


On the Highway to Hell: Slippery Slope Perceptions in Judgments of Moral Character

Personality and Social
Psychology Bulletin
1–15
© 2023 by the Society for Personality
and Social Psychology, Inc
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01461672221143022
journals.sagepub.com/home/pspb


Rajen A. Anderson¹ , Benjamin C. Ruisch²,
and David A. Pizarro³

Abstract

Across four studies, we test the hypothesis that people exhibit “slippery slope” thinking in their judgments of moral character—that is, do observers judge that a person who behaves immorally will become increasingly immoral over time? In Study 1, we find that a person who commits an immoral act is judged as more likely to behave immorally and as having a worse character in the future than in the past. In Study 2, we find that it is the *commission* of an immoral act specifically—rather than merely attempting an immoral act—that drives this slippery slope effect. In Study 3, we demonstrate that observers judge the moral agent as more likely to commit acts of greater severity further in time after the initial immoral act. In Study 4, we find that this effect is driven by an anticipated corrupting of moral character, related to perceptions of the agent’s guilt.

Keywords

morality, slippery slope, intentionality, moral character

Received April 20, 2022; revision accepted November 14, 2022

“It can’t be overstressed how dangerous a person is if they can do that at the age of 16. What can they do at the age of 26 or 36?”

—Bob Arthur, *journalist, referring to a teenager convicted of murdering his girlfriend*

Lay theories of moral character often involve the notion that a character follows a trajectory or a path, with a narrative arc that can change course over time. For example, when an otherwise good person commits an immoral act, people often express concern that the person may be “on a slippery slope,” “falling from grace,” or being on a “highway to hell.” Similarly, a person who chooses *not* to transgress is “sticking to the straight and narrow” or “choosing the high road.” Central to these lay conceptualizations of morality is the importance of individual behaviors in shaping the trajectory of a person’s character: Moral behaviors propel a person toward future moral behavior, whereas immoral behaviors push a person toward subsequent immoral acts. Although these ideas are common in metaphors and lay discussion, there has not, to our knowledge, been any systematic empirical examination of how (im)moral acts shape the expected trajectory of a person’s future moral character.

In the present research, we examine how an agent’s immoral behavior shapes people’s predictions of the trajectory of that agent’s future behavior and moral character. Specifically, we test for slippery slope thinking in people’s expectations

regarding others’ moral character and behavior, answering the question of whether, when, and why people believe that committing a single immoral act will propel an agent toward committing other immoral acts in the future. Existing theories on character attribution are relatively agnostic regarding whether (and how) people judge that others will morally change over time, and thus we aim to fill this theoretical gap in whether people judge either change or stability in character based on that agent’s (im)moral behavior. We hypothesize that people do not simply expect consistency in moral character but instead predict that future character and behavior can change as a function of the actions taken by an individual.

The Slippery Slope

The slippery slope has primarily been discussed in the domains of philosophy and law in the context of argumentation. Slippery slope arguments (SSAs) are typically used to

¹Northwestern University, Evanston, IL, USA

²University of Kent, Canterbury, UK

³Cornell University, Ithaca, NY, USA

Corresponding Author:

Rajen A. Anderson, Northwestern University, 633 Clark St., Evanston, IL 60208, USA.

Email: rajen.anderson@kellogg.northwestern.edu

argue against changes to the status quo, often in legal matters (for a review on slippery slope arguments in judicial reasoning, see Schauer, 1985). Although they can take many forms, SSAs typically adhere to the following general structure: If *relatively innocuous Action A* occurs, *more negative Effect B* will occur in the future; so, to prevent the occurrence of B we should avoid performing A (Lode, 1999; Schauer, 1985). By connecting a small, seemingly inoffensive change to a more severe and egregious future outcome, SSAs seek to make the initial small change seem potentially dangerous or immoral, thereby discouraging others from enacting it. The persuasive strength of an SSA depends on the perceived similarity between the current action or behavior (*Innocuous Action A*) and the posited end state (*More Negative Effect B*; Corner et al., 2011; Volokh, 2003).

In the present research, we use the “slippery slope” as a metaphor to understand and describe the general patterns of lay cognition regarding how a person’s future moral character is expected to change following the commission of an immoral action. Building on the legal and philosophical literature, we propose that moral evaluations often exhibit a slippery slope pattern such that agents are expected to subsequently increase in their immorality after committing an initial immoral act. We use the term “slippery slope” to refer to the idea that when an individual commits an immoral act, observers will judge that individual as more likely to commit subsequent immoral acts *because of the very act that they committed*. In the present work, we are interested in people’s judgments of changes in moral behavior *regardless of whether those judgments accurately track changes in moral behavior* (for work on changes in moral behavior over time, see Baack et al., 2000; Garrett et al., 2016; Jennings, 2011; Tenbrunsel & Messick, 2004; Welsh et al., 2015). That is, we make no claims as to the accuracy of slippery slope judgments but are instead interested in deepening our understanding of the nature and process of moral evaluation and social prediction.

