypothesis that remembered past actions judged to be more morally right (relative to less morally right) are also judged as more central to personal identity and to the life story. For the sake of completeness, we also investigated the effect of moral intensity for remembered morally wrong behaviors only. We found that the most morally wrong remembered actions were more central than slightly morally wrong remembered actions ($b = .84, SE = .07, t = 11.86, p < .001, 95\% \text{ CI} = [.71, .99]$). This effect held after statistically controlling for the emotional arousal associated with each remembered event ($b = .18, SE = .08, t = 2.14, p = .033, 95\% \text{ CI} = [.01, .35]$).

Next, we computed separate follow-up LMEMs within the different moral intensities. That is, we first compared CES scores for participants’ most morally right versus most morally wrong remembered actions. This addresses the primary hypothesis in Study 2: that participants’ most morally right remembered actions will be more
central to personal identity and the life story than their most morally wrong remembered actions. Then, for the sake of completeness, we compared CES scores for the slightly morally right versus the slightly morally wrong remembered actions. Corroborating our primary hypothesis, the results indicate that the most morally right remembered actions were more central than the most morally wrong remembered actions ($b = -.36, SE = .08, t = -4.60, p < .001, 95\% CI = [-.51, -.21])$. This effect held after statistically controlling for the emotional arousal associated with each remembered event ($b = -.35, SE = .07, t = -4.68, p < .001, 95\% CI = [-.50, -.20]$). Second, slightly morally right remembered actions were more central than slightly morally wrong remembered actions ($b = -.13, SE = .05, t = -2.96, p = .003, 95\% CI = [-.23, -.04]$). This effect held after statistically controlling for the emotional arousal associated with each remembered event ($b = -.16, SE = .04, t = -3.80, p < .001, 95\% CI = [-.24, -.08]$).

### 2.4 General discussion

In two studies, we investigated the role of remembered past events involving moral and immoral actions in constructing personal identity. For morally right actions, we found that remembered actions judged to be more morally right relative to less morally right were more central to personal identity (Study 1). We then found that remembered morally right actions were more central to personal identity than remembered morally wrong actions—even after statistically controlling for emotional arousal (Study 2). Taking these studies together, specific memories of our especially
morally good actions are more likely to help in defining who we believe ourselves to be than specific memories of our less morally good actions and our moral transgressions. These findings provide some insights into longstanding questions about memory, morality, and personal identity.

Accumulating empirical evidence suggests that people tend to believe that moral traits and qualities are more central, or essential, to personal identity than other features of our mental lives, such as preferences, desires, emotions, and even memories (Strohminger et al., 2017; De Freitas et al., 2017). Some have argued that the belief in a “good true self” is a form of psychological essentialism, or “the basic cognitive tendency to assume that all entities have deep, unobservable, inherent properties that comprise their true nature” (De Freitas et al., 2017, p. 634; see also, De Freitas et al., 2018; De Freitas & Cikara, 2018; Strohminger et al., 2017). We suggest that theorists can readily accept this theoretical position regarding the morally good true self while still maintaining that memory plays an important role in constructing and fostering personal identity. Remembering, reflecting on, and interpreting the significance of past events may serve as a complementary means of fostering a morally good sense of self. Our memories of past events provide the concrete evidence for our beliefs about our moral goodness.

The fact that participants rated their morally praiseworthy deeds as more central than their moral transgressions may also reflect a self-enhancement or self-protection
bias. People tend to remember the personal past in a way that facilitates the construction of a favorable view of the self in the present (Alicke & Sedikides, 2009; Gregg, Sedikides, & Gebauer, 2011; Sedikides & Gregg, 2008; Wilson & Ross, 2001, 2003). These self-enhancement and self-protection motives play a role in how we remember our moral transgressions and how we come to interpret their significance (Stanley & De Brigard, 2019). For instance, when remembering immoral actions from their personal pasts, people judge their own transgressions to be less morally wrong than those of others, and they perceive personal moral improvement over time (Stanley et al., 2017, 2019; Stanley & De Brigard, 2019). Judging past morally praiseworthy actions as more central to the life story and personal identity than immoral actions may foster a favorable view of the self as morally good, helping people to self-enhance or self-protect.

Our experimental design and research questions raise a question about the relationship between emotional valence and moral valence. We suggest that it should not be assumed a priori that emotional valence and moral valence are capturing the same variance. Most of us have had emotionally positive experiences during which we did something morally wrong. To provide a specific example, a participant described an event in which they took a tip at work that they were not allowed to take, and it was described as a positive experience; nevertheless, the participant believed that the action was morally wrong. Moreover, most of us have had an emotionally negative experience during which we did something morally right. For example, a participant described
lending a helping hand to a friend who was repairing his damaged roof on a hot summer day. This experience might have been difficult and negative, but the helping behavior was reported to be morally right. In practice, moral and emotional valence might be correlated, but there are many cases in which moral and emotional valence come apart. Future research will more systematically explore cases in which moral and emotional valence come apart, and cases in which they do not.

Our sample comprised American residents recruited on AMT, and we included no individual difference measures in our studies. Future work will address individual differences and cross-cultural differences in understanding the relationship between remembered moral actions and personal identity. For instance, the particular kinds of remembered morally right behaviors that are most important for constructions of personal identity may vary across cultures and individuals. Research has shown that there are at least two clusters of moral values, or foundations: “individualizing” values of caring and fairness, and “binding” values of loyalty, obedience, and purity (Graham, Haidt & Nosek, 2009; Graham et al., 2011). Because liberals tend to care more about individualizing values than conservatives (Graham et al., 2009), remembered behaviors involving individualizing values may be most important for constructing personal identity for liberals than for conservatives.
3. Cheaters claim they knew the answers all along

3.1 Introduction

Cheating scandals have become commonplace in recent years. For example, in 2012, roughly 125 Harvard students were investigated for cheating in a government class; in 2014, 92 Air Force Officers were suspended for cheating on a missile exam; in 2019, parents allegedly paid for the correction of their children’s SAT test answers (Operation Varsity Blues). These high-profile cases are unusual in that cheating was caught. The vast majority of cases likely go uncovered and unnoticed, given that some surveys show that more than ninety percent of undergraduates admit to such behavior (Evans & Craig, 1990; McCabe & Trevino, 1997; McCabe, Trevino, & Butterfield, 2001; McKibban & Burdsal, 2013; Stuber-McEwen, Wiseley, & Hoggatt, 2009; Williams, Tanner, Beard, & Hale, 2012).

Despite the pervasiveness of cheating, people are strongly motivated to enhance and protect their favorable self-concepts, believing that they truly exemplify positive traits and virtues like honesty, trustworthiness, and integrity (Alicke & Sedikides, 2009; Aquino & Reed, 2002; Stanley & De Brigard, 2019; Stanley, Henne, & De Brigard, 2019; Wojciszke, 2005). The pervasiveness of cheating stands in stark contrast to people’s widespread beliefs that they embody positive traits and virtues like honesty,

---

trustworthiness, and integrity. We offer evidence in favor of an explanation for why people can cheat without damaging their favorable self-concepts. This explanation involves systematic alterations in how past events are recollected and re-evaluated in light of what has transpired since the events occurred.

Converging lines of evidence suggest that people often have difficulty retrospectively determining what they knew prior to acquiring new information (e.g., Fischhoff, 1975; Hasher, Attig, & Alba, 1981; Wood, 1978; Hawkins & Hastie, 1990; Roese & Vohs, 2012). In seminal research on the hindsight bias (Fischhoff, 1975), for example, participants read historical scenarios and answered multiple-choice questions about several possible outcomes. After being informed of the actual outcomes, participants were asked to indicate which outcome they would have guessed, if they had not been provided the correct answer. Participants were consistently biased toward reporting they would have selected the correct answer all along (see also, Guilbault, Bryant, Brockway, & Posavac, 2004). Similar effects occur across a variety of domains, including: economic decisions, political elections, and sports events (Christensen-Szalanski & Willham, 1991; Guilbault, Bryant, Brockway, & Posavac, 2004; Hawkins & Hastie, 1990). Building off early work on the hindsight bias, other research has found that, after attempting to answer general knowledge questions and then receiving the correct answers to the questions, people often come to believe that they just knew the correct
answers all along (i.e., a “knew-it-all-along effect”; Fischhoff, 1977; Arnold & Lindsay, 2007; Hasher et al., 1981; Jacoby & Kelley, 1987; Metcalfe & Finn, 2011; Wood, 1978).

Returning to our focus on cheating, we suggest that a knew-it-all-along effect might offer one way for people to explain away their cheating behavior in a manner that protects their favorable self-concepts. That is, cheating might not be as distasteful (and might even be justified) if people believe that they already knew the answers. In this sense, cheating might be considered accidental or unnecessary, and thus, not a real threat to a favorable self-concept. Furthermore, to the extent that this illusion is protective of one’s identity, the knew-it-all-along effect accompanying cheating may be larger than would be expected if people simply received the answers without cheating (as feedback, for example).

In three studies we tested the hypothesis that cheating is associated with amplified estimates of prior knowledge. These studies depended on two procedural choices. First, we conducted extensive pre-testing to identify questions from geography, history, science, and sports for which no participants knew the correct answers. The goal was to identify items for which participants likely cheated if they answered them correctly. Second, cheating was made possible in some conditions by providing the correct answer upside-down at the bottom of the screen, in the same way that magazine quizzes often provide answers for self-scoring. In Study 1, we found that participants reported higher estimates of prior knowledge for questions they correctly answered than
for questions they did not—even though no participants correctly answered any of the questions in our pre-tests (this effect replicated in Studies 2 and 3). Studies 2 and 3 included control conditions in which participants did not have the opportunity to cheat but were exposed to the correct answers before making judgments about whether they knew the correct answers prior to the experimental session. The control conditions were included to demonstrate that cheating inflated estimates of prior knowledge, beyond what was expected from a general knew-it-all-along effect. In both Studies 2 and 3, we found that participants assigned to the experimental condition who had the opportunity to cheat exhibited a larger knew-it-all-along effect relative to participants in the control conditions who were exposed to the correct answers but had no opportunity to cheat in the study.

3.2 Study 1

3.2.1 Materials and method

3.2.1.1 Participants

A total of 150 American residents with at least 50 completed HITs and an approval rating above 90% voluntarily participated in this study on Amazon’s Mechanical Turk (AMT) for monetary compensation. Three participants were excluded from analyses for failing at least one of the two attention checks (see below for details); so, data were analyzed with the remaining 147 individuals ($M_{age} = 36.13$ years, $SD = 9.89$, age range = [19-63], 80 males, 66 females). To determine our sample size in Study 1, we
sought to match or exceed the sample sizes of seminal research on the knew-it-all-along effect (Hasher et al., 1981; Wood, 1978; Hawkins & Hastie, 1990). In all studies reported herein, data were analyzed only after the required sample size target was met, and the sample size was determined prior to data collection in each study (see pre-registrations). In Studies 1 and 2, we recruited participants through AMT to obtain more representative samples of the United States population than traditional convenience samples obtained through undergraduate participant pools. In Study 3, we collected a nationally representative sample through Lucid.

3.2.1.2 Materials

Stimuli consisted of 24 general knowledge questions that systematically differed as a function of difficulty (some of these stimuli have been used in published work from Tauber et al. (2013) and from Wang et al. (2016)). The questions came from several domains, including: geography, history, science, and sports.

Pre-testing was conducted to ensure that participants on AMT were unlikely to know the answers to 12 target questions (i.e., the difficult questions) and were likely to know the answers to 12 filler questions (i.e., the easy questions). To this end, a separate sample of 100 participants (5 participants were excluded for not answering all questions) were asked to produce answers to a large set of questions. We adapted instructions from a large-scale norming study on general knowledge statements (Tauber et al., 2013). Briefly, participants were instructed to answer general knowledge questions and to
search memory for the correct answer. Participants were also told that there would be no penalty for guessing. For example, the following question was presented to participants:
What was the first genomically sequenced fish? The correct response to this question is the pufferfish. As in Tauber et al. (2013), these responses were scored with leniency for misspellings. No participants correctly answered any of the 12 difficult questions (e.g., “Which nation was the first to ratify the United Nations charter in 1945?”), and at least 70% of participants correctly recalled the answers to each of the 12 easy questions (e.g., “How many legs does a spider have?”). All 24 questions and answers are provided in Appendix A. The easy questions were included only to help conceal the aims of the study; only the difficult questions were used to answer our research questions.

3.2.1.3 Procedure

After providing informed consent, participants were told that they would be presented with many different general knowledge questions of varying difficulty, and they were instructed to answer each question by typing it in. These 24 questions were presented one at a time in a randomized order; this comprised the testing phase of the study. Prior to beginning the testing phase, participants were told that the correct answer to each question would be presented in small print (size 10 font, Times New Roman) and upside-down in the bottom-right corner of the screen. When answering the questions, participants were explicitly instructed not to look at the correct answers on the screen. They were told that it would be cheating to look at the correct answers on the
screen. Participants were also explicitly told that it would be cheating to use outside resources (for example: other people, the internet, or books). We tested participants on these instructions with the following true/false question: When answering the trivia questions, it is cheating to look at the correct answers on the screen or to use outside resources (for example: other people, the internet, or books). Participants who failed this first attention check (by responding “false”) were excluded from our analyses. This attention check was meant to ensure that the included participants understood it would be cheating to look at the correct answers on the screen or to use outside resources.

After attempting to answer all 24 questions, participants completed a brief, unrelated, distractor task. In a subsequent block, participants were then presented with the same 24 questions seen in the testing phase of the experiment, one at a time in a randomized order. This time, the correct answer was presented with each question. In this phase, participants reported whether they knew the answer to each question prior to the beginning of the study (1 = this was new to me, 7 = I actually knew this all along). This prior knowledge measure was adapted from Metcalfe and Finn (2011). Participants were instructed to answer each question honestly, told that their responses would be confidential, and assured that there would be no negative repercussions no matter how they responded.

At the end, participants completed an attention check: “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 didn’t take the study seriously; (4) no, something else affected my participation negatively; or (5) yes. We assured participants that their responses would not affect payment or eligibility for future studies. Only those participants who selected (5) were included in the analyses. This same attention check question has been used in published research (Stanley, Marsh, & Kay, 2020; Stanley, Yin, & Sinnott-Armstrong, 2019). Upon completion, participants were monetarily compensated for their time.

3.2.1.4 Statistical analyses

Data were analyzed using R (R Development Core Team, 2009) with the ‘lme4’ software package (Bates, Maechler, Bolker, & Walker, 2014) and with the ‘lmerTest’ software package (Kuznetsova, Brockhoff, & Christensen, 2017). Data were fit to a linear mixed-effects model (LMEM) with the ‘optimix’ optimizer. Significance for fixed effects was assessed using Satterthwaite approximations to degrees of freedom, and 95% confidence intervals around beta-values were computed using bootstrapping (n simulations = 1000).

3.2.2 Results and discussion

Overall, participants correctly answered 14% of the difficult questions, suggesting that participants cheated on roughly 14% of the questions given that the expected answer rate was zero based on our pre-testing. We tested our hypothesis that, for those cases in which participants likely cheated to answer the question (i.e.,
answered correctly even though no participants in the pre-test answered correctly), they would be more likely to report that they actually knew the answers all along (prior to beginning the study). To this end, we computed a LMEM with the knew-it-all-along judgments as the outcome variable. Whether or not participants reported the correct answer to each of the difficult questions was modeled as a binary fixed factor (yes or no). Participant and item were included as crossed random effects (random intercepts only). For questions for which participants provided the correct answer, relative to questions for which participants did not provide the correct answer, participants were more likely to report that they knew the answer all along ($b = 3.20$, $SE = .10$, $t = 31.53$, $p < .001$, 95% CI = [2.99, 3.42]). As an exploratory analysis, the same pattern of findings was obtained when we computed the proportion of questions on which each participant cheated and the average of the knew-it-all-along judgments for each participant, and then correlated these resultant values: $r = .69$, $p < .001$. Average knew-it-all-along judgments were 4.40 for cases in which participants reported the correct answer, and average knew-it-all-along judgments were 1.28 for cases in which participants did not provide the correct answer.

### 3.3 Study 2

Study 1 provides correlational evidence consistent with our hypothesis that, for those particular cases in which participants likely cheated (i.e., provided the correct answer to questions that no participants answered correctly in the pre-test), they were
more likely to report that they actually knew the answers prior to beginning the study.

In Study 2, we offer experimental evidence to show that this effect is not just a typical knew-it-all-long effect. To this end, we instituted a control condition in which participants received feedback on their answers but were not given the opportunity to cheat. Knew-it-all-long judgments made by participants in this control condition were compared to those made by participants in the critical experimental condition in which they had the opportunity to cheat.

3.3.1 Materials and method

3.3.1.1 Participants

A total of 150 American residents with at least 50 completed HITs and an approval rating above 90% voluntarily participated in this study on Amazon’s Mechanical Turk (AMT) for monetary compensation. No participants failed either of the attention checks, so data were analyzed with all 150 individuals ($M_{age} = 35$ years, $SD = 11$, age range $= [18-70]$, 79 males, 69 females).

3.3.1.2 Materials

The 24 questions (12 easy, 12 difficult) from Study 1 were also used in Study 2.

3.3.1.3 Procedure

After providing informed consent, participants were randomly assigned to one of two conditions in a between-subjects fashion. In the experimental condition (the cheating condition), the procedure was identical to the procedure experienced by all
participants in Study 1. The critical differences between the experimental and control conditions were in the testing phase. Participants in the cheating condition were provided with the correct answers to the questions in small print and upside-down in the bottom-right corner of the screen (as in Study 1), but participants in the control condition were not presented with the correct answers on the same screen as the questions. Instead, participants in the control condition attempted to answer each question, and then after submitting their response, they were shown the correct answer to the question on the following page. Thus, participants in the control condition had no opportunity to cheat on any of the questions, but they received similar (or even more) exposure to the answers as did the participants in the cheating condition. Otherwise, the control condition was identical to the cheating condition.

3.3.1.4 Statistical analyses

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

3.3.2 Results and discussion

First, a non-parametric Independent-Samples Mann-Whitney U Test revealed that participants in the cheating condition ($n = 73$, $Mean = .17$, $SD = .31$, $Median, = .00$, IQR = .17) answered more of the difficult questions correctly than the participants in the control condition ($n = 77$, $Mean = .02$, $SD = .04$, $Median, = .00$, IQR = .00; $p < .001$). This
suggests that more participants cheated in the cheating condition than in the control condition. Figure 1 depicts these results.

![Boxplots for the proportion of questions correctly answered (across only the 12 difficult questions) in Study 2.](image)

Next, we attempted to replicate our findings from Study 1. To this end, for only those participants assigned to the cheating condition, we computed a LMEM with knew-

---

2 Based on these descriptive statistics, it is worth noting that participants in control condition almost never answered the difficult questions correctly. This accords with what we expected based on pre-testing. These descriptive statistics further suggest that participants in the control did not cheat by searching for the answers on the internet. We suspect this is because searching for the answers on the internet requires time and effort, and far more time and effort than just glancing at the bottom of the screen as in the cheating condition. As such, our control condition is sufficient to prevent cheating, at least in the vast majority of cases.
it-all-along judgments as the outcome variable. Whether or not participants reported the correct answer to each of the difficult questions was modeled as a binary fixed factor (yes or no). For questions for which participants provided the correct answer, relative to questions for which participants did not provide the correct answer, participants were more likely to report that they knew the answer all along ($b = 2.96$, $SE = .15$, $t = 19.83$, $p < .001$, $95\% CI = [2.66, 3.25]$). As an exploratory analysis, the same pattern of findings was obtained when we computed the proportion of questions participants got correct (i.e., likely cheated) and the average of the knew-it-all-along judgments for each participant, and then correlated the resultant values: $r = .80$, $p < .001$. Average knew-it-all-along judgments were 4.54 for cases in which participants provided the correct answer, and average knew-it-all-along judgments were 1.30 for cases in which participants did not provide the correct answer.

Finally, we tested our second hypothesis that participants in the cheating condition would be more likely to report that they knew the answers all along (prior to beginning the study)—relative to participants in the control condition. To this end, we computed a LMEM with condition (cheating vs. control) as a binary fixed factor and with knew-it-all-along judgments as the outcome variable. Participant and item were included as crossed random effects in the model. The results indicate that participants in the cheating condition, relative to participants in the control condition, reported higher knew-it-all-along judgments ($b = .63$, $SE = .17$, $t = 3.76$, $p < .001$, $95\% CI = [.33, .95]$).
Average knew-it-all-along judgments were 1.89 for participants in the cheating condition, and average knew-it-all-along judgments were 1.25 for participants in the control condition. Figure 2 graphically depicts these results.

**Figure 2: Boxplots for average knew-it-all-along ratings (across the 12 difficult questions) in Study 2.**

### 3.4 Study 3

Study 3 further investigates whether cheating amplifies the knew-it-all-along effect while accounting for two possible alternative explanations that could have produced the observed effects in Study 2. One possibility is that cheaters in Study 2 consciously lied when making their prior knowledge judgments, so that the
experimenters would not detect their earlier cheating and subject them to negative repercussions (e.g., not paying them, blocking them from participating in future studies). To address this possibility, we explicitly told participants in Study 3 that their estimates of prior knowledge would not affect their payment or their eligibility for future studies. Participants were tested to ensure they understood this instruction, and the analyses included only those participants who reported understanding that their responses would not have any negative repercussions. To preview, even after these exclusions, we still found that participants in the cheating condition exhibited higher knew-it-all-along judgments than participants in the control condition.

A second possibility is that a difficulty in source monitoring drove the effects in Study 2, as opposed to the motivational account that we hypothesized. That is, estimating one’s prior knowledge requires one to discriminate what (if anything) one learned in the experiment from what one knew beforehand. The design of Study 2 may have made this decision more difficult for participants in the cheating condition than for participants in the control condition. In the cheating condition, the answers appeared on the same screen as the general knowledge questions, and cheating involved typing in the answer as when answering the easy filler questions without cheating. In contrast, the control condition used a traditional feedback design, with participants first attempting to answer the question followed by answer feedback on a separate screen. This means that the correct answer was slightly delayed in time, and when participants did not
know the answer, they did not type it in. While it would still be interesting if cheating produces conditions that lead to difficult source decisions, this would be a cognitive explanation rather than a motivational explanation. To address this concern, Study 3 equated the difficulty of the source judgments across the cheating and control conditions, to rule out that explanation for our results. To preview, we still found that participants in the cheating condition exhibited higher knew-it-all-along judgments than participants in the control condition, consistent with our hypothesized motivational account.

3.4.1 Materials and method

3.4.1.1 Participants

A total of 603 American residents were recruited through Lucid. For a recent analysis of the Lucid platform for participant recruitment for experimental research, see Coppock and McClellan (2019). We aimed to recruit 600 participants through Lucid with the expectation that we would end up with roughly 2.5 times as many participants as in Study 2, after exclusions. One-hundred ninety-one participants failed at least one attention check, so data were analyzed with 412 individuals ($M_{age} = 49$ years, $SD = 17$, age range = [18-82], 167 males, 235 females). The number of participants who failed each individual attention check are indicated below (note that some participants failed multiple attention checks, so adding up number of participants who failed each
attention check will result in a number greater than the number of participants recruited).

### 3.4.1.2 Materials

The 24 questions (12 easy, 12 difficult) from the previous two studies were also used in Study 3 (see Appendix A for all questions and answers).

### 3.4.1.3 Procedure

After providing informed consent, participants were randomly assigned, in a between-subjects fashion, to the cheating condition or to the control condition. The cheating condition in Study 3 was the same as in Study 2. However, the control condition in Study 3 was changed so that control participants also saw the correct answers at the bottom of the screen—but critically, control participants were told they could look at the answer at the bottom of the screen if they could not generate the answer themselves. Thus, participants in both conditions saw exactly the same screens; what differed was whether they were told that looking at the answers was allowed vs. prohibited. Participants in both conditions were explicitly told that it would be cheating to use outside resources (for example: other people, the internet, or books).

We tested participants on these instructions. Participants in the cheating condition were presented with the following true/false question: When answering the trivia questions, it is cheating to look at the correct answers on the screen or to use outside resources (for example: other people, the internet, or books). Participants in the
cheating condition who failed this attention check (by responding “false”) were excluded from our analyses (14 participants failed this attention check). Participants in the control condition were presented with the following true/false question: When answering the trivia questions, it is not cheating to look at the correct answers on the screen, but it is cheating to use outside resources (for example: other people, the internet, or books). Participants in the control condition who failed this attention check (by responding “false”) were excluded from our analyses (41 participants failed this attention check). Participants in both conditions received feedback about their responses. These attention checks were meant to ensure that the included participants understood what did and did not constitute cheating in the study. These attention checks were again presented immediately after participants finished the testing phase of the study. Again, participants who failed these attention checks were excluded from our analyses (19 participants in the cheating condition failed the attention check, and 19 participants in the control condition failed the attention check). And again, participants received feedback about their responses.

Participants then completed a brief, unrelated distractor task. Then, participants were instructed that they would be presented with the same 24 questions seen earlier in the experiment, one at a time in a randomized order, and that each question would be paired with its correct answer. In this phase, and as in the previous studies, participants were instructed to report whether they knew the answer to each question prior to the
beginning of the study (1 = *this was new to me*, 7 = *I actually knew this all along*). We explicitly told participants that how they answered the questions would not affect their payment or eligibility for future studies, that there would be no negative repercussions regardless of how they answered, and that their responses would be confidential. Two attention check questions were presented before participants made any prior knowledge judgments. First, participants were presented with the following true/false question: In this next part, I am indicating whether I knew the answer to each question before I started this study. Participants were then presented with this second true/false question: Your answers to the following questions will not affect your payment or eligibility for future studies. No matter how you answer, we will not penalize you and you will be able to participate in the future studies from our lab. Participants who failed either of these attention checks (by responding “false”) were excluded from our analyses (32 participants failed the first attention check, and 36 participants failed the second attention check). Participants received feedback about their responses to the attention check questions.

After making prior knowledge judgments for all 24 questions, participants answered demographics questions and were presented with the final attention check question that we provided in the previous studies. The 63 participants who failed this final attention check were also excluded from our analyses.
3.4.1.4 Statistical analyses

The same statistical software and packages used the previous studies were also used in Study 3.

3.4.2 Results and discussion

Descriptively, the average proportion of difficult questions answered correctly by participants in the cheating condition was .22 (n = 225, SD = .35, Median = .00, IQR = .29), and the average proportion of difficult questions answered correctly by participants in the control condition was .79 (n = 187, SD = .34, Median = 1.00, IQR = .25). See Figure 3 for a graphical representation of these results.

Figure 3: Boxplots for the proportion of questions correctly answered (across only the 12 difficult questions) in Study 3.
We first attempted to replicate our finding from Studies 1 and 2 that participants reported higher estimates of prior knowledge for questions they correctly answered than for questions they did not—even though no participants correctly answered any of the questions in our pre-tests. To this end, for only those participants assigned to the cheating condition, we computed a LMEM with knew-it-all-along judgments as the outcome variable. Whether or not participants reported the correct answer to each of the difficult questions was modeled as a binary fixed factor (yes or no). For questions for which participants provided the correct answer, relative to questions for which participants did not provide the correct answer, participants were more likely to report that they knew the answer all along ($b = 2.13$, $SE = .09$, $t = 24.01$, $p < .001$, 95% CI = [1.96, 2.32]). As an exploratory analysis, for participants in the cheating condition, the same pattern of findings was obtained when we computed the proportion of questions participants got correct (i.e., likely cheated) and the average of the knew-it-all-along judgments for each participant, and then correlated the resultant values: $r = .70$, $p < .001$. For participants in the cheating condition, average knew-it-all-along judgments were 4.79 for cases in which participants provided the correct answer, and average knew-it-all-along judgments were 1.65 for cases in which participants did not provide the correct answer. In contrast, for participants in the control condition, there was no statistically significant relationship between the proportion of questions participants got correct and average knew-it-all-along judgments ($r = .06$, $p = .43$). For participants in the control
condition, average knew-it-all-along judgments were 2.02 for cases in which participants provided the correct answer, and average knew-it-all-along judgments were 1.53 for cases in which participants did not provide the correct answer.

Next, we tested our second hypothesis that participants in the cheating condition would be more likely to report that they knew the answers all along (prior to beginning the study) — relative to participants in the control condition. To this end, we computed a LMEM with condition (cheating vs. control) as a binary fixed factor and with knew-it-all-along judgments as the outcome variable. Participant and item were included as crossed random effects in the model. The results indicate that participants in the cheating condition, relative to participants in the control condition, reported higher knew-it-all-along judgments ($b = .43$, $SE = .16$, $t = 2.77$, $p = .006$, 95% CI = [.13, .74]). Average knew-it-all-along judgments were 2.35 for participants in the cheating condition, and average knew-it-all-along judgments were 1.91 for participants in the control condition. Figure 4 graphically depicts these results.
Overall, the results from Study 3 support our hypothesis that cheating amplifies the knew-it-all-along effect. The control condition instituted in Study 3 eliminated one possible alternative explanation for our results: by using the exact same procedures across conditions (other than the instructions about cheating) we equated the two conditions in source monitoring difficulty. Furthermore, our extensive instructions and the new instructional attention check in Study 3 should help to alleviate the concern that cheaters consciously lied when making prior knowledge judgments to prevent the experimenters from subjecting them to negative repercussions (e.g., not paying them,
blocking them from participating in future studies). Of course, we cannot entirely rule out the possibility that cheaters consciously lied about their prior knowledge to save face with the experimenter – but we believe it is unlikely that participants were concerned with a nameless experimenter’s opinion of them in a brief online study, especially since they knew their compensation was safe.

3.5 General discussion

In three studies, we investigated whether cheating on a general knowledge test is associated with increased belief of having known the answers prior to the experimental session. In all three studies, we found correlational evidence that, for those particular cases in which participants provided the correct answer, they were more likely to report that they actually knew the answers all along (prior to beginning the study). In Studies 2 and 3, we found that this effect is not simply a typical knew-it-all-along effect; participants in control conditions who received the correct answers, but were not given the same opportunity to cheat, showed a significantly smaller knew-it-all-along effect than participants in the experimental condition who had the opportunity to cheat. These experimental findings are particularly surprising given that participants only cheated on a minority of questions in the cheating condition; this minority was enough to obtain higher average knew-it-all-along judgments in the cheating condition relative to the control condition in which participants were provided with the correct answer to all questions before making knew-it-all-along judgments. Overall, our results are consistent
with the idea that people are motivated to believe that they already knew the answers to
the questions they cheated on. Cheating poses no threat to one’s favorable self-concept if
one already “knew” the information.

These findings contribute to a nascent but growing literature on how memory
can support people’s self-concepts and allow them to believe they are morally good
despite committing moral transgressions frequently and repeatedly. Prior research has
shown that people forget damaging details about their moral transgressions (Kouchaki
& Gino, 2016; Reczek et al., 2017; Shu et al., 2011), and other research suggests that
people recall their transgressions in a way that distances themselves from their more
egregious past transgressions while perceiving moral improvement over time (Stanley et
al., 2017, 2019; Stanley & De Brigard, 2019). Our work shows a metacognitive
contribution. Our findings broaden the scope of the roles that memory and
metacognition play in protecting and fostering a morally good self-concept.

Overall, these results fit well with the larger literature, lending direct evidence to
a mechanism that was postulated (but not explicitly tested) in a paper investigating self-
deception (Chance, Norton, Gino & Ariely, 2011; see also, Chance, Gino, Norton, &
Ariely, 2015). The procedure in that paper was similar to the initial testing phase in our
experiment: Participants took a general knowledge test, with a subset of participants
able to view the answer key while taking the test (although looking at the answer key
was not explicitly labeled as cheating). Participants then predicted their future
performance on a similar test. Critically, participants who had been given the opportunity to use an answer key on the initial test made higher estimates for their performance on the subsequent, similar test. This work did not directly test whether a knew-it-all-along effect drove the inflated estimates, but the authors posited it as a possible explanation for their results (Chance et al., 2011).

Our findings also provide insight into the relationship between the knew-it-all-along effect, hindsight bias, and the self. Considerable research has shown that the hindsight bias is reduced or eliminated following negative (as opposed to positive) outcomes that are of direct relevance to the self (Hölzl, Kirchler, & Rodler, 2002; Louie, 1999; Louie, Curren, & Harich, 2000; Mark, Boburka, Eyssell, Cohen, & Mellor, 2003; Mark & Mellor, 1991; Pezzo, 2011; Pezzo & Pezzo, 2007). For example, when an individual is laid off from a job, they might believe that they could not have possibly seen it coming (Mark & Mellor, 1991). After a negative self-relevant outcome, people who strategically come to believe that they could not have foreseen that outcome occurring can absolve themselves from blame and wrongdoing in a way that protects their favorable self-concepts (Pezzo, 2011; Roese & Vohs, 2012). In contrast, our results show an increase in knew-it-all-along judgments as a function of threat to a favorable self-concept (i.e., having cheated relative to not cheating). This might suggest that, at least in some cases, we can boost the knew-it-all-along effect to absolve ourselves of blame and wrongdoing in a way that protects our favorable self-concepts.
Cheating behavior has become commonplace and widespread (Williams, Tanner, Beard, & Hale, 2012; McKibban & Burdsal, 2013). Committing these transgressions can be deleterious to individuals, groups, universities, and society-at-large, so it is important to better understand why people so frequently commit them. Our findings provide evidence for a new mechanism that might help in understanding why people can so frequency commit these transgressions, even while they continue to believe they are morally good, honest, trustworthy, and of high integrity.
4. I’m not the person I used to be: The self and autobiographical memories of immoral actions¹

4.1 Introduction

People maintain a positive identity by evaluating themselves in more favorable terms than they evaluate other people (Alicke & Sedikides, 2009) and by judging themselves to be better now than they were in the past (Wilson & Ross, 2003). Rapidly accumulating evidence also suggests that morality is essential for the construction and perception of one’s identity over time (Strohminger, Newman, & Knobe, 2017).

Nevertheless, few studies have integrated these diverse research programs to explore how autobiographical memories of moral and immoral actions shape one’s identity. To fill this gap in the literature, we investigated how individuals, when recalling personal, autobiographical memories of actions of moral relevance, compare themselves to other people and to their past selves.

Disparate lines of research have produced convergent evidence suggesting that people evaluate themselves in more positive terms than they evaluate others (Alicke & Sedikides, 2009; Taylor & Brown, 1988, 1994). In particular, biases and distortions in the way people remember the past may help to maximize the positivity and minimize the negativity of their self-assessments relative to others (Conway, Singer, & Tagini, 2004;

D’Argembeau & Van der Linden, 2008). People are more likely to recall and vividly re-experience positive information about themselves than they are to recall such information about others (D’Argembeau & Van der Linden, 2008; Rogers, Kuiper, & Kirker, 1977). Complementary evidence has shown that people believe they are more virtuous, intelligent, talented, and compassionate than the average person (Alicke & Govorun, 2005; Alicke & Seikides, 2009; Batson & Collins, 2011). Critically, compared to beliefs in other domains such as intelligence, competence, or ambition, desirable moral traits (e.g., honesty) are associated with the largest difference between judgments of self and the average person (Alicke, Dunning, & Krueger, 2005; Alicke, Vredenburg, Hiatt, & Govorun, 2001; Tappin & McKay, 2017). Thus, a sense of moral superiority obtained through evaluating oneself in more favorable terms than others is a particularly strong bias that helps people maintain a positive personal identity (Tappin & McKay, 2017).

People not only compare themselves with others to maintain a positive view of self; they also compare their present selves with their past selves. Some evidence even suggests that people more frequently compare themselves with their past selves than with other people (Wilson & Ross, 2000). At least one theory, temporal self-appraisal theory, explains why we compare our present selves with past selves. According to this theory, assessments of the subjective distance between past experiences and the present facilitates the self-enhancement and self-protection functions of autobiographical memory (Ross & Wilson, 2002; Wilson & Ross, 2001, 2003). People often compare
themselves to past selves because they perceive themselves as improving over time, regardless of whether the perceived improvement is accurate (Ross & Wilson, 2000; Ryff, 1991). Notably, evaluations of former selves, compared to the current self, become increasingly unfavorable over time, and people are most critical of subjectively distant past selves on the traits they believe are most important (Wilson & Ross, 2001, 2003). However, this work on temporal self-appraisal theory has not investigated the role of the perceived morality or immorality of past actions. Comparisons with perceived inferior past selves—specifically through autobiographical recollections of immoral actions in the distant past—may be particularly important for the perception of self-improvement and the maintenance of a positive identity.

Despite the importance of morality and memory for the construction and perception of one’s identity over time (Strohminger, Knobe, & Newman, 2017), autobiographical memories of moral events have not been extensively studied. Some evidence suggests that memories of immoral actions are particularly susceptible to biases and distortions (Kouchaki & Gino, 2016; Pizzaro, Laney, Morris, & Loftus, 2006). Of particular relevance to the present investigation, Escobedo and Adolphs (2010) identified a pronounced bias in recollections of both moral and immoral memories by discovering that memories of events that occurred in the nearer past (nearer memories) tend to be rated as less negative than memories of events that occurred in the more distant past (distant memories). This finding accords with work showing that
autobiographical recollections often serve a self-enhancement function (Demiray & Janssen, 2015; Wilson & Ross, 2003). It remains unclear, however, whether less negative memories with moral content correspond to nearer events for both actions committed by oneself and those committed by others to oneself. Relatedly, it is unclear whether more distant memories are judged to be more morally wrong than nearer memories. Additionally, it is unclear whether certain kinds of moral transgressions disproportionately produced the effects obtained by Escobedo and Adolphs (2010).

Finally, because Escobedo and Adolphs (2010) only assessed the relationship between the affective qualities of memories and the time that they occurred in the past, it remains unclear whether these effects still hold when accounting for subjective, psychological distance from current selves. The current paper investigates these unanswered questions.

To this end, we distinguish between actor and recipient autobiographical memories. Actor memories concern actions that the participant committed, and recipient memories concern actions for which the participant was the target. Furthermore, recent work has shown that situations involving emotional harm and dishonesty are among the most commonly experienced moral violations (Hofmann, Wisneski, Brandt, & Skitka, 2014), suggesting that individuals should be more likely to recall multiple memories of lying and emotional harm than memories for other moral violations (e.g., sanctity). As
such, Study 1a investigates memories about lying, and Study 1b investigates memories about emotional harm; Study 2 investigates both kinds of memories.

We report results from three studies that collectively build upon prior work showing that memory biases and distortions enable people to evaluate themselves in more positive terms than they evaluate others and their past selves. We investigate whether there are actor-recipient differences in the emotional qualities and perceived moral wrongness of remembered actions involving lying and emotional harm. We also investigate whether the perceived emotional qualities and moral wrongness of these remembered actions differ as a function of temporal distance and of perceived subjective, psychological distance from past selves.

4.2 Study 1a

Study 1 is divided into two parts and investigates the emotional and moral content of autobiographical memories involving lying (Study 1a) and emotional harm (Study 1b) from actor and recipient conditions. We make four specific predictions in Study 1. First, consistent with previous results showing that people are motivated to evaluate themselves in more positive terms than they evaluate others (Alicke & Sedikides, 2009; Taylor & Brown, 1994), we predict that remembered actions in which the participant is the actor (i.e., actor condition) will be rated as less morally wrong and less negative than remembered events in which the participant is the recipient of the action (i.e., recipient condition). Our additional predictions build upon previous
research on temporal self-appraisal theory (Wilson & Ross, 2003), autobiographical memories with moral content (Escobedo & Adolphs, 2010), and the importance of morality in constructing and perceiving the self (Strohminger, Knobe, & Newman, 2017).

Second, we predict that more temporally distant remembered actions will be rated as more negative than temporally nearer remembered actions when the participant is the actor but not the recipient of the action. Third, we predict that more temporally distant remembered actions will be judged to be more morally wrong than temporally nearer remembered actions when the participant is the actor but not the recipient of the action. Given that people compare themselves to their past selves to perceive self-improvement and maintain a positive self-image, actions in more distant memories from the actor condition may be judged as more morally wrong than actions in nearer memories.

Fourth, we predict that the more people believe they have changed since the event occurred, the more likely they are to judge their actions during that time in the past as more morally wrong in the actor condition but not the recipient condition.

4.2.1 Materials and method

4.2.1.1 Participants

A total of 51 adults voluntarily participated in this study. Sample sizes in Study 1a and 1b were determined in order to obtain a similar number of participants and autobiographical memories to those acquired by Escobedo and Adolphs (2010) for their analyses. Three participants were excluded due to failures in following instructions.
Data were analyzed with the remaining 48 participants ($M_{age} = 25$ years, $SD = 3$, age range = [20, 30]; 30 females, 18 males), and each participant was tested individually. All participants were fluent English speakers.

4.2.1.2 Procedure

Participants were seated in a room alone in front of a computer, on which they typed all of their responses. The setting was private, and participants were assured that their responses would be confidential. Each participant was given 30 min to recall as many specific, autobiographical memories as possible in which he/she lied to another person (i.e., the actor condition) and another 30 min to come up with as many memories as possible in which he/she discovered that another person lied to him/her (i.e., the recipient condition). The order in which participants completed actor and recipient conditions was counterbalanced across subjects. Participants were told that all memories must be of events that have occurred within the past 10 years, must be of one particular episode specific in time and place, and must have been personally experienced.

For each memory, participants described the event in 2-6 sentences and listed the time and place of the event. Table 3 provides several examples of memories recalled by participants. To record when the event took place relative to now, participants were provided with the following options from which they were instructed to choose exactly one: within the past day, within the past week, within the past month, within the past year, within the past two years, within the past three years, within the past four years,
within the past five years, within the past six years, within the past seven years, within the past eight years, within the past nine years, within the past ten years. Then, participants provided the following ratings for the actions in each memory in this order: overall, how well do you remember the event? (1 = hardly, 7 = very well); how well do you remember how you felt during the event? (1 = not at all, 7 = very well); how morally right or morally wrong was the action performed? (1 = very morally wrong, 7 = very morally right); what were your emotions associated with the event (i.e., valence)? (1 = very negative, 7 = very positive); what was the intensity of the emotion you felt during the event? (1 = not at all intense, 7 = very intense); and to what extent do you believe you are the same person now compared to the person you were around the time that the remembered event occurred (1 = very similar, 7 very different). Although these data were not analyzed, we included measures of memory vividness, detail, and ease of imagining to further ensure that participants were remembering specific episodes of relatively short duration from the past. We have reported all measures, conditions, and data exclusions. Upon completion of the study, participants were monetarily compensated for their time.

Table 3: Examples of memories of lying in actor and recipient conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Recalled Event</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

58
Actor: I told my (now ex) boyfriend that I had stopped talking to another guy that I had dated in the past. I had actually made a secret, private email account to message him. My boyfriend had a history of going through my phone, so I hid the messages.

Actor: I told a guy I was dating that I turned down an acceptance to a prestigious university even though I had actually never applied in the first place.

Recipient: I was told by my boss that I had been doing a very good job working on one of the experimental projects that I am on. When I talked to the other graduate students in the lab, I found out that she was actually complaining that I was not working fast enough on it.

Recipient: A friend and I were texting about hanging out. She said she was free to come over on a Saturday morning to spend some time together. It turned out that she wasn’t available when I followed up with her later in the week.

4.2.1.3 Statistical analyses

Data were analyzed using R (R Development Core Team, 2009) with the ‘lme4’ software package (Bates, Maechler, Bolker, & Walker, 2015). Data were fit to linear mixed-effects regression (LMER) models with subject always modeled as a random effect. Random slopes and intercepts were included in each model. Fixed effects and outcome variables differed between models depending upon the hypothesis being tested. Significance for models with continuous outcome variables was assessed using the Kenward-Roger corrected $F$-test with the ‘pbkrtest’ package (Halekoh & Hojsgaard, 2014). 95% confidence intervals were computed using parametric bootstrapping. To assess the relationship between temporal distance (i.e., time relative to the present when
the event in the memory occurred) and our other variables of interest within the LMER framework, the time variable was coded as follows: 0 = within the past day, 1 = between the past week, 2 = within the past month, 3 = within the past year, 4 = within the past two years, 5 = within the past three years, 6 = within the past four years, 7 = within the past five years, 8 = within the past six years, 9 = within the past seven years, 10 = within the past eight years, 11 = within the past nine years, 12 = within the past ten years. We created these time bins in such a way to closely resemble the methods from Escobedo and Adolphs (2010).

**4.2.2 Results and discussion**

On average, participants generated 6.13 memories ($SD = 2.19$) in the actor condition and 4.83 memories ($SD = 1.85$) in the recipient condition. Across all participants, a total of 294 memories were generated in the actor condition, and a total of 232 memories were generated in the recipient condition. Figure 5a-c depicts frequency statistics for each variable split by actor and recipient conditions.
Our first prediction was that remembered events in which the participant is the actor will be rated as less morally wrong and less negative than remembered events in which the participant is the recipient of the action. A LMER of condition (binary factor: actor, recipient) on moral wrongness revealed a significant effect of condition ($b = -.789$, $SE = .131$, $F(1, 46.069) = 20.451$, $p < .0001$, 95% CI = [-1.50, -.527]) such that actions in memories in the recipient condition were rated as more morally wrong than those in memories in the actor condition (Figure 5a). A second LMER of condition on valence revealed a significant effect of condition ($b = -.743$, $SE = .122$, $F(1, 46.946) = 18.590$, $p < .0001$, 95% CI = [-1.40, -.507]) such that actions in memories in the recipient condition were rated as more negative than those in memories in the actor condition (Figure 5b).
.0001, 95% CI = [-.978, -.516]) such that memories in the recipient condition were more negative than memories in the actor condition (Figure 5b). A final LMER of condition on emotional intensity revealed a significant effect of condition ($b = 1.343, SE = .161, F(1, 43.813) = 63.310, p < .0001, 95% CI = [.995, 1.649]) such that memories in the recipient condition were more emotionally intense than memories in the actor condition (Figure 5c). Even after controlling for differences in how well each event was remembered and how well participants remembered how they felt during each event, condition was still significantly related to moral wrongness ($p < .0001), valence ($p < .0001) and emotional intensity ($p < .0001). Table 4 depicts overall means for moral wrongness, valence, and emotional intensity ratings split by actor and recipient conditions. Supporting our first prediction, these results suggest that remembered actions that involve lying are perceived as more morally wrong, more negative, and more emotionally intense in the recipient condition than in the actor condition.

<table>
<thead>
<tr>
<th></th>
<th>Memories of Lying</th>
<th>Memories of Emotional Harm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actor</td>
<td>Recipient</td>
</tr>
<tr>
<td>Morality Rating</td>
<td>3.721</td>
<td>2.940</td>
</tr>
<tr>
<td>Valence</td>
<td>3.500</td>
<td>2.775</td>
</tr>
<tr>
<td>Emotional Intensity</td>
<td>3.401</td>
<td>4.750</td>
</tr>
</tbody>
</table>

Our second prediction was that more temporally distant remembered actions will be rated as more negative than temporally nearer remembered actions when the participant is the actor but not the recipient of the action. An initial LMER of time and
condition on memory valence revealed a significant interaction between time and condition ($b = .099$, $SE = .038$, $F(1, 452.321) = 6.558$, $p = .011$, 95% CI = [.029, .171]). Another LMER of time and condition on emotional intensity revealed no significant interaction between time and condition ($b = -.038$, $SE = .050$, $F(1, 429.378) = .556$, $p = .456$, 95% CI = [-.136, .059]). To better assess these relationships, separate follow-up LMERs for actor and recipient conditions were computed.

A LMER of time on memory valence revealed a significant effect of time ($b = -.085$, $SE = .025$, $F(1, 30.515) = 10.199$, $p = .003$, 95% CI = [-.135, -.033]) such that actions in more temporally distant memories in the actor condition were more negative than temporally nearer memories. Even after controlling for differences in how well each event was remembered and how well participants remembered how they felt during each event, time was significantly related to valence ($p = .005$). Another LMER of time on memory valence revealed no significant effect of time ($b = -.003$, $SE = .029$, $F(1, 39.459) = .051$, $p = .822$, 95% CI = [-.064, .056]) in the recipient condition. Supporting our second prediction, these results suggest that actions in more temporally distant memories that involve lying tend to be more negative than actions in nearer memories in the actor condition but not the recipient condition.

Our third prediction was that more temporally distant remembered actions will be judged to be more morally wrong than temporally nearer remembered actions when the participant is the actor but not the recipient. An initial LMER of time and condition
on moral wrongness revealed a significant interaction between time and condition ($b = \cdot100, SE = .041, F(1, 429.532) = 5.888, p = .016, 95\% CI = [.022, .182]$). Therefore, separate follow-up LMERs for actor and recipient conditions were computed.

In the actor condition, a LMER of time on moral wrongness revealed a significant effect of time ($b = -.110, SE = .028, F(1, 35.563) = 10.882, p = .002, 95\% CI = [-.172, -.052]$) such that actions in more temporally distant memories were rated as more morally wrong than nearer ones. Even after controlling for differences in how well each event was remembered and how well participants remembered how they felt during each event, time was still significantly related to moral wrongness ($p = .003$). Another LMER of time on moral wrongness revealed no significant effect of time ($b = -.020, SE = .030, F(1, 38.084) = 0.287, p = .595, 95\% CI = [-.080, .040]$) in the recipient condition. Supporting our third prediction, these results suggest that actions in more temporally distant memories of lying tend to be rated as more morally wrong than actions in nearer memories in the actor condition but not the recipient condition.

Our fourth prediction was that the more people believe they have changed since the event occurred, the more likely they are to judge their actions during that time in the past as more morally wrong in the actor condition but not the recipient condition. Because ratings of personal change severely violated the assumption of normality, we binarized the personal change variable with a median split. An initial LMER of personal change (high change or low change) and condition on moral wrongness ratings revealed
a marginally significant interaction between personal change and condition ($b = .561$, $SE = .292$, $F(1, 481.585) = 3.612, p = .057$, 95% CI = [.017, 1.126]). Therefore, separate follow-up LMER models for actor and recipient conditions were computed.

In the actor condition, a LMER of personal change on moral wrongness revealed a significant effect of personal change ($b = -.855$, $SE = .258$, $F(1, 32.846) = 10.520, p = .003$, 95% CI = [-1.331, -.339]) such that the more people believe they have changed since the event occurred, the more likely they are to judge their actions as more morally wrong. Even after controlling for differences in how well each event was remembered and how well participants remembered how they felt during each event, personal change was still significantly related to moral wrongness ($p = .004$). Another LMER of personal change on moral wrongness revealed no significant effect of time ($b = -.273$, $SE = .260$, $F(1, 33.190) = 1.083, p = .306$, 95% CI = [-.786, .276]) in the recipient condition. Supporting our fourth prediction, the more people believe they have changed since the event occurred, the more likely they are to judge their actions during that time in the past as more morally wrong in the actor condition but not the recipient condition.

4.3 Study 1b

4.3.1 Materials and method

4.3.1.1 Participants

A total of 48 adults voluntarily participated in this study. Four participants were excluded due to failures in following instructions. Data were analyzed with the
remaining 44 participants ($M_{age} = 22$ years, $SD = 3$, age range = [18, 30]; 30 females, 14 males), and each participant was tested individually. All participants were fluent English speakers.

4.3.1.2 Procedure

The procedures employed in Study 1b are identical to those in Study 1a with one exception: each participant was cued to recall as many specific, autobiographical memories as possible of emotional harm from actor and recipient conditions.

Accordingly, while Study 1a investigated memories of lying, Study 1b investigates memories of emotional harm. Table 5 provides several examples of memories recalled by participants.

Table 5: Examples of memories of emotional harm from both actor and recipient conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Recalled Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actor</td>
<td>I did not go to my friend’s wedding, even though I was free that day. He really wanted me to be there, but I lied to him and said that I had an unexpected family emergency.</td>
</tr>
<tr>
<td>Actor</td>
<td>My sister and I forgot that it was Mother’s Day, and we spent all day fighting and arguing with each other and our mom. My mom started crying and left the house.</td>
</tr>
<tr>
<td>Recipient</td>
<td>A person racially abused me by calling me “Apu” which is a racial slur used for Indians. She did that when I told her that I did not support her candidate for the presidency. She got extremely annoyed and said “Apu, you are moron. Go back to eating cow dung in India”</td>
</tr>
</tbody>
</table>
Recipient

I had been taking care of my dad who had been going through chemotherapy for weeks. One day I told my brother that I wasn’t going to visit my dad at the hospital that day. My brother got mad at me and screamed at me for changing the schedule. I was hurt because I had been taking care of dad alone for a week, and I just needed break. My brother didn’t seem to understand how exhausted I was.

4.3.1.3 Statistical analyses

All data were analyzed using the same software and methods as in Study 1a.

4.3.2 Results and discussion

On average, participants generated 5.36 memories (SD = 1.64) in the actor condition and 5.91 memories (SD = 1.41) in the recipient condition. Across all participants, a total of 236 memories were generated in the actor condition, and a total of 260 memories were generated in the recipient condition. Figure 5d-f depicts frequency statistics for each variable split by actor and recipient conditions.

Our first prediction was that remembered events in the actor condition will be rated as less morally wrong and less negative than those in the recipient condition. A LMER of condition (binary factor: actor, recipient) on moral wrongness showed that condition did not reach significance (b = -.377, SE = .200, F(1, 45.153) = 3.555, p = .066, 95% CI = [-.766, -.005]). Actions in the recipient condition were rated as somewhat more morally wrong than those in the actor condition (Figure 5d). Another LMER of condition on valence revealed a significant effect of valence (b = -.737, SE = .139, F(1, 46.931) = 28.115, p < .0001, 95% CI = [-1.001, -.446]) such that remembered actions in the recipient
condition were more negative than those in the actor condition (Figure 5e). A final LMER of condition on emotional intensity revealed a significant effect of emotional intensity \( (b = .852, SE = .172, F(1, 44.132) = 24.415, p < .0001, 95\% \text{ CI} = [.521, 1.170]) \) such that remembered actions in the recipient condition were more emotionally intense than those in the actor condition (Figure 5f). After controlling for differences in how well each event was remembered and how well participants remembered how they felt during each event, condition was significantly related to moral wrongness \( (p = .047) \), valence \( (p < .0001) \), and emotional intensity \( (p = .0002) \). Table 4 summarizes overall means for morality, valence, and emotional intensity ratings split by actor and recipient conditions. Supporting our first prediction, these results suggest that memories of emotional harm are perceived as more morally wrong (after controlling for quality of memory), more negative, and more emotionally intense in the recipient condition than in the actor condition.

Our second prediction was that more temporally distant remembered actions will be rated as more negative than temporally nearer remembered actions in the actor condition but not the recipient condition. An initial LMER of time and condition on memory valence revealed no significant interaction between time and condition \( (b = .000, SE = .036, F(1, 474.154) = .000, p = .994, 95\% \text{ CI} = [-.079, .066]) \). Similarly, the LMER of time and condition on emotional intensity revealed no significant interaction between time and condition \( (b = .021, SE = .048, F(1, 450.528) = .192, p = .662, 95\% \text{ CI} = [-.073,
.114]). Unlike the results for memories of lying, there were no actor-recipient differences in the relationship between the temporal distance of the remembered actions and their affective qualities.

Our third prediction was that more temporally distant remembered actions will be judged to be more morally wrong than temporally nearer remembered actions when the participant is the actor but not the recipient. An initial LMER of time and condition on moral wrongness revealed a significant interaction between time and condition ($b = .097, SE = .043, F(1, 455.534) = 5.009, p = .026, 95\% CI = [.010, .189]$). Therefore, separate follow-up LMER models for actor and recipient conditions were computed.

In the actor condition, an LMER of time on moral wrongness revealed a significant effect of time ($b = -.158, SE = .037, F(1, 33.924) = 17.875, p = .0002, 95\% CI = [-.227, -.093]$) such that actions in more temporally distant memories in the actor condition were rated as more morally wrong than those in nearer memories. Even after controlling for differences in how well each event was remembered and how well participants remembered how they felt during each event, time was still significantly related to moral wrongness ($p = .0004$). Another LMER of time on moral wrongness revealed no significant effect of time ($b = -.029, SE = .029, F(1, 27.221) = .881, p = .356, 95\% CI = [-.083, .028]$) in the recipient condition. Supporting our third prediction, these results suggest that actions in more temporally distant memories of emotional harm tend to be rated as
more morally wrong than actions in nearer memories in the actor condition but not the recipient condition.

Our fourth prediction was that the more people believe they have changed since the event occurred, the more likely they are to judge their actions during that time in the past as more morally wrong in the actor condition but not the recipient condition. Because ratings of personal change severely violated the assumption of normality, we binarized the personal change variable with a median split. An initial LMER of personal change (high change or low change) and condition on moral wrongness ratings revealed a significant interaction between personal change and condition \( (b = .564, SE = .264, F(1, 459.931) = 4.523, p = .034, 95\% CI = [.038, 1.086]) \). Therefore, separate follow-up LMER models for actor and recipient conditions were computed.

In the actor condition, a LMER of personal change on moral wrongness revealed a significant effect of personal change \( (b = -1.025, SE = .248, F(1, 33.078) = 16.473, p = .0003, 95\% CI = [-1.499, -.539]) \) such that the more people believe they have changed since the event occurred, the more likely they are to judge their actions as more morally wrong in the actor condition. Even after controlling for differences in how well each event was remembered and how well participants remembered how they felt during each event, personal change was still significantly related to moral wrongness \( (p = .0005) \). Another LMER of personal change on moral wrongness revealed no significant effect of personal change \( (b = -.066, SE = .188, F(1, 34.301) = .118, p = .733, 95\% CI = [-.477,
.268) in the recipient condition. Supporting our fourth prediction, the more people believe they have changed since the event occurred, the more likely they are to judge their actions during that time in the past as more morally wrong in the actor condition but not the recipient condition.

4.4 Study 2

In Study 2 we sought to directly assess the effect of perceived differences in the self over time on ratings of morality, valence, and emotional intensity for memories of lying and emotional harm from both actor and recipient conditions. While temporal distance serves as a useful proxy for psychological or subjective distance, temporal self-appraisal theory specifically posits that it is the subjective distance between past experiences and the present that underlies the self-enhancement and self-protection functions of autobiographical memory (Ross & Wilson, 2002; Wilson & Ross, 2001, 2003). Accordingly, we cued participants to provide autobiographical memories (1) when they believed they were very different people than they are now (different-self condition) and (2) when they believed they were very similar to or the same as who they are now (similar-self condition). Additionally, despite being cued to generate memories of lying and emotional harm in Study 1, participants still rated some actions in these memories as morally right and others as morally neutral. In Study 2, we sought to investigate only memories for morally wrong actions. Hence, we cued participants to provide autobiographical memories of specifically immoral actions from their personal pasts.
We make two specific predictions in Study 2 that extend prior research on temporal self-appraisal theory (Wilson & Ross, 2003), expand upon work from Escobedo and Adolphs (2010), and adding to a growing literature on the central role of morality in the construction and perception of the self (Strohminger, Knobe, & Newman, 2017). First, we predict that memories of lying and emotional harm from the actor condition in the different-self condition will be judged as more negative than memories in the similar-self condition. Second, we predict that memories of lying and emotional harm from the actor condition in the different-self condition will be judged as more morally wrong than memories in the similar-self condition.

4.4.1 Materials and method

4.4.1.1 Participants

A total of 105 adults 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 85%. The sample size in Study 2 was chosen in order to obtain a similar number of participants and memories to those acquired by Escobedo and Adolphs (2010). Two participants were excluded because they were unable to recall memories. Three participants were excluded due to failures following instructions. Data were analyzed with the remaining 100 participants ($M_{age} = 32.96, SD = 8.45$, age range = [19, 65]; 38 females, 60 males). All participants reported being fluent English speakers.
4.4.1.2 Procedure

Participants were asked to recall a total of eight events from their personal pasts and to provide several ratings for each recalled event. They were instructed to provide only memories of actions they believed to be morally wrong. Participants were instructed to provide exactly one memory from each of eight cues: (1) during a period of time in which you felt you were a very different person than the person you are now, recall a specific memory of an event in which you lied to someone else; (2) during a period of time in which you felt you were a very different person than the person you are now, recall a specific memory of an event in which you found out that someone else lied to you; (3) during a period of time in which you felt you were a very different person than the person you are now, recall a specific memory of an event in which you emotionally harmed someone else; (4) during a period of time in which you felt you were a very different person than the person you are now, recall a specific memory of an event in which you were emotionally harmed by someone else; (5) during a period of time in which you felt you were very similar to or the same as the person you are now, recall a specific memory of an event in which you lied to someone else; (6) during a period of time in which you felt you were very similar to or the same as the person you are now, recall a specific memory of an event in which you found out that someone else lied to you; (7) during a period of time in which you felt you were very similar to or the same as the person you are now, recall a specific memory of an event in
which you emotionally harmed someone else; (8) during a period of time in which you felt you were very similar to or the same as the person you are now, recall a specific memory of an event in which you were emotionally harmed by someone else. The order of the eight cues was presented randomly. Participants were told that all memories must be of events that have occurred when they were older than 10 years of age, must be of one particular episode specific in time and place, and must have been personally experienced.

For each memory, participants described the event in 2-6 sentences and listed the time when the event occurred (month and year). Then, participants provided the following ratings for the actions in each memory in this order: how confident are you that the event occurred when you believe it did? (1 = not at all confident, 7 = very confident); how morally wrong was the action performed? (1 = slightly morally wrong, 7 = very morally wrong); what were your emotions associated with the event? (1 = very negative, 7 = very positive); what was the intensity of the emotion you felt during the event? (1 = not at all intense, 7 = very intense); and to what extent do you believe you are the same person now compared to the person you were around the time that the remembered event occurred (1 = very similar, 7 very different). We have reported all measures, conditions, and data exclusions. Upon completion of the study, participants were monetarily compensated for their time.
4.4.2 Results and discussion

Participant responses for each dependent variable were not normally distributed. Accordingly, we implemented non-parametric statistical tests for related samples. Sign Tests for paired observations were used to assess significance and the direction of effects. To assess effect sizes, we used $PS_{dep}$ (Grissom & Kim, 2012), which is the probability that in a randomly sampled pair of scores, the score from one condition (the condition which most frequently has the higher score) will be greater than the score from the other condition (the condition which most frequently has the lower score). Ties were discarded to obtain $PS_{dep}$ values. Possible effect sizes ranged from 0.50 to 1.00 with higher values indicating larger effects.

Verifying that participants followed instructions, across all conditions, ratings of personal identity change (on the 7-pt scale) when instructed to generate memories in the different-self condition were much higher than when instructed to generate memories in the similar-self condition (Sign Tests: all $p$’s < .001). Additionally, memories of lying and emotional harm cued in the different-self condition typically occurred in the more temporally distant past than those cued in the similar-self condition (Sign Tests: all $p$’s < .001), and participants were very confident in their assessments of when the events occurred (range of median confidence ratings across all conditions: [6-7]).

For memories of lying in the actor condition, we found a significant difference in moral wrongness ($Z = 3.604, p < .001, PS_{dep} = 0.72$) and valence ($Z = 2.988, p = .003, PS_{dep} =$
0.69) but not emotional intensity ($Z = 1.354, p = .176, PS_{dep} = 0.59$) as a function of perceived similarity to the current sense of self. For memories of events in the different-self condition, participants judged their remembered actions as more morally wrong and more negative as compared with memories of events in the similar-self condition. For memories of lying in the recipient condition, we found no significant differences in moral wrongness ($Z = 1.488, p = .137, PS_{dep} = 0.60$), valence ($Z = 1.793, p = .073, PS_{dep} = 0.62$), or emotional intensity ($Z = 1.424, p = .154, PS_{dep} = 0.59$) as a function of perceived similarity to the current sense of self. Supporting our predictions, these results indicate that participants are more likely to rate their lying as more morally wrong when they felt that they were very different than—as opposed to very similar to—who they are now. Figure 6a summarizes the results involving lying from the Sign Tests: The percentage of negative differences (i.e., different-self condition minus similar-self condition), positive differences (i.e., similar-self condition minus different-self), and ties (i.e., no difference in ratings between different-self and similar-self conditions) are depicted. Figure 7 depicts frequency statistics for each measure split by condition (different-self versus similar-self) and condition.
Figure 6: Frequency statistics for morality, valence, and emotional intensity ratings for memories of lying (a) and emotional harm (b) in both actor and recipient conditions.
Figure 7: For memories of lying, frequency statistics are displayed for ratings of morality, valence, and emotional intensity.

For memories of emotional harm in the actor condition, we found a significant difference in moral wrongness ($Z = 4.002, p < .001, PS_{dep} = 0.75$) and valence ($Z = 2.757, p = 0.006, PS_{dep} = 0.69$) but not emotional intensity ($Z = .125, p = .901, PS_{dep} = 0.52$) as a function of perceived similarity to the current sense of self. For memories in the different-self condition, participants perceived those actions as more morally wrong and more negative compared with memories in the similar-self condition. For memories of emotional harm in the recipient condition, we found no significant differences in moral wrongness ($Z = .889, p = .374, PS_{dep} = 0.56$), valence ($Z = 1.143, p = .253, PS_{dep} = 0.59$), or
emotional intensity \((Z = .265, p = .791, PS_{dep} = 0.53)\) as a function of perceived similarity to the current sense of self. Supporting our predictions, these results indicate that participants are more likely to rate their emotional harming as more morally wrong when they felt that they were very different than—as opposed to very similar to—who they are now. Figure 6b summarizes the results involving emotional harm from the Sign Tests: The percentage of negative differences (i.e., different-self condition minus similar-self condition), positive differences (i.e., similar-self condition minus different-self), and ties (i.e., no difference in ratings between different-self and similar-self conditions) are depicted. Figure 8 depicts frequency statistics for each measure split by condition and condition.
Figure 8: For memories of emotional harm, we display frequency statistics for ratings of morality, valence, and emotional intensity.

4.5 General discussion

We report three studies investigating whether there are actor-recipient differences in the emotional qualities and perceived moral wrongness of remembered actions from one’s personal past, and whether the perceived emotional and moral content of these remembered actions differs as a function of temporal distance and perceived changes in the self over time. These studies yielded four main findings. First, remembered actions in the actor condition were rated as less morally wrong and less negative than remembered actions in the recipient condition. These findings held for
memories of lying and of emotional harm. Second, remembered actions rated as more negative were more temporally distant in the actor condition than in the recipient condition—but only for memories of lying. Third, the more temporally distant remembered actions were, the more likely they were to be rated as more morally wrong in the actor condition but not in the recipient condition. These findings held both for memories of lying and emotional harm. Fourth, during periods of time when participants felt they were very different than—as opposed to very similar to—who they are now, they judged their own acts of lying and emotional harm as more morally wrong and more negative.

Consistent with research showing that people evaluate themselves in more positive terms than they evaluate others (Alicke & Sedikides, 2009; Taylor & Brown, 1988, 1994), our results suggest that for autobiographical memories of lying and emotional harm, people recall events that are rated as being less morally wrong, less negative, and less emotionally intense when they, as opposed to someone else, commit the act. Autobiographical memory research has consistently shown that recollections of past events are easily and often distorted (Schacter, Guerin, & St. Jacques, 2011; Schacter & Slotnick, 2004). Moreover, memories of immoral events are often emotionally salient (Escobedo & Adolphs, 2010), and emotional memories tend to be recollected vividly and readily but often inaccurately (Christianson & Loftus, 1991; Heuer & Reisberg, 1990). Given the frequent facilitation of self-enhancement and self-protection through viewing
one’s actions as more favorable than others’ and the frequent distortion of emotional memories, our results can be interpreted in two non-exclusive ways: (1) people are motivated to recall specific events that reduce or mitigate a negative self-image and/or enhance a positive self-image, or (2) people reconstruct or reinterpret existing memories to reduce or mitigate a negative self-image and/or enhance a positive self-image. Further work needs to determine the circumstances under which either of these two possibilities hold.

Consistent with Escobedo and Adolphs (2010) regarding the relationship between the affective qualities of memories and their temporal distance, we also found that actor condition memories of lying that were more negative tended to be more temporally distant than less negative memories (Study 1a). Moreover, assessing subjective, psychological distance between current and past selves, we found that memories of lying in the different-self condition were rated as more negative than memories of lying in the similar-self condition (Study 2). Our results for memories of emotional harm are more nuanced. When participants were cued to provide memories of emotional harm, we found no relationship between valence and temporal distance in the actor condition (Study 1b). However, when we indexed subjective, psychological distance and specifically cued participants to provide memories involving immoral emotional harm, these memories in the different-self condition were rated as more negative than those in the similar-self condition (Study 2). These differences between
our studies are easily explained. When participants were not specifically cued only to provide memories of immoral actions, they sometimes recalled memories of lying and emotional harm that they rated as being morally neutral. In Study 1b, temporal distance was used as a proxy for subjective, psychological distance. Subjective, psychological distance between past experiences and the present underlies the self-enhancement and self-protection functions of autobiographical memory in temporal self-appraisal theory (Ross & Wilson, 2002; Wilson & Ross, 2001, 2003). The inclusion of these morally neutral memories and the use of temporal distance as a proxy for subjective, psychological distance increases noise, which, in turn, decreases the likelihood of uncovering the predicted effects. In Study 2, we removed this unwanted noise by assessing subjective, psychological distance and by specifically asking participants to recall memories of immoral actions, and we found a significant difference in both valence and moral wrongness between the different-self condition and the similar-self condition.

Like Escobedo and Adolphs (2010), our results are also largely consistent with and expand upon work on temporal self-appraisal theory, which maintains that people view themselves as improving over time in order to enhance or protect the current self (Wilson & Ross, 2003). Work supporting this theory shows that people evaluate their current self more favorably than earlier selves and perceive themselves as having improved over time (Wilson & Ross, 2001, 2003). Expanding upon this work, our results show that more temporally distant memories were rated as more morally wrong than
nearer memories. Similarly, assessing subjective, psychological distance from the current self, we found that remembered actions of lying and emotional harm in the subjectively distant past were rated as more morally wrong than memories in the subjectively nearer past. Given the relative differences in effect sizes, our results also suggest that moral evaluations of past actions may be more important than valence for evaluating the current self more favorably than earlier selves and for perceiving improvement in oneself over time.

One might be tempted to argue that more morally wrong acts are more negative, emotionally intense, and encoded more deeply in memory, so they tend to be remembered longer and recollected more easily. Accordingly, one might suggest that our results are simply the product of the differential rate that memories fade over time as a function of their emotional valence and intensity. If this were the case, however, then we would expect that more negative, emotionally intense, and morally wrong memories would be more temporally and subjectively distant from both actor and recipient conditions. Our data show that there is no relationship between temporal or subjective distance and valence, emotional intensity, or moral wrongness in the recipient condition. Furthermore, research on the fading affect bias suggests that affect associated with negative memories fades more quickly than affect associated other kinds of memories (Skowronski, Walker, Henderson, & Bond, 2014). Therefore, it is unlikely that
our findings are the product of a more general phenomenon that memories with different emotional qualities are forgotten at different rates over time.

Autobiographical memory and morality are essential in the construction and perception of one’s identity over time. Many researchers have found that some parts of the self are more authentic, genuine, or central than others (Strohminger, Knobe, & Newman, 2017). Interestingly, when different characteristics and qualities are compared with each other (e.g., perception, memories, preferences, personality), people typically report the greatest identity discontinuity when moral characteristics have been altered or removed (Strohminger & Nichols, 2014, 2015; Heiphetz, Strohminger, & Young, 2016), suggesting that moral traits are perhaps the most prominent and central part of a person’s identity (Chen, Urminsky, & Bartels, 2016). Our results provide a novel contribution to this burgeoning literature on the role of morality in constructing and perceiving the self by showing that participants judge their moral transgressions during periods of time when they felt very different than—as opposed to very similar to—their current selves as more morally wrong. Doing so likely facilitates a greater sense of self-improvement over time and the ability to view oneself as moral and good in the present. Although what purportedly comprises the self includes core beliefs and events from the personal past, few researchers have investigated recollections of moral transgressions from the personal past or how those recollections interact with their conceptions of the self. In part due to the inherent difficulty in systematically
investigating autobiographical memories (St. Jacques & De Brigard, 2015), the majority of moral psychology research has employed designs that present moral content in the form of artificial vignettes created in the laboratory (e.g., Chituc, Henne, Sinnott-Armstrong, & De Brigard, 2016; Clifford, Iyengar, Cabeza, & Sinnott-Armstrong, 2015). Comparatively few studies have investigated prototypical, everyday instances of immorality that have been personally experienced (however, see Escobedo & Adolphs, 2010; FeldmanHall et al. 2012; Hofmann et al. 2014; Knutson et al. 2010). This imbalance is surprising given our prioritization of moral traits and values when judging the overall favorability of others (Goodwin, Piazza, & Rozin, 2014) and the fact that we define personal identity largely in terms of moral characteristics (Strohminger & Nichols, 2014; Strohminger et al., 2017). We provide novel contributions by investigating how memories of lying and emotional harm are associated with the maintenance of a positive identity. Moreover, we hope that the studies reported in the current paper encourage researchers to continue exploring different ways of integrating research on autobiographical memory, moral cognition, and personal identity.
5. Remembering moral and immoral actions in constructing the self

5.1 Introduction

A critical function of remembering the personal past is to construct a sense of self: remembering events from the past informs who we are and who we wish to be (Bluck, 2003; Conway, 2005; Fivush, Habermas, Waters, & Zaman, 2011; McAdams, 2013; Wilson, Gunn, & Ross, 2009). But the person we believe ourselves to be in the present influences which events we readily recall, how we recall those events, and how we come to interpret the significance of those events (Greenwald, 1980; Ross, 1989; Conway, 2005; Wilson & Ross, 2003). At least in Western cultures, most people are motivated to maintain a positive view of themselves in the present (Alicke & Sedikides, 2009; Baumeister, 1998; Sedikides, 1993). Systematic biases and distortions in recollecting specific past events and in constructing a life narrative help to maintain, enhance, and protect this positive view of the self.

In certain contexts, people construct a positive view of the self in the present by readily and selectively recalling past behaviors that showcase their positive traits and qualities (Pasupathi, Mansour, & Brubaker, 2007; Markus & Wurf, 1987; Ross, 1989; Sanitioso, Kunda, & Fong, 1990). In one line of research, experimenters led participants to believe that either extroverts or introverts tend to be more successful (Kunda, Fong, 2019).

---

Sanitioso, & Reber, 1993; Kunda & Sanitioso, 1989; Sanitioso, Kunda, Fong, 1990). Those led to believe that extroversion is conducive to success more readily recalled past events showcasing their extroversion relative to those led to believe that introversion is conducive to success. Other research has indicated that people tend to preferentially forget negative feedback about themselves, particularly when that negative feedback is about something of personal importance (Sanitioso & Wlodarski, 2004; Sedikides & Green, 2000). Importantly, those with the strongest motivations to view themselves positively are most likely to interpret their positive experiences as characteristic of a stable sense of self that persists over time (Kunda, 1990; Wilson & Ross, 2003). For instance, a person strongly motivated to believe she is intelligent might selectively seek out, readily accept, and successfully retain information substantiating this positive view of herself as intelligent. She might then easily and frequently recall events from her personal past that support this view of herself, and those recollections would strengthen the belief that she currently is, and steadfastly has been, an intelligent person. In this way, autobiographical memories facilitate impressions of similarity (or consistency) over time in service of fostering a favorable view of the self.

The readiness with which we recall events showcasing our positive traits does not imply that we are more likely to forget experiences that portray the self negatively (Stanley, Yang, & De Brigard, 2018). When confronted by our own past shortcomings and failures, it is still possible to achieve a positive view of the self in the present by
constructing a life narrative in which negative experiences represent turning points or indications of self-improvement over time (Conway, 2005; Libby & Eibach, 2002; Wilson & Ross, 2003). Individuals compare their current selves to their past selves to perceive positive change and improvement over time, which, in turn, helps to maintain, or even to enhance, a favorable view of the current self (Demiray & Janssen, 2015; Gebauer, Broemer, Haddock, & von Hecker, 2008; Ross & Wilson, 2003). For example, Wilson and Ross (2000) manipulated the objectives of participants when instructing them to describe themselves: some were encouraged to adopt the goal of evaluating themselves favorably, and others were encouraged to evaluate themselves accurately. Participants instructed to evaluate themselves favorably were more likely to describe inferior past selves than were participants pursuing accuracy goals. In another study, Ross and Wilson (2002) found that participants felt more psychologically distant from past negative experiences, but they felt psychologically nearer to past positive experiences. By perceiving a past achievement as psychologically close, people can continue to relish their success and establish a positive self-image; by perceiving a past failure as psychologically distant, people can dismiss that failure as belonging to a very different past self. In this way, a positive view of the self can be cultivated in the present by strategically interpreting negative past experiences within a life narrative.

When constructing a positive sense of self from autobiographical memories, memories showcasing certain traits and qualities are presumably more important than
others. Recent evidence suggests that perceiving alterations in another person’s moral traits leads to attributions of greater personal change for that person than perceiving alterations in other mental features such as perception, desires, or emotions (Heiphetz, Strohminger, & Young, 2017). Critically, people also treat moral traits as defining features of their own selves (Molouki & Bartels, 2017)—just as they do when making judgments about the identities of other people. More than any other feature of our mental lives, positive, moral traits seem to be the most deeply rooted and causally central components of the self (Strohminger & Nichols, 2014, 2015; Strohminger, Newman, & Knobe, 2017; De Freitas et al., 2017).

Despite the now considerable evidence indicating that positive moral traits are most central to personal identity, there is little research investigating how remembering moral and immoral actions from the personal past creates a morally good self-image. Most research on morality and the self has utilized vignettes and hypothetical thought experiments, not actual memories of personal past experiences. Nevertheless, many argue that autobiographical memories of one’s own personal history play a critical role in constructing the self (Conway, Singer, & Tagini, 2004; Wilson, Gunn, & Ross, 2009). Some evidence does suggest that autobiographical memories of immoral actions are particularly susceptible to biases and distortions in ways that look favorably on the rememberer (Kouchaki & Gino, 2016; Escobedo & Adolphs, 2010; Stanley et al., 2017). In particular, Stanley et al. (2017) found that participants judged their own moral
transgressions from the distant past to be more morally wrong than their more recent transgressions. Participants advantageously utilized time to buttress a belief in personal moral improvement. This finding is consistent with the notion that people frequently compare their current selves to their past selves to perceive personal improvement over time, regardless of whether that perceived improvement is accurate (D’argembeau & Van der Linden, 2008; Demiray & Janssen, 2015; Ross & Wilson, 2000; Wilson & Ross, 2003; Ryff, 1991).

If people are especially motivated to view themselves as morally good, then recollecting their past moral and immoral behaviors might produce differential perceptions of change in the self since those events occurred. People might perceive greater change in the self after recalling their own past immoral behaviors relative to moral behaviors, even if objective, calendar time cannot be readily utilized to produce an impression of moral improvement over time. By perceiving greater change and dissimilarity in the self since committing a moral transgression, people can form an impression of personal moral improvement over time. This, in turn, might foster a positive, morally good view of the self in the present. Furthermore, these perceptions of similarity or change in the self after reflecting on past immoral relative to moral behaviors might manifest in general and specific ways. More generally, people might report perceiving greater global change in their selves after recalling their past moral transgressions relative to their past morally praiseworthy deeds. More specifically,
people might report perceiving greater positive change in the extent to which they exemplify particular moral traits (e.g., honesty, loyalty) over time after recalling their past immoral behaviors relative to moral behaviors. In four studies, we investigate the extent to which people perceive change in the self after recalling moral and immoral behaviors from the personal past. We hypothesize that people strategically come to perceive meaningful change in the self after reflecting on their own immoral behaviors from the past, but that people tend to perceive similarity (lack of change) in the self over time after reflecting on their own moral behaviors from the past. In this way, people use their autobiographical memories of moral and immoral actions in complementary ways to forge a morally good sense of self in the present.

5.2 Study 1

In Study 1, we investigate whether people perceive greater dissimilarity and change in the self after recollecting their own immoral relative to moral behaviors. We further investigate whether these perceptions of dissimilarity and change persist after accounting for when participants report that the events actually occurred in the past (i.e., objective, calendar time).

5.2.1 Materials and method

5.2.1.1 Participants

One-hundred ten individuals 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 85%. Twenty participants were excluded for failing to provide at least one memory, for recalling an event that occurred more than 10 years ago, or for providing the incorrect type of memory based on the cue (i.e., a memory of a morally wrong behavior was provided when the participant was cued to recall a morally right behavior, or vice versa). As such, data were analyzed with the remaining 90 participants ($M_{\text{age}} = 36$ years, $SD = 10$, age range = [20, 71], 33 females, 56 males). All participants reported being fluent English speakers.

5.2.1.2 Procedure

The study was self-paced. Participants were asked to recall a total of six distinct behaviors, one at a time, from their personal pasts that occurred within the past 10 years. Three of these memories involved morally wrong actions committed by the participant; the other three memories involved morally right actions committed by the participant. Participants were told that these remembered actions must have occurred on a particular day in a particular place.

For each remembered behavior, participants described the event in 2-5 sentences. They then typed in the month and year that 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 two weeks, within the past month, within the past two months, within the past six months, within the past year, within the past two years, within the
past five years, within the past ten years. As a manipulation check, participants answered the following: how morally wrong or morally right was your behavior in this instance? \((1 = \text{very morally wrong}, 7 = \text{very morally right})\). Participants then answered the following three questions in random order: as you think about this memory now, do you feel like you are a different person now than you were then? \((1 = \text{definitely no}, 7 = \text{definitely yes})\); as you think about this memory now, do you feel like you are the same person now that you were then? \((1 = \text{definitely no}, 7 = \text{definitely yes})\); how much have you changed as a person since this event occurred? \((1 = \text{not at all}, 7 = \text{a lot})\). It is worth noting that these last three questions assess perceptions of psychological change over time, not whether the former self ceased to exist after the creation of a new self (Starmans & Bloom, 2018). Upon completion of the study, participants were monetarily compensated for their time.

### 5.2.1.3 Statistical analyses

Data were analyzed using R (R Development Core Team, 2009) with the ‘lme4’ software package (Bates, Maechler, Bolker, & Walker, 2015) and with the ‘lmerTest’ software package (Kuznetsova, Brockhoff, & Christensen, 2017). Data were fit to linear mixed-effects models, and subject was included as a random effect (random intercepts only) in all models. 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. Objective time was assessed in two
complementary ways. One objective time variable (hereafter referred to as timeA) was coded as follows: 0 = within the past day; 1 = within the past week; 2 = within the past two weeks; 3 = within the past month; 4 = within the past two months; 5 = within the past six months; 6 = within the past year; 7 = within the past two years; 8 = within the past five years; 9 = within the past ten years. Similar methods have been implemented to characterize the actual time that events occurred in the past (e.g., Escobedo & Adolphs, 2010; Stanley et al., 2017). The other objective time variable (hereafter referred to as timeB) 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.

5.2.2 Results and discussion

Figure 9 depicts means for each variable of interest as a function of the memory cue (i.e., morally right versus morally wrong). An initial linear mixed-effects model was computed to ensure that the remembered actions generated from the morally wrong cue were, in fact, judged to be more morally wrong on the 7-pt scale than the remembered actions generated from the morally right cue. This expectation was confirmed ($b = 3.80, SE = .08, t = 48.80, p < .001, 95\% CI = [3.64, 3.95]$).
Figure 9: Means and standard error bars are depicted for the judged morality of the remembered actions, different judgments, same judgments, and change judgments.

We then investigated the relationship between the cued morality of the action (binary, fixed factor: morally wrong versus morally right) and the extent to which participants judged themselves to be different people now than they were when the event occurred. A linear mixed-effects model revealed a significant effect of the cued morality of the action on the extent to which participants believe they are different people now than they were when the event occurred ($b = 1.59, SE = .14, t = 11.06, p < .001, 95% CI = [1.28, 1.88]$). Participants judged that, when the remembered action occurred, they were more dissimilar to their current selves for the morally wrong condition relative to the morally right condition. This effect held at a similar magnitude even after statistically controlling for time ($b = 1.34, SE = .13, t = 10.34, p < .001, 95% CI = [1.07,$
and times ($b = 1.25, SE = .13, t = 9.55, p < .001, 95\% CI = [1.01, 1.50]$) in separate models.

Next, we investigated the relationship between the cued morality of the action (binary, fixed factor: morally wrong versus morally right) and the extent to which participants judged themselves to be the same people now that they were when the event occurred. A linear mixed-effects model revealed a significant effect of the cued morality of the action on the extent to which participants believe they are the same people now as they were when the event occurred ($b = -1.72, SE = .14, t = -12.28, p < .001, 95\% CI = [-2.00, -1.45]$). Participants judged that, when the remembered action occurred, they were more similar to their current selves in the morally right condition relative to morally wrong condition. This effect held at a similar magnitude even after statistically controlling for time ($b = -1.48, SE = .13, t = -11.52, p < .001, 95\% CI = [-1.74, -1.25]$) and time ($b = -1.37, SE = .13, t = -10.66, p < .001, 95\% CI = [-1.62, -1.12]$) in separate models.

Finally, we investigated the relationship between the cued morality of the action (binary, fixed factor: morally wrong versus morally right) and the extent to which participants judged themselves to have changed since the event occurred. An initial linear mixed-effects model revealed a significant effect of the cued morality of the action on the extent of perceived change since the event occurred ($b = 1.30, SE = .14, t = 9.38, p < .001, 95\% CI = [1.04, 1.60]$). Participants judged that they had changed more since the remembered morally wrong actions occurred than they had since the morally right
actions occurred. This effect held at a similar magnitude even after statistically controlling for time\textsubscript{A} (\(b = 1.03, SE = .12, t = 8.50, p < .001, 95\% CI = [.78, 1.27]\)) and times (\(b = .95, SE = .12, t = 7.64, p < .001, 95\% CI = [.72, 1.21]\)) in separate models.

Taken together, the results of Study 1 suggest that after remembering their own morally wrong actions from the personal past, participants perceive their current selves to be more dissimilar to those past selves than they do after recollecting morally right actions; they also perceive themselves to have undergone greater change since those past transgressions occurred. These effects remained at a similar magnitude even after statistically controlling for objective, calendar time.

\textbf{5.3 Study 2}

The purpose of Study 2 is to provide further support for the roles of recalling morally right and wrong actions in service of fostering a morally good view of the self across diverse domains of morality. To this end, we investigate individuals’ perceptions of change in the self over time by directly comparing memories involving honesty versus dishonesty, helping versus harming, fairness versus unfairness, and loyalty versus disloyalty. We chose these matched pairs because people report frequently engaging in morally right and wrong actions within these domains in everyday life (Hofmann et al., 2014).
5.3.1 Materials and method

5.3.1.1 Participants

One-hundred ten individuals 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 85%. Eleven participants were excluded for failing to provide at least one memory, for providing a memory of an event that occurred more than 10 years ago, or for providing the incorrect type of memory based on the cue. As such, data were analyzed with the remaining 99 participants ($M_{age} = 34$ years, $SD = 10$, age range = [20, 69], 47 females, 51 males). All participants reported being fluent English speakers. Those who participated in Study 1 were automatically prevented from participating in Study 2.

5.3.1.2 Procedure

The study was self-paced. Participants were asked to recall a total of eight distinct actions, 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 eight memories: (1) recall a specific past experience in which you were honest with another person, and you believe your action was morally right; (2) recall a specific past experience in which you were dishonest with another person, and you believe your action was morally wrong; (3) recall a specific past experience in which you helped another person, and you believe your action was morally right; (4) recall a specific past experience in which you harmed
another person, and you believe your action was morally wrong; (5) recall a specific past experience in which you treated another person fairly, and you believe your action was morally right; (6) recall a specific past experience in which you treated another person unfairly, and you believe your action was morally wrong; (7) recall a specific past experience in which you were loyal to another person, and you believe your action was morally right; and (8) recall a specific past experience in which you were disloyal to another person, and you believe your action was morally wrong. We randomized the order in which these cues were presented across participants. This cueing procedure ultimately produced four distinct matched pairs of remembered actions—one morally right and the other morally wrong—for each particular kind of behavior: (1) honesty-dishonesty; (2) helping-harming; (3) fairness-unfairness; and (4) loyalty-disloyalty.

For each remembered action, participants described the event in 2-5 sentences. Participants then reported when the event occurred (the same two measures used in Study 1 were also used in Study 2). As a manipulation check, participants answered the following: how morally wrong or morally right was your behavior in this instance? (1 = very morally wrong, 7 = very morally right). Participants then answered the following three questions in random order: as you think about this memory now, do you feel like you are a different person now than you were then? (1 = definitely no, 7 = definitely yes); as you think about this memory now, do you feel like you are the same person now that you were then? (1 = definitely no, 7 = definitely yes); how much have you changed as a person
since this event occurred? (1 = not at all, 7 = a lot). After completing the study, participants were monetarily compensated for their time.

5.3.1.3 Statistical analyses

Data were analyzed in two ways. First, separate paired-samples t-tests were computed to investigate mean differences in the four outcome variables (i.e., judged morality, different, same, and change) as a function of memory cue (morally right versus morally wrong). Second, using R (R Development Core Team, 2009), the ‘lme4’ software package (Bates, Maechler, Bolker, & Walker, 2015), and the ‘lmerTest’ software package (Kuznetsova, Brockhoff, & Christensen, 2017), data were fit to linear mixed-effects models that included the memory cue as a predictor of the different, same, and change outcome variables in separate models. \( Time_A \) and \( time_B \) were included as controls in separate models (due to multi-collinearity, they could not both be included in the same model), and subject was included as a random effect (random intercepts only) in all models. As in Study 1, 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.

5.3.2 Results and discussion

Four separate paired-samples t-tests were initially computed to verify that the remembered actions generated from the morally wrong cue were, in fact, judged to be more morally wrong than the remembered actions generated from the morally right cue.
Confirming this expectation, remembered actions generated from the morally wrong cue were rated as more morally wrong on the 7-pt scale than those generated from the morally right cue for each matched pair: honesty-dishonesty ($M_{diff} = 3.75$, $t(98) = 24.41$, $p < .001$, 95% CI = [3.44, 4.05]), helping-harming ($M_{diff} = 4.08$, $t(98) = 29.06$, $p < .001$, 95% CI = [3.80, 4.36]), fairness-unfairness ($M_{diff} = 3.82$, $t(98) = 30.04$, $p < .001$, 95% CI = [3.57, 4.07]), and loyalty-disloyalty ($M_{diff} = 3.99$, $t(98) = 23.70$, $p < .001$, 95% CI = [3.66, 4.32]). All reported 95% CIs are for mean differences. See Table 6 for descriptive statistics.

Table 6: Summary of means (SDs) for each measured variable as a function of memory cue.

<table>
<thead>
<tr>
<th>Memory Cue</th>
<th>Judged Morality</th>
<th>Different (SD)</th>
<th>Same (SD)</th>
<th>Change (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honesty</td>
<td>6.05 (0.81)</td>
<td>2.97 (1.98)</td>
<td>5.16 (1.92)</td>
<td>3.08 (2.02)</td>
</tr>
<tr>
<td>Dishonesty</td>
<td>2.30 (1.14)</td>
<td>4.18 (2.18)</td>
<td>3.93 (2.15)</td>
<td>4.07 (2.12)</td>
</tr>
<tr>
<td>Helping</td>
<td>6.27 (0.75)</td>
<td>2.39 (1.78)</td>
<td>5.91 (1.65)</td>
<td>2.51 (1.81)</td>
</tr>
<tr>
<td>Harming</td>
<td>2.19 (1.08)</td>
<td>4.49 (2.10)</td>
<td>3.55 (2.05)</td>
<td>4.48 (2.12)</td>
</tr>
<tr>
<td>Fairness</td>
<td>6.11 (0.87)</td>
<td>2.53 (1.79)</td>
<td>5.56 (1.69)</td>
<td>2.59 (1.74)</td>
</tr>
<tr>
<td>Unfairness</td>
<td>2.29 (0.80)</td>
<td>4.52 (2.07)</td>
<td>3.60 (2.04)</td>
<td>4.13 (2.01)</td>
</tr>
<tr>
<td>Loyalty</td>
<td>6.20 (0.94)</td>
<td>3.08 (1.92)</td>
<td>5.07 (1.84)</td>
<td>3.28 (1.86)</td>
</tr>
<tr>
<td>Disloyalty</td>
<td>2.21 (1.04)</td>
<td>4.36 (2.08)</td>
<td>3.68 (2.08)</td>
<td>4.38 (2.11)</td>
</tr>
</tbody>
</table>

For each matched pair, we investigated the extent to which participants believed they were different people when they recalled the event than they were when the event actually occurred as a function of recalling a morally right or wrong past action. Paired-samples $t$-tests revealed that participants felt more dissimilar to their past selves after recalling morally wrong actions relative to morally right actions for honesty-dishonesty ($M_{diff} = 1.21$, $t(98) = 4.32$, $p < .001$, 95% CI = [.66, 1.77]), helping-harming ($M_{diff} = 2.11$, $t(98) =$
7.86, \( p < .001, 95\% \text{ CI} = [1.58, 2.64]\)), fairness-unfairness \( (M_{\text{diff}} = 1.99, t(98) = 7.71, p < .001, 95\% \text{ CI} = [1.48, 2.50]) \), and loyalty-disloyalty \( (M_{\text{diff}} = 1.28, t(98) = 4.48, p < .001, 95\% \text{ CI} = [.71, 1.85]) \). See Table 6. Subsequent LMERs were then computed, each with the cued morality of the past action (morally wrong versus morally right) predicting the extent to which participants believed they were different people now than they were when the event actually occurred, all while statistically controlling for when the past experiences actually occurred. For each remembered pair of events, participants felt more dissimilar to their past selves in the morally wrong condition relative to the morally right condition, even after statistically controlling for timeA and timeB in separate models (all ps < .001).

Next, for each matched pair, we investigated the extent to which participants believed they were the same people upon recalling the event as they were when the event actually occurred as a function of recalling a morally right or wrong past behavior. Paired-samples \( t \)-tests revealed that participants were more likely to feel the same as their past selves after recalling morally right actions relative to morally wrong actions for honesty-dishonesty \( (M_{\text{diff}} = 1.23, t(98) = 4.42, p < .001, 95\% \text{ CI} = [.68, 1.79]) \), helping-harming \( (M_{\text{diff}} = 2.36, t(98) = 9.25, p < .001, 95\% \text{ CI} = [1.86, 2.87]) \), fairness-unfairness \( (M_{\text{diff}} = 1.96, t(98) = 7.63, p < .001, 95\% \text{ CI} = [1.45, 2.47]) \), and loyalty-disloyalty \( (M_{\text{diff}} = 1.39, t(98) = 4.99, p < .001, 95\% \text{ CI} = [.84, 1.95]) \). See Table 6. Subsequent LMERs were then computed, each with the cued morality of the past action (morally wrong versus morally
right) predicting the extent to which participants believed they were the same people now that they were when the event occurred, all while statistically controlling for when these past experiences actually occurred. For each remembered pair of events, participants were more likely to feel the same as their past selves in the morally right condition relative to the morally wrong condition, even after statistically controlling for timeA and timeB in separate models (all p < .001).

Finally, for each matched pair, we investigated how much participants believed they had changed since the remembered event occurred as a function of whether a morally right or wrong past action was recalled. Paired-samples t-tests revealed that participants believed that they had changed more since the remembered event occurred after recalling morally wrong actions relative to morally right actions for honesty-dishonesty (Mdiff = .99, t(98) = 3.59, p = .001, 95% CI = [.44, 1.54]), helping-harming (Mdiff = 1.98, t(98) = 7.61, p < .001, 95% CI = [1.46, 2.50]), fairness-unfairness (Mdiff = 1.55, t(98) = 6.26, p < .001, 95% CI = [1.06, 2.04]), and loyalty-disloyalty (Mdiff = 1.10, t(98) = 3.90, p < .001, 95% CI = [.54, 1.66]). See Table 6. Subsequent LMERs were then computed, each with the cued morality of the past action (morally wrong versus morally right) predicting the extent to which participants believed they had changed since the remembered event occurred, all while statistically controlling for when these past experiences actually occurred. For each remembered pair of events, participants reported that they had changed more since the remembered events occurred in the
morally wrong condition relative to the morally right condition, even after statistically controlling for time\textsubscript{A} and time\textsubscript{B} in separate models (all \( p < .004 \)).

The results from Study 2 corroborate and extend our findings from Study 1. As in Study 1, we found that after recalling their own morally wrong actions from the personal past, individuals perceived greater change and dissimilarity in the self since the events occurred than they did after recalling their own morally right actions. This pattern of results was obtained for each matched pair: honesty versus dishonesty, helping versus harming, fairness versus unfairness, and loyalty versus disloyalty. In all cases, these effects remained even after statistically controlling for objective, calendar time.

### 5.4 Study 3

The results from Studies 1 and 2 provide some evidence that recollecting past immoral actions is associated with stronger feelings of dissimilarity and change in the sense of self over time than remembering past moral actions. Building upon our findings in Studies 1 and 2, the purpose of Study 3 is two-fold. First, the within-subjects design in the previous studies could have created a demand characteristic: participants might have assumed that the experimenter wanted them to answer the outcome measures differently for remembered morally wrong versus morally right behaviors. To circumvent this issue, Study 3 implements a between-subjects design. Second, the effects in Study 1 could have been produced by some other process related to how people are
perceived more generally, rather than a motivational process about the self. In order to provide more direct support for the explanation that these effects resulted from a motivational process about the self, an additional condition was included: some participants were asked to provide memories of witnessing other people commit morally right or wrong actions. If the difference between morally right and wrong conditions in different, same, and change measures are exclusive to memories about the self committing morally right and wrong actions, then the explanation for our pattern of results is likely motivational.

5.4.1 Materials and method

5.4.1.1 Participants

Two-hundred sixty individuals 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 50 HITs on AMT and had obtained an approval rating above 85%. Twenty-two participants were excluded for failing to provide at least one memory, for recalling an event that occurred more than 10 years ago, or for providing the incorrect type of memory based on the cue. As such, data were analyzed with the remaining 238 participants ($M_{age} = 36$ years, $SD = 11$, age range = [18, 70], 118 females, 119 males). All participants reported being fluent English speakers. Those who participated in the previous studies were prevented from participating in this study.
5.4.1.2 Procedure

The study was self-paced. We manipulated both the nature of the remembered behavior (morally wrong versus morally right) and the person who committed the act (self versus other). Participants were randomly assigned to one of these four conditions in a between-subjects fashion: (1) participants recalled three morally right behaviors that they themselves committed, (2) three morally wrong behaviors that they themselves committed, (3) three morally right behaviors committed by another person, or (4) three morally wrong behaviors committed by another person. Participants were told that these remembered actions must have occurred within the past 10 years on a particular day in a particular place.

For each remembered behavior, participants described the event in 2-5 sentences. Participants then reported when the event occurred (the same two measures used in Studies 1 and 2 were also used in Study 3). As a manipulation check, participants answered the following: how morally wrong or morally right was the remembered behavior in this instance? (1 = very morally wrong, 7 = very morally right). Participants then answered the different, same, and change questions in a random order. Different, same, and change judgments were always made about the individual who behaved morally or immorally. Upon completion of the study, participants were monetarily compensated for their time.
5.4.1.3 Data analyses

Using R (R Development Core Team, 2009), the ‘lme4’ software package (Bates, Maechler, Bolker, & Walker, 2015), and the ‘lmerTest’ software package (Kuznetsova, Brockhoff, & Christensen, 2017), data were fit to linear mixed-effects models. Fixed effects and outcome variables differed depending upon the model, but subject was included as a random effect (random intercepts only) in all models. TimeA and timeB were included as controls in separate models (due to multi-collinearity, they could not both be included in the same model). As in previous studies, 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.

5.4.2 Results and discussion

An initial linear mixed-effects model was computed to ensure that the remembered actions generated from the morally wrong cue were, in fact, judged to be more morally wrong on the 7-pt scale than the remembered actions generated from the morally right cue. This expectation was confirmed ($b = 3.36, SE = .16, t = 21.45, p < .001, 95\% \text{ CI} = [3.04, 3.68]$).

We investigated the effects of the cued morality of the action (binary, fixed factor: morally wrong versus morally right) and the person who committed the action (binary, fixed factor: self versus other) on different, same, and change judgments in separate linear mixed-effects models. The interaction between the cued morality of the action
(morally wrong coded as 1, morally right coded as 0) and the person committing the action (self coded as 1, other coded as 0) was significant for all three different outcome variables (i.e., different, same, and change; all $p$s < .005). Table 7 depicts full results from these three models.

Table 7: Results of three linear mixed-effects models with cued morality and the person doing the action as predictors of different, same, and change judgments, respectively.

<table>
<thead>
<tr>
<th>Outcome: Different</th>
<th>$b$</th>
<th>SE</th>
<th>$t$-value</th>
<th>$p$-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cued Morality</td>
<td>-1.18</td>
<td>.29</td>
<td>-1.60</td>
<td>.10</td>
<td>[.77, .41]</td>
</tr>
<tr>
<td>Person Doing the Action</td>
<td>-1.53</td>
<td>.31</td>
<td>-1.73</td>
<td>.08</td>
<td>[-1.16, .06]</td>
</tr>
<tr>
<td>Cued Morality x Person Doing the Action</td>
<td>1.63</td>
<td>.43</td>
<td>3.76</td>
<td>&lt; .001</td>
<td>[.82, 2.45]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome: Same</th>
<th>$b$</th>
<th>SE</th>
<th>$t$-value</th>
<th>$p$-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cued Morality</td>
<td>-1.12</td>
<td>.26</td>
<td>-1.46</td>
<td>.15</td>
<td>[-.66, .40]</td>
</tr>
<tr>
<td>Person Doing the Action</td>
<td>.43</td>
<td>.27</td>
<td>1.55</td>
<td>.12</td>
<td>[-1.14, .94]</td>
</tr>
<tr>
<td>Cued Morality x Person Doing the Action</td>
<td>-1.13</td>
<td>.39</td>
<td>-2.93</td>
<td>.004</td>
<td>[-1.87, -.35]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome: Change</th>
<th>$b$</th>
<th>SE</th>
<th>$t$-value</th>
<th>$p$-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cued Morality</td>
<td>-.26</td>
<td>.31</td>
<td>-1.38</td>
<td>.17</td>
<td>[-.86, .40]</td>
</tr>
<tr>
<td>Person Doing the Action</td>
<td>-.20</td>
<td>.32</td>
<td>-1.63</td>
<td>.10</td>
<td>[-.98, .45]</td>
</tr>
<tr>
<td>Cued Morality x Person Doing the Action</td>
<td>1.31</td>
<td>.45</td>
<td>2.89</td>
<td>.004</td>
<td>[.34, 2.33]</td>
</tr>
</tbody>
</table>

Note. All 95% CIs are for the beta-estimates.

To further interrogate these interaction effects, we computed follow-up linear mixed-effects models for self and other conditions taken separately. In the self condition, there was a significant effect of the cued morality of the action on the extent to which people believe they are different now relative to who they were when the event occurred ($b = 1.43$, $SE = .30$, $t = 4.71$, $p < .001$, 95% CI = [.81, 2.00]), the extent to which people believe they are the same now relative to who they were when the event occurred ($b = -1.27$, $SE = .27$, $t = -4.76$, $p < .001$, 95% CI = [-1.83, -.76]), and the extent to which they
believe they have changed since the event occurred ($b = 1.05, SE = .32, t = 3.26, p = .001$, 95% CI = [.50, 1.67]; see Figure 10). When participants recalled their own morally wrong actions relative to their own morally right actions, they judged themselves to be more dissimilar to their past selves who committed the actions and they judged themselves to have changed more since the event occurred. These effects persisted at a similar magnitude even after statistically controlling for timeA and timeB in separate models (all $ps < .004$).
Figure 10: For the self (a) and other (b) conditions, means and SE bars are depicted for the four outcome variables.

In the other condition, linear mixed-effects models revealed no significant effects of the cued morality of the action on different ($b = -17, SE = .31, t = -.55, p = .58, 95\% CI = [-.80, .51]$), same ($b = -19, SE = .28, t = -.67, p = .51, 95\% CI = [-.77, .36]$), or change ($b = -21, SE = .32, t = -.66, p = .51, 95\% CI = [-.83, .42]$) judgments. So, the effect of recalling morally
wrong versus morally right actions on different, same, and change judgments are evident in the self but not on the other condition.

The pattern of results from Studies 1 and 2 was replicated in Study 3 using a between-subjects design instead of a within-subjects design. That is, participants’ recollections of their own past immoral actions were accompanied by stronger feelings of dissimilarity and change in the self over time relative to recollections of their own past moral actions. This effect persisted even after statistically controlling for objective, calendar time. Study 3 also provides more direct support for a motivational explanation for our pattern of results that is about the self. Participants perceived greater change and dissimilarity in their selves after recalling their own past moral transgressions relative to their own past morally praiseworthy deeds, but there were no significant differences between remembered morally right and wrong actions committed by other people.

5.5 Study 4

In the previous three studies, we found that people report perceiving greater change and dissimilarity in their selves after recalling their own past immoral behaviors relative to moral behaviors. In Study 4, we investigate whether people specifically report greater positive change in particular moral traits (e.g., honesty, loyalty) over time after recollecting their past immoral behavior relative to moral behaviors. While recalling an immoral behavior from the personal past in which they were dishonest, participants might be motivated to believe they are more honest now than they were when the event
occurred. By believing that they have become more honest since they acted dishonestly, they can still foster a positive view of the current self as morally good by perceiving their moral improvement over time. In contrast, after recalling a morally good behavior from the personal past in which they were honest, there would be no motivation to perceive positive change since the event occurred. We further expect that the perceived changes in the possession of particular moral traits after recalling morally right versus wrong past behaviors will remain after accounting for variance attributable to objective, calendar time.

5.5.1 Materials and method

5.5.1.1 Participants

Eight-hundred sixty individuals 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 50 HITs on AMT and had obtained an approval rating above 85%. Fifty-six participants were excluded for failing to provide at least one memory, for providing a memory of an event that occurred more than 10 years ago, or for providing the incorrect type of memory based on the cue. As such, data were analyzed with the remaining 804 participants (\(M_{\text{age}} = 35\) years, \(SD = 11\), age range = \([19, 86]\), 322 females, 480 males). All participants reported being fluent English speakers. Those who participated in any of the previous studies were automatically prevented from participating in Study 4.
5.5.1.2 Procedure

The study was self-paced. Participants were asked to recall one event from the personal past that occurred within the past 10 years. There were eight different conditions to which participants were randomly assigned in a between-subjects fashion: (1) recall a specific past experience in which you were honest with another person, and you believe your action was morally right; (2) recall a specific past experience in which you were dishonest with another person, and you believe your action was morally wrong; (3) recall a specific past experience in which you helped another person, and you believe your action was morally right; (4) recall a specific past experience in which you harmed another person, and you believe your action was morally wrong; (5) recall a specific past experience in which you treated another person fairly, and you believe your action was morally right; (6) recall a specific past experience in which you treated another person unfairly, and you believe your action was morally wrong; (7) recall a specific past experience in which you were loyal to another person, and you believe your action was morally right; and (8) recall a specific past experience in which you were disloyal to another person, and you believe your action was morally wrong. As in Study 2, this cueing procedure ultimately produced four distinct matched pairs of remembered actions—one morally right and the other morally wrong—for each particular kind of behavior: (1) honesty-dishonesty; (2) helping-harming; (3) fairness-
unfairness; and (4) loyalty-disloyalty. However, unlike Study 2, all comparisons between matched pairs of remembered actions were between-subjects.

After recalling an event from the personal past, participants described the event in 2-5 sentences. Participants then reported when the event occurred (the same two measures used in the three previous studies were also used in Study 4). As a manipulation check, participants answered the following: how morally wrong or morally right was your behavior in this instance? (1 = very morally wrong, 7 = very morally right). For our outcome variables of interest, participants made one of four different ratings depending upon the particular kind of behavior remembered. Participants who recalled an event involving honesty or dishonesty answered the following question: do you believe you are a more honest person now than the person you were when this event occurred? (1 = definitely no, 7 = definitely yes). Participants who recalled an event involving helping or harming answered the following question: do you believe you are a more helpful person now than the person you were when this event occurred? (1 = definitely no, 7 = definitely yes). Participants who recalled an event involving fairness or unfairness answered the following question: do you believe you are a more fair person now than the person you were when this event occurred? (1 = definitely no, 7 = definitely yes). And participants who recalled an event involving loyalty or disloyalty answered the following question: do you believe you are a more loyal person now than the person you were when this event occurred? (1 = definitely no, 7 = definitely yes). All participants
also answered the following questions about the phenomenology of their memories:
how detailed is your memory of this event? (1 = not at all detailed, 7 = very detailed); how vivid is your memory of this event? (1 = very vague, 7 = very vivid); how clear is your memory of this event (1 = very unclear, 7 = very clear). These phenomenology measures were included to help conceal the purpose of the study. After completing the study, participants were monetarily compensated for their time.

5.5.1.3 Statistical analyses

Data were analyzed in two ways. First, separate independent-samples t-tests were computed to investigate mean differences in the outcome variables (i.e., perceived changes in honesty, helpfulness, fairness, and loyalty since the remembered event occurred) as a function of memory cue (morally right versus morally wrong). Second, data were fit to regression models that included the memory cue (morally wrong coded as 1, morally right coded as 0) as a predictor of perceived changes in honesty, helpfulness, fairness, and loyalty in separate models. TimeA and timeB were then included as controls in separate regression models (due to multi-collinearity, they could not both be included in the same model).

5.5.2 Results and discussion

Four separate independent-samples t-tests were initially computed to verify that the remembered actions generated from the morally wrong cue were, in fact, judged to be more morally wrong than the remembered actions generated from the morally right
cue. Confirming this expectation, remembered actions generated from the morally wrong cue were rated as more morally wrong on the 7-pt scale than those generated from the morally right cue for each matched pair: honesty-dishonesty ($n = 199, \text{M}_{\text{diff}} = 3.62, t(197) = 22.34, p < .001, 95\% \text{ CI} = [3.30, 3.94])$, helping-harming ($n = 197, \text{M}_{\text{diff}} = 3.95, t(195) = 25.58, p < .001, 95\% \text{ CI} = [3.65, 4.26]$), fairness-unfairness ($n = 203, \text{M}_{\text{diff}} = 3.51, t(201) = 22.51, p < .001, 95\% \text{ CI} = [3.21, 3.82]$), and loyalty-disloyalty ($n = 205, \text{M}_{\text{diff}} = 3.73, t(203) = 20.49, p < .001, 95\% \text{ CI} = [3.37, 3.38]$).

For memories involving honesty and dishonesty, we investigated the effect of the cued morality of the action (morally wrong versus morally right) on the extent to which participants judged themselves to be more honest at the time of recalling the event than they were when the event actually occurred. Relative to participants cued to provide a memory of a morally right, honest behavior ($M = 4.27, SD = 1.72$), those participants cued to provide a memory of a morally wrong, dishonest behavior ($M = 4.80, SD = 1.79$) judged themselves to be more honest at the time of recalling the event than they were when the event actually occurred ($n = 199, \text{M}_{\text{diff}} = .53, t(197) = 2.13, p = .028, 95\% \text{ CI} = [.04, 1.02]$). This effect remained at a similar magnitude even after statistically controlling for timeA and timeB in separate regression models (both $ps < .05$).

For memories involving helping and harming, we investigated the effect of the cued morality of the action (morally wrong versus morally right) on the extent to which participants judged themselves to be more helpful at the time of recalling the event than
they were when the event actually occurred. Relative to participants cued to provide a memory of a morally right, helpful behavior (\(M = 4.68, SD = 1.57\)), those participants cued to provide a memory of a morally wrong, harmful behavior (\(M = 5.21, SD = 1.65\)) judged themselves to be more helpful at the time of recalling the event than they were when the event actually occurred (\(n = 197, M_{\text{diff}} = .53, t(195) = 2.30, p = .022, 95\% \text{ CI} = [.08, .98]\)). However, this effect was rendered non-significant after statistically controlling for time\(_A\) (\(p = .052\)) and time\(_B\) (\(p > .10\)) in separate regression models.

For memories involving fairness and unfairness, we investigated the effect of the cued morality of the action (morally wrong versus morally right) on the extent to which participants judged themselves to be more fair at the time of recalling the event than they were when the event actually occurred. Relative to participants cued to provide a memory of a morally right, fair behavior (\(M = 4.33, SD = 1.66\)), those participants cued to provide a memory of a morally wrong, unfair behavior (\(M = 5.11, SD = 1.74\)) judged themselves to be more fair at the time of recalling the event than they were when the event actually occurred (\(n = 203, M_{\text{diff}} = .78, t(201) = 3.25, p = .001, 95\% \text{ CI} = [.31, 1.25]\)). This effect remained at a similar magnitude even after statistically controlling for time\(_A\) and time\(_B\) in separate regression models (both \(ps < .006\)).

For memories involving loyalty and disloyalty, we investigated the effect of the cued morality of the action (morally wrong versus morally right) on the extent to which participants judged themselves to be more loyal at the time of recalling the event than
they were when the event actually occurred. Relative to participants cued to provide a memory of a morally right, loyal behavior ($M = 4.56$, $SD = 1.66$), those participants cued to provide a memory of a morally wrong, disloyal behavior ($M = 5.16$, $SD = 1.72$) judged themselves to be more loyal at the time of recalling the event than they were when the event actually occurred ($n = 205$, $M_{diff} = .61$, $t(203) = 2.56$, $p = .011$, 95% CI = [.14, 1.07]). This effect remained at a similar magnitude even after statistically controlling for timeA and timeB in separate regression models (both $ps < .03$).

Overall, the results from Study 4 indicate that people report greater positive change in particular moral traits over time after recalling their past immoral relative to moral behaviors. This effect was obtained for all four moral traits investigated: honesty, helpfulness, fairness, and loyalty. After recollecting memories of morally wrong behaviors, there seems to be a motivation to perceive positive change in the self since the events occurred. That is, to counteract a threat to a morally good view of self, participants tend to perceive positive change over time in the extent to which they possess moral traits. In contrast, after recollecting their morally right behaviors, there is no need to perceive moral improvement over time in service of repairing a negative view of the self. Furthermore, for memories involving honesty-dishonesty, fairness-unfairness, and loyalty-disloyalty, the perceived changes in the possession of these moral traits (after recalling morally right versus wrong past events) remained at a similar magnitude even after statistically controlling for objective, calendar time. So,
regardless of how long ago the event actually occurred, people still perceived meaningful, positive psychological change in the self.

### 5.6 General discussion

Autobiographical memories are not literal records of the personal past (Conway, 2005; D’Argembeau & Van der Linden, 2008; Schacter, 1999; Schacter, Guerin, & St. Jacques, 2011). Current motivations affect which personal past experiences are recalled, how those events are reconstructed, and how those events are interpreted within a life narrative. In four studies, we found consistent evidence for a particularly strong bias in how past experiences are exploited to foster a positive moral self-image. After recollecting their own moral transgressions from the personal past, participants reported feeling dissimilar to their past selves and perceiving significant change in the self since those events occurred. In contrast, after recalling their own morally praiseworthy past actions, participants reported feeling more similar to those past selves and perceiving less change in the self since those events occurred. Not only did these effects hold for diverse domains of morality (i.e., honesty/dishonesty, helping/harming, fairness/unfairness, and loyalty/disloyalty), but they also persisted after statistically controlling for when the events actually occurred in the past (objective, calendar time). Supporting a motivational explanation, we consistently found effects of recalling morally right versus wrong actions on judgments of change over time when those past actions were committed by the participants themselves; however, there were no
significant effects when participants recalled morally right and wrong actions committed by other people.

Overall, the current findings support our hypothesis that people strategically construct a positive, morally good sense of self in the present (1) by perceiving similarity in the self after reflecting on their own morally praiseworthy actions from the past and (2) by perceiving meaningful change or transformation in the self after reflecting on their own immoral actions from the past. Moreover, these perceptions of similarity and change in the self after recollecting past moral and immoral behaviors manifested in both general and specific ways. More generally, participants reported perceiving greater global change and dissimilarity in their selves after recalling their past moral transgressions relative to their past morally praiseworthy deeds (Studies 1-3). More specifically, participants reported perceiving greater positive change in the extent to which they exemplified particular moral traits (e.g., honesty, loyalty) over time after recalling their past immoral relative to moral behaviors (Study 4). These results add to a substantial literature suggesting that the ways in which we remember and interpret our past experiences serve self-enhancement and self-protective functions (Leary, 2007; Wilson & Ross, 2003). For example, people tend to believe they have undergone personal improvement over time to a greater extent than other people (Wilson & Ross 2001), people tend to take greater personal responsibility for their past successes than their past failures (Blaine & Crocker 1993), and people are more likely to engage in self-
enhancing deceptions involving their current selves than their past selves (Robinson & Ryff, 1999). Going beyond this existing literature, our findings suggest that how we remember and interpret our past experiences helps to construct a current sense of self that is not just positive but also morally good. This result is particularly important because moral traits and characteristics are considered to be the most important, fundamental, or central part of personal identity (Chen, Urminsky, & Bartels, 2016; Strohminger & Nichols, 2014; Strohminger et al., 2017; Molouki & Bartels, 2017; De Freitas et al., 2017).

Philosophers and psychologists have used the terms ‘self’ and ‘identity’ in several distinct ways. For our purposes, there are at least two ways of thinking about the self, or personal identity, that are worth differentiating: (1) numerical identity is the sense in which a single thing persists over time and (2) qualitative identity refers to the sharing of particular properties over time (Starmans & Bloom, 2018). In assessing numerical identity, we might say that infant Neil Armstrong is, for example, identical to adult Neil Armstrong. But in assessing qualitative identity over time, we might say that Neil Armstrong prior to walking on the moon is dissimilar to, or not the same as, Neil Armstrong after walking on the moon, because walking on the moon must be a particularly transformative experience. As in our previous work (Stanley et al., 2017), we believe that the current findings specifically inform individuals’ judgments about qualitative identity over time. That is, in one sense, a given person is still the same
person before and after committing a serious moral transgression, so numerical identity remains fixed over time. But, in another sense, after reflecting on a particular moral transgression several months later, a person might report feeling very different from his past self or feeling as though he has undergone significant change since that remembered transgression occurred; so, qualitative identity changes in meaningful, predictable, and systematic ways over time.

Almost all existing research on morality and the self has utilized stylized vignettes, not actual memories of personal past experiences. Although this existing research has produced valuable insights into the relationship between morality and the self, we suggest that investigating autobiographical memories of moral and immoral behaviors that actually occurred in the past is necessary for obtaining a complete picture of how people construct and understand their moral selves. By remembering and reflecting on events from the personal past, we come to construct and understand who we believe ourselves to be in the present (Conway, 2005; McAdams, 2013; Wilson, Gunn, & Ross, 2009). Biases, distortions, and motivations in how we remember our personal pasts play critical roles in how we view our selves in the present, and these biases, distortions, and motivation are unlikely to play a role in vignette-based investigations of morality and the self.

There are positive consequences of viewing the self accurately and truthfully, even if that that view of the self isn’t particularly flattering. Maintaining an accurate,
truthful view of the self can facilitate sensible future planning, enable goal success, and cultivate personal and social well-being (Strube, 1990; Trope, 1986). Nevertheless, converging lines of evidence now indicate that enhancing or protecting a positive view of the self is frequently pursued at the expense of obtaining an accurate, truthful view of the self. Of course, views of the self do not always have to be veridical to positively affect intentions and behavior: an unrealistically favorable self-view induces feelings of efficacy and a willingness to engage in decisive action (Taylor, Lerner, Sherman, Sage, & McDowell, 2003). But sometimes there are significant costs that accompany an overriding desire view the self in a positive way. In the moral domain, people seem particularly susceptible to forgetting, dismissing, or post-hoc justifying their own moral transgressions for the sake of maintaining a positive, but inaccurate, view of the self as morally upstanding. Many people who consider themselves to be morally upstanding still cheat, steal, lie, deceive, and harm others. In fact, utilizing ecological momentary assessment, Hofmann and colleagues (2014) recently showed that people behave immorally with surprising frequency. Moral transgressions, even those that are relatively minor, often have deleterious social and financial consequences (Kouchaki & Gino, 2016; Shalvi, Gino, Barkan, & Ayal, 2015). An important goal of our future research will be to develop strategies for encouraging people to develop accurate, truthful views of the moral self, specifically in those contexts where accurate, truthful
views of the moral self are better suited to impact intentions and behavior in a positive way.

Finally, it is worth mentioning a possible alternative interpretation of our results. Because the tasks we employed instructed participants to recall moral and immoral autobiographical memories, it is possible that reported impressions of change (or lack thereof) in the self across time simply reflect the extent to which they embrace or disavow the remembered action. As such, our results may reflect participants’ feelings toward the remembered action rather than a definitive perception of change or sameness in the self across time. The problem with this alternative account is that it is not clear whether embracing or disavowing a remembered action is independent of, or even explanatorily prior to, a person’s perception of change in the self over time. Do we disavow immoral actions because we feel that we are different than we were before, or do we feel that we are different than we were before because we disavow such immoral actions? It is reasonable to think that people are as motivated to disavow a past immoral action because, upon reflection, they do not think of themselves as the same person who committed it, as they are to think of themselves as a different person than they were before precisely because they are motivated to disavow the remembered action. If these two explanations are truly distinct, then it is an empirical question to uncover right order of explanation. This, we think, is another fruitful avenue for future research on the self and moral memories.
6. The roles of memory and counterfactual simulation for past wrongdoings in fostering moral learning

6.1 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; 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 Adapted from: Stanley, M. L., Cabeza, R., Smallman, R., & De Brigard, F. (under review). Memory and counterfactual simulations for past wrongdoings foster moral learning and improvement.
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 and colleagues (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 their own past moral transgressions instead of forgetting them (Huang, Stanley, & De Brigard, 2020; Stanley, Bedrov, Cabeza, & De Brigard, 2020; Stanley & De Brigard, 2019). 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 et al., 2014; Stanley et al., 2017, 2019, 2020). The process of retrieving these moral transgressions is commonly accompanied by intense negative affect and distress (Escobedo & Adolphs, 2010; Huang, Stanley, & De Brigard, 2020; Stanley et al., 2017). Critically, Huang and colleagues (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 tend to be 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 these kinds of events tends to be negative, distressing, and threatening to their favorable self-concepts.

If, 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. 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.

When we recall our personal past events, we often consider alternative ways that past events could have transpired instead—i.e., we engage in episodic counterfactual thinking (De Brigard & Giovanello, 2012; De Brigard & Parikh, 2019). In other words,
thinking about our past experiences often entails mentally simulating “what if” or “only if” possibilities (Byrne, 2016; De Brigard & Parikh, 2019; Kahneman & Miller, 1986). Such 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; 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 et al., 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 (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, then remembering those events and simulating relevant counterfactuals could serve as guides for forming intentions and goals to behave differently in the future.

Across three 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 unethical behaviors, and then they rated 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 upon them since the event occurred. We also hypothesized that memories of past behaviors judged to be more morally wrong would also be retrieved with stronger negative emotions, more often, and thought about 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. The results from Study 1 corroborate these hypotheses.

In Study 2, we first attempted to replicate our 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 assessed participants’ intentions to behave differently in the future after recalling each event. The purpose of assessing intentions was to test the following additional hypothesis: that people would tend to more frequently generate 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. We successfully replicated the findings from Study 1 using this different cueing procedure. Moreover, the pattern of
results obtained in Study 2 is consistent with the hypothesis that people tend to more frequently generate morally upward counterfactuals after committing more severe moral transgressions, which, in turn, are associated with stronger intentions to behave differently and better in the future.

In Study 3, we implemented an experimental manipulation to further investigate the possible directive 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 re-simulated 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). Corroborating our hypothesis, we found that making accessible a morally upward counterfactual strengthened intentions for future moral improvement—relative to re-simulating the remembered event and making accessible a morally downward counterfactual.

**6.2 Study 1**

The primary purpose of Study 1 is to test the hypothesis that memories of past behaviors judged to be more morally wrong would also be retrieved with stronger negative emotions, more often, and thought about counterfactually with greater frequency.
6.2.1 Materials and method

6.2.1.1 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 ($M_{age} = 35$ years, $SD = 11$, 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.

6.2.1.2 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 2-5 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 two weeks, within the past month, within the past two months, within the past six months, within the past year, within the past two years, within the past five years, within the past ten 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.
6.2.1.3 Statistical 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 fit 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.

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 timeA) was coded as follows: 0 = within the past day; 1 = within the past week; 2 = within the past two weeks; 3 = within the past month; 4 = within the past two months; 5 = within the past six months; 6 = within the past year; 7 = within the past two years; 8 = within the past five years; 9 = within the past ten 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 (hereafter referred to as time$^2$) 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, & De Brigard (2019) for a similar methodological approach).

### 6.2.2 Results and discussion

Table 8 depicts descriptive statistics for our variables of interest. 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 on emotional valence revealed a significant effect of the severity of the moral transgressions ($b = -0.32, SE = 0.03, t = -10.17, p < .001, 95\% CI [-0.38, -0.26]$) 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 on emotional intensity revealed a significant effect of the severity of the moral transgressions ($b = 0.43, SE = 0.04, t = 11.13, p < .001, 95\% CI [0.36, 0.50]$) 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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moral Wrongness</td>
<td>4.59</td>
<td>1.75</td>
</tr>
</tbody>
</table>
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 \\

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 \((b = .42, SE = .05, t = 8.39, p < .001, 95\% CI [.32, .51])\) and time B \((b = 19.20, SE = 2.52, t = 7.63, p < .001, 95\% CI [14.28, 24.10])\) in separate models such that 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 \((b = -.31, SE = .03, t = -9.31, p < .001, 95\% CI [-.38, -.24])\) and time B \((b = -.30, SE = .03, t = -9.10, p < .001, 95\% CI [-.37, -.24])\) 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 A \((b = .45, SE = .04, t = 11.03, p < .001, 95\% CI [.37, .53])\) and time B \((b = .44, SE = .04, t = 10.76, p < .001, 95\% CI [.35, .52])\) in separate models.

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 on the frequency of
voluntary recall revealed a significant effect of the severity of the moral transgressions ($b = .25, SE = .04, t = 5.87, p < .001, 95\% CI [.16, .33])$. 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 byproduct of when the events actually occurred in the past, we computed two additional LMEMs with the judged severity of the moral transgressions on the frequency of voluntary recall after statistically controlling for timeA and timeB. This effect remained significant after statistically controlling for timeA ($b = .28, SE = .04, t = 6.31, p < .001, 95\% CI [.19, .36]$) and timeB ($b = .27, SE = .04, t = 6.05, p < .001, 95\% CI [.17, .36]$) in separate models. In addition, average ratings for voluntary recall were significantly above floor ($p < .001$; Table 8), indicating that participants voluntarily thought about their moral transgressions with some frequency.

Similarly, another LMEM with the judged severity of the moral transgressions on the frequency of involuntary recall revealed a significant effect of the severity of the moral transgressions ($b = .31, SE = .04, t = 7.34, p < .001, 95\% CI [.22, .39]$). As in the case of voluntary recall, 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 on the frequency of voluntary recall after statistically controlling for time\textsubscript{A} and time\textsubscript{B}. This effect remained significant after statistically controlling for time\textsubscript{A} ($b = .35$, $SE = .04$, $t = 7.85$, $p < .001$, 95% CI [.26, .43]) and time\textsubscript{B} ($b = .35$, $SE = .04$, $t = 7.99$, $p < .001$, 95% CI [.26, .44]) in separate models. In addition, average ratings for involuntary recall were significantly above floor ($p < .001$; Table 8), indicating that participants involuntarily thought about their moral transgressions with some frequency.

Finally, we tested whether the judged severity of the moral transgression predicted the frequency of generating morally upward counterfactuals about the event. A LMEM with the judged severity of the moral transgression on the frequency of thinking about morally upward counterfactuals since the event occurred revealed a significant effect of the severity of the moral transgression ($b = .41$, $SE = .05$, $t = 8.42$, $p < .001$, 95% CI [.31, .50]). Past transgressions judged to be more morally wrong tended to elicit more frequent counterfactual thoughts. More specifically, participants 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 on the frequency
of morally upward counterfactual generation after statistically controlling for time\textsubscript{A} and times\textsubscript{B}. This effect remained significant after controlling for time\textsubscript{A} ($b = .47$, $SE = .05$, $t = 9.11$, $p < .001$, 95% CI [.36, .56]) and times\textsubscript{B} ($b = .46$, $SE = .05$, $t = 8.94$, $p < .001$, 95% CI [.36, .56]) in separate models.

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 experience more extreme negative emotions when recalling their more severe moral transgressions, they nevertheless tend to more frequently recall and think 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.

6.3 Study 2

In Study 2, we more directly investigated whether the function of frequently simulating morally upward counterfactuals is directive. That is, we investigate 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.

6.3.1 Materials and method

6.3.1.1 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 ($M_{age} = 33$ years, $SD = 9$, age range = $[20, 69]$, 68 females, 101 males).

6.3.1.2 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 2-5 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 two weeks, within the past month, within the past two months, within the past six months, within the past year, within the past two years, within the past five years, within the past ten 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.

6.3.1.3 Statistical analyses

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

6.3.2 Results and discussion

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-pt 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 ($b = 3.34, SE = .15, t = 22.19, p < .001, 95\% CI [3.04, 3.64]$), for those involving harm ($b = 3.00, SE = .15, t = 19.53, p < .001, 95\% CI [2.68, 3.31]$), and for those involving unfairness ($b = 2.98, SE = .15, t = 19.46, p < .001, 95\% CI [2.69, 3.27]$). So, this binary variable indexing the judged severity of remembered moral transgressions will be used in subsequent analyses.

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 on valence revealed a significant effect of the severity of the moral transgression for dishonesty violations ($b = -1.03, SE = .12, t = -8.55, p < .001, 95\% CI [-1.26, -0.79]$), harm violations ($b = - .80, SE = .12, t = -6.86, p < .001, 95\% CI [-1.02, -0.57]$), and unfairness violations ($b = - .70, SE = .11, t = -6.40, p < .001, 95\% CI [-.92, -.48]$; see Figure 11a) In all 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 on emotional intensity revealed a significant effect of the severity of the moral transgression for dishonesty violations ($b = 1.68, SE = .16, t = 10.77, p < .001, 95\% CI [1.36, 2.00]$), harm violations ($b = 1.11, SE = .16, t = 6.88, p < .001, 95\% CI [.81, 1.43]$), and unfairness violations ($b = 1.20, SE = .14, t = 8.53, p < .001, 95\% CI
[.93, 1.47]; see Figure 11b) 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.

![Figure 11: Means and standard error bars are depicted for valence (a) and emotional intensity ratings (b) for remembered transgressions.](image)

The cued morality (slightly morally wrong versus 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). 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 did in fact remain
significant after controlling for timeA and timeB for violations involving dishonesty, harm, and unfairness (all ps < .001).

We also 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 on the frequency of voluntary recall revealed a significant effect of the severity of the moral transgressions for dishonesty violations (b = 1.34, SE = .17, t = 7.82, p < .001, 95% CI [.99, 1.65]), harm violations (b = .96, SE = .15, t = 6.35, p < .001, 95% CI [.66, 1.25]), and unfairness violations (b = 1.05, SE = .14, t = 7.33, p < .001, 95% CI [.78, 1.34]; See Figure 12a) In all cases, voluntary recall was more frequent for transgressions judged to be more morally wrong. Moreover, LMEMs with the judged severity of the moral transgressions on the frequency of involuntary recall revealed a significant effect of the severity of the moral transgressions for dishonesty violations (b = 1.51, SE = .15, t = 9.87, p < .001, 95% CI [1.23, 1.81]), harm violations (b = 1.05, SE = .15, t = 7.07, p < .001, 95% CI [.74, 1.37]), and unfairness violations (b = 1.22, SE = .14, t = 8.86, p < .001, 95% CI [.95, 1.47]; see Figure 12b) In all cases, involuntary recall was more frequent for transgressions judged to be more morally wrong.
Figure 12: Means and standard error bars are depicted for the frequency of voluntary (a) and involuntary (b) recall for remembered transgressions.

As before, we computed additional LMEMs to ensure that the effects of the judged severity of the moral transgressions on the frequency of voluntary and involuntary recall were not merely a byproduct of differences in when the events actually occurred in the past. These additional LMEMs revealed that past transgressions judged to be more morally wrong tended to be more frequently recalled—both voluntarily and involuntarily—even after statistically controlling for time\textsubscript{A} and time\textsubscript{B} in separate models (all ps < .001).

In addition, 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 on the frequency of morally upward counterfactual thinking revealed a significant effect of the severity of the moral transgression for dishonesty violations (b =
1.64, $SE = .18$, $t = 9.03$, $p < .001$, 95% CI [1.29, 1.98]), harm violations ($b = .99$, $SE = .18$, $t = 5.41$, $p < .001$, 95% CI [.63, 1.34]), and unfairness violations ($b = 1.22$, $SE = .17$, $t = 7.06$, $p < .001$, 95% CI [.88, 1.55]; see Figure 13a). In all cases, for past transgressions judged to be more morally wrong, participants tended to more frequently consider morally upward counterfactuals about those events.

![Figure 13](image)

**Figure 13**: Means and standard error bars are depicted for the frequency of morally upward counterfactual thinking (a) and intention strength (b).

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 byproduct 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 frequency of counterfactual thinking even after controlling for time and times in separate models (all $p$s < .001). In all cases, participants tended to more frequently consider morally upward counterfactuals for more morally wrong
remembered behaviors, relative to less morally wrong remembered behaviors, even after controlling for objective temporal distance.

Moreover, we hypothesized that, when reflecting on their more severe past moral transgressions, participants would form stronger intentions to behave differently and better in the future. LMEMs with the judged severity of the moral transgressions on the strength of intentions to behave differently in the future revealed a significant effect of the severity of the moral transgression for dishonesty violations ($b = 1.44$, $SE = .19$, $t = 7.53$, $p < .001$, 95% CI [1.07, 1.81]), harm violations ($b = .87$, $SE = .18$, $t = 4.97$, $p < .001$, 95% CI [.51, 1.19]), and unfairness violations ($b = .93$, $SE = .16$, $t = 5.71$, $p < .001$, 95% CI [.62, 1.26]; see Figure 13b). In all cases, for past transgressions judged to be more morally wrong relative to less morally wrong, participants tended to form stronger intentions to behave differently in the future.

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, respectively. 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). 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 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. Figure 14 depicts full results from all three mediation analyses. These results are consistent with our hypothesis that people tend to more frequently generate morally upward counterfactuals after committing more severe moral transgressions, which, in turn, is associated with stronger intentions to behave differently and better in the future.
A. Dishonesty

Frequency of Morally Upward Counterfactual Thinking

B = 1.64***
95% CI = [1.29, 1.98]

Severity of Transgression

Total Effect: B = 1.44***, 95% CI = [1.07, 1.81]
Direct Effect: B = .83***, 95% CI = [.44, 1.25]

B = 45***
95% CI = [.36, .54]

Strength of Intention

<table>
<thead>
<tr>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACME</td>
<td>.61*** [43, .84]</td>
</tr>
<tr>
<td>Proportion Mediated</td>
<td>.42** [28, .61]</td>
</tr>
</tbody>
</table>

B. Harm

Frequency of Morally Upward Counterfactual Thinking

B = .99***
95% CI = [.63, 1.34]

Severity of Transgression

Total Effect: B = .87***, 95% CI = [.51, 1.19]
Direct Effect: B = .53**, 95% CI = [.20, .87]

B = .39***
95% CI = [.29, .48]

<table>
<thead>
<tr>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACME</td>
<td>.34** [20, .52]</td>
</tr>
<tr>
<td>Proportion Mediated</td>
<td>.39** [23, .68]</td>
</tr>
</tbody>
</table>

C. Unfairness

Frequency of Morally Upward Counterfactual Thinking

B = 1.22***
95% CI = [.88, 1.55]

Severity of Transgression

Total Effect: B = .93***, 95% CI = [.62, 1.26]
Direct Effect: B = .54**, 95% CI = [.23, .84]

B = .37***
95% CI = [.28, .46]

<table>
<thead>
<tr>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACME</td>
<td>.39** [24, .55]</td>
</tr>
<tr>
<td>Proportion Mediated</td>
<td>.42** [26, .66]</td>
</tr>
</tbody>
</table>
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 more minor moral transgressions; participants reported having more frequently retrieved—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 more minor moral transgressions. All these effects remained statistically 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 frequency of simulating morally better ways of
acting also predicted intentions to behave differently and better in the future. The more frequently participants simulated morally better ways in which they could have acted instead, the stronger the intentions were to behave differently in the future. Ultimately, we found consistent support for our hypothesis 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.

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. Frequently recalling our moral transgressions and considering morally better ways in which we could have acted predicts the strength of intentions to behave in a morally better way in the future.

6.4 Study 3

Study 3 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 3 is to test the hypothesis that making accessible a morally upward counterfactual strengthens intentions for moral improvement—relative to re-simulating 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 re-simulated the remembered event. Participants then reported their intentions to cheat in the future.

Our central hypothesis in Study 3 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 a content-neutral pathway. 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 (McMullen & Markman, 2000; Markman et al., 2008). 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.

**6.4.1 Materials and method**

**6.4.1.1 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 ($M_{age} = 37$ years, $SD = 10$, 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).

**6.4.1.2 Procedure**

The study was self-paced. Participants were asked to recall and describe, in 2-5 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 2-5 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 2-5 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 2-5 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 and colleagues (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 Appendix B 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.

### 6.4.2 Results and discussion

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 re-simulation) 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 \((F(2, 413) = 4.01, p = .019, \eta^2_p = .02)\). 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 re-simulation condition \((M_{diff} = .41, SE_{diff} = .18, p = .025, 95\% CI [.05, .76], \text{Cohen’s} \, d = .27)\) and participants in the morally downward counterfactuals conditions \((M_{diff} = .48, SE_{diff} = .18, p = .009, 95\% CI [.12, .84], \text{Cohen’s} \, d = .33)\). There was no difference in cheating intentions between re-simulation and morally downward counterfactual conditions \((M_{diff} = .07, SE_{diff} = .18, p = .70, 95\% CI [-.29, .43], \text{Cohen’s} \, d = .04)\). Table 9 presents means and standard deviations for each condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Cheating Intention</th>
<th>Average Moral Intention</th>
<th>Average Social Norm Intention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morally Upward Counterfactual</td>
<td>2.52 (1.39)</td>
<td>2.20 (.94)</td>
<td>2.30 (.92)</td>
</tr>
<tr>
<td>Morally Downward Counterfactual</td>
<td>3.00 (1.54)</td>
<td>2.34 (1.01)</td>
<td>2.42 (1.05)</td>
</tr>
<tr>
<td>Re-Simulation</td>
<td>2.93 (1.64)</td>
<td>2.26 (1.06)</td>
<td>2.32 (1.06)</td>
</tr>
</tbody>
</table>

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. These analyses
were exploratory. 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 re-simulation) on intentions for general moral improvement. There was no significant effect of condition on average moral intention judgments ($F(2, 413) = .73, p = .48, \eta^2_p = .004$). See Table 9 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 re-simulation) on intentions for obeying social norms. There was no significant effect of condition on average social norm intention judgments ($F(2, 413) = .57, p = .56, \eta^2_p = .003$). See Table 9 for means and standard deviations for each condition.

Overall, the results from Study 3 support our primary hypothesis: that making accessible a morally upward counterfactual strengthens intentions for moral improvement—relative to re-simulating 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.

6.5 General discussion

In three studies, we investigated the role of remembering and reflecting on our past moral transgressions in service of learning from those past mistakes to facilitate personal 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 serves a directive function. When participants recalled their moral transgressions, particularly those judged to be seriously morally wrong, they tended to frequently think about morally better ways in which they could have acted. The more that participants 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 unethical actions 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 upon 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 counterfactual thinking—morally upward counterfactual thinking—that occurs after committing moral transgressions. Second, we found that the frequency of simulating morally upward counterfactuals predicts the strength of intentions to behave differently in the future. The more that people simulate upward counterfactuals after having negative experiences, the stronger intentions are to behave differently in the future.

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.

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 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; Smallman, 2013; Roese & Epstude, 2017).

Second, 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 explore 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.

We have taken a functional theoretical approach to remembering past moral transgressions, finding that people do remember 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 destined to repeatedly commit the same moral transgressions throughout their lives.
7. General Conclusions

Almost everyone believes they are fundamentally and truly morally good. For most people, positive moral traits and qualities play an integral role in defining who they believe themselves to be in the present (Aquino & Reed, 2002; Stanley & De Brigard, 2019; Strohminger et al., 2017). Nevertheless, people tend to commit moral transgressions with surprising frequency in everyday life, and they can readily remember many of their past transgressions. This dissertation has provided evidence for several complementary strategies involving memory that may help people to construct, protect, and maintain the belief that they are morally good (even if they frequently behave immorally). People tend to believe that their memories of their most morally right actions—relative to less morally right actions and immoral actions—play a more central role in constructing their own personal identities. After cheating on a test, people tend to report that they just knew the answers all along. People also strategically compare their current selves to their past selves to foster a sense of positive moral change over time. Yet, there are likely a host of other strategies that are implemented—implicitly or explicitly—to satisfy people’s need to be believe they are morally good. Future research identifying other possible strategies for maintaining a morally good sense of self will prove valuable from both basic and applied conditions.

Even if people have a strong motivation to construct, protect, and maintain a morally good self-concept, they are still capable of learning from their moral mistakes.
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. Even if people distance themselves from their particularly serious past transgressions, they can still remember those transgressions and frequently recall them. Psychological distancing does not preclude moral learning, intention formation, or moral improvement facilitated via memory processes. In fact, the capacity to attribute our more serious past transgressions to a distant, dissimilar self might render our memories of those transgressions less uncomfortable to think about in service of forming intentions for future improvement. This dissertation has provided evidence that people remember and frequently ruminate upon their more severe moral transgressions—even though they evoke strong and unpleasant negative emotions. The more that people ruminate on morally better ways in which they could have acted over time, the stronger their intentions are to behave differently and better in the future. Making accessible morally upward counterfactuals when reflecting on our past moral transgressions can strengthen intentions for moral improvement in the future. 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 may explore the content of morally upward counterfactual in moderating the relationship between counterfactual thinking and behavioral intentions (and even actual behavior).
Appendix A

Easy Questions and Answers:

1. What is the last name of the first president of the United States?
   Answer: Washington

2. How many legs do spiders have?
   Answer: Eight

3. In which country is the Taj Mahal located?
   Answer: India

4. What is the name for a group of lions called?
   Answer: Pride

5. In which state is the Grand Canyon located?
   Answer: Arizona

6. What is the name of the largest planet in our Solar System?
   Answer: Jupiter

7. What is the name of the rubber object that is hit back and forth by hockey players?
   Answer: Puck

8. What is the last name of the author who wrote “Romeo and Juliet”?
   Answer: Shakespeare

9. What is the name of the largest ocean?
Answer: Pacific Ocean

10. What is the capital of France?
   Answer: Paris

11. What is the first letter of the English alphabet?
   Answer: A

12. What is the name for a dried grape?
   Answer: Raisin

**Difficult Questions and Answers:**

Difficult Questions and Answers for Studies 1, 2, and 3 in the main manuscript:

1. What is the name for a plant that pales due to lack of sun?
   Answer: Etiolated

2. What is the name for an object shaped like a bean?
   Answer: Fabiform

3. What is the name for marine animals that move without currents?
   Answer: Nektons

4. What is the name for an anchoring point near a harbor?
   Answer: Roadstead

5. What was the first genomically sequenced fish?
   Answer: Pufferfish

6. What is the Scottish term for threat against a householder during a robbery?
7. What is the name for separation of bones without a fracture?
   Answer: Diastasis

8. What is the name for a vast desert sand dune?
   Answer: Draa

9. Which nation was the first to ratify the United Nations charter in 1945?
   Answer: Nicaragua

10. Which country has the most Nobel prize winners per capita?
    Answer: Iceland

11. What is the name of the tallest mountain on the planet Mercury?
    Answer: Caloris Montes

12. What is the name for a group of hedgehogs?
    Answer: Prickle

**Appendix B**

All intention items used in Chapter 6, Study 3:

1. In the future I will cheat on something.

2. In the future I will be disloyal to a friend.

3. In the future I will disrespect an authority figure.

4. In the future I will steal something that does not belong to me.
5. In the future I will destroy property that does not belong to me.

6. In the future I will punch somebody just because I feel like it.

7. In the future I will insult another person.

8. In the future I will talk to myself in public.

9. In the future I will wear sunglasses at night.

10. In the future I will eat dessert before the main entrée arrives.

11. In the future I will eat soup with a fork.

12. In the future I will drink coffee with a spoon.

13. In the future I will wear my clothes backwards.

14. In the future I will wear clothes that are several sizes too big.
References


Biography

Matthew Lawrence Stanley attended Wake Forest University for his undergraduate studies, graduating summa cum laude in 2014 with a B.A. in Philosophy and a B.S. in Psychology. As an undergraduate student, he worked at the Lab for Complex Brain Networks where he published three first-author, peer-reviewed papers: “Defining nodes in complex brain networks,” “Changes in global and regional modularity associated with increasing working memory load,” and “Changes in brain network efficiency and working memory performance in aging.”

While at Duke University, Matthew has published a total of 35 peer-reviewed papers in journals that include: Psychonomic Bulletin & Review (x2), Memory & Cognition (x3), Cognition, Cognitive Science (x2), Journal of Applied Research in Memory & Cognition (x3), Memory, Journal of Experimental Psychology: General (x3), Journal of Personality, Journal of Experimental Social Psychology, Political Behavior, Thinking & Reasoning, Judgment and Decision Making, Philosophy of Science, Current Directions in Psychological Science, Trends in Cognitive Sciences, Language, Cognition, & Neuroscience, Cerebral Cortex (x2), Human Brain Mapping, and Neurobiology of Aging. The titles of these papers include: “Structure-seeking as a psychological antecedent of beliefs about morality,” “Students’ reasoning about dilemmas in business ethics,” “Perceived similarity of imagined possible worlds affects judgments of counterfactual plausibility,” “Truncating bar graphs persistently misleads viewers,” “Age-related compensatory reconfiguration of PFC connections during
episodic memory retrieval,” “Cultural identity changes the accessibility of knowledge,”
“False beliefs: Byproducts of an adaptive knowledge base?,” “Analytic-thinking predicts
hoax beliefs and helping behaviors in response to the COVID-19 pandemic,” “Cheaters
claim they knew the answers all along,” “Intellectual humility and perceptions of
political opponents,” “Closed-minded cognition: Right-wing authoritarianism is
negatively related to belief updating,” “Exposure to opposing reasons reduces negative
impressions of ideological opponents,” “Prediction and topological models in
neuroscience,” “Moral values reveal the causality implicit in verb meaning,” “The
centrality of remembered moral and immoral actions in constructing personal identity,”
“Age-related differences in recognition in associative memory,” “Resistance to position
change, motivated reasoning, and polarization,” “The phenomenology of remembering
our moral transgressions,” “Moral memories and the belief in the good self,” “A reason-
based explanation for moral dumbfounding,” “When the unlikely becomes likely:
Qualifying language does not influence later truth judgments,” “Remembering moral
and immoral actions in constructing the self,” “Reasons probably won’t change your
mind: The role of reasons in revising moral decisions,” “Process-specific alliances (PSAs)
in Cognitive Neuroscience,” “Medial temporal, prefrontal, and parietal contributions to
the functional networks underlying item and source memory in older adults,” “No
evidence for unethical amnesia for imagined actions: A failed replication and extension,”
“I’m not the person I used to be: The self and autobiographical memories of the immoral
past,” “From hippocampus to whole-brain: The role of integrative processing in episodic memory retrieval,” “Emotional intensity in episodic autobiographical memory and counterfactual thinking,” “Resting state networks do not determine cognitive function networks: A commentary on Campbell and Schacter (2016),” “Comparative similarity and counterfactual plausibility,” “Hippocampal contributions to the large-scale episodic memory network predict vivid visual memories,” “Modularity in network neuroscience and neural reuse,” “Network modularity as a foundation for neural reuse,” and “Toward a more integrative cognitive neuroscience of episodic memory.”

Matthew was awarded the National Science Foundation Graduate Research Fellowship (2016), the Kenan Institute for Ethics Graduate Fellowship (2019), and the Lafitte Foundation Graduate Research Grant (2019 and 2020).