Predicting Moral Character and Behavior

In examining how observers predict the trajectory of an agent’s future character and behavior, we build on a growing literature in moral psychology focused on the evaluation of moral character (for reviews, see Hartman et al., 2022; Helzer & Critcher, 2018; Pizarro & Tannenbaum, 2011; Uhlmann et al., 2015). Psychological research has demonstrated that people make judgments not just of particular acts (e.g., “is this a moral action?”) but also of the people who commit those acts (e.g., “is this a moral person?”; Goodwin et al., 2014)—and these judgments of moral character play a central role in person perception (Hartley et al., 2016; Newman et al., 2014; Strohminger & Nichols, 2014). In sum, evaluations of an individual are heavily influenced by perceptions of that individual’s moral character.

Furthermore, a key function of social cognition and moral evaluation is not just to understand but also to predict other people’s behavior (Dennett, 1989; Saxe, 2009). When presented with an agent who commits an immoral act, what predictions do observers make about that agent in the future? For example, if we hear that someone shoplifts something small from a store today, what unethical behavior do we predict that person will perform in the future?

To our knowledge, there has been no systematic empirical examination of how people perceive that moral character and behavior will change over time. That is, although there has been research on how character judgments themselves can change (e.g., Cone & Ferguson, 2015; Gawronski & Bodenhausen, 2006; Mann & Ferguson, 2015; Rydell & McConnell, 2006; Siegel et al., 2018), we are aware of no empirical research on how a person’s (im)moral behavior influences judgments of how that person will morally change over time. The closest extant research we are aware of concerns how people evaluate existing change, finding that people are especially quick to diagnose moral decline versus moral improvement (Klein & O’Brien, 2016). However, much of this work focuses on changes that have already occurred, as opposed to predicting future changes. This empirical gap is particularly important given that existing models of moral attribution and character (e.g., Cushman, 2008; Helzer & Critcher, 2018; Shaver, 1985; Uhlmann et al., 2015; Weiner, 1995) are relatively agnostic on the question of when, whether, and why people will predict change in others’ moral character over time.

We formulated two competing hypotheses regarding whether observers predict changes in future immoral behavior and character. One possibility is that observers will evaluate an agent who performs an immoral act as simply being a “bad person”—someone who is consistently immoral in their character and behavior. Observers may simply believe that bad people perform bad acts and that the degree of a person’s (im)morality does not necessarily change over time (e.g., as a result of committing individual immoral acts). This prediction is consistent with much past research and theory, which generally finds that observers predict that a person’s future behavior will be similar to their past behavior (e.g., Baxter & Goldberg, 1987; Buehler et al., 1994; Helzer & Dunning, 2012; Kelley, 1967; Vazire & Mehl, 2008). In short, people often believe that how someone is *now* is how they will be *in the future* (e.g., Quoidbach et al., 2013). Therefore, according to this perspective, if an agent commits an immoral act, this should simply affect people’s *global* evaluations of that person’s moral character.

The second possibility is that people’s moral character judgments may exhibit a slippery slope effect—consistent with lay theories of moral character and the predictions that we have outlined above. According to this perspective, the anticipated future trajectory of an agent’s moral character and behavior will change based on that agent’s behavior in the present. Specifically, an agent who commits an immoral

act should be seen as subsequently being more likely to perform immoral acts in the future, relative to the past.

While this predicted change could be the result of several different factors (e.g., being rewarded for unethical behavior), we predicted and tested one potential explanation: that of perceived changes in the agent's character. That is, observers will predict that an agent who commits an unethical act will experience psychological changes, becoming desensitized to that unethical act and experiencing less guilt for doing so. Past research has highlighted the important role of an agent's perceived tendencies and emotions in judgments and predictions of that agent's character and behavior (e.g., Ames & Johar, 2009; Anderson et al., 2021; Critcher et al., 2013; Pizarro et al., 2003). Such psychological changes will then make subsequent unethical behavior more likely, creating a perceived feedback loop. This "corruption" of conscience will then be used to predict how that agent will behave in the future, creating the hypothesized slippery slope in judgments of moral behavior.

In the present studies, we fill these empirical and theoretical gaps by examining the inferences that observers make when an agent commits an immoral act. We hypothesize that observers will perceive a single immoral act to be a signal of future negative moral change such that this immoral act will worsen an agent's moral character and increase the likelihood that the agent will commit further immoral acts in the future.

Predictions From a Slippery Slope Effect

In sum, past research on the moral character has elucidated many of the processes underlying how people make judgments about others' character. This work has shown that people are quick to make global judgments about others' moral character based on the actions that they perform—even when only limited information is provided—and that an individual who behaves immorally is perceived to have a more negative moral character than one who behaves morally, neutrally, or about whom no information is provided (e.g., Ames & Johar, 2009; Chakroff et al., 2017; Critcher et al., 2013; Klein & O'Brien, 2016; Reeder & Brewer, 1979; Uhlmann et al., 2015). This research has also revealed that more nuanced considerations can also shape people's judgments of others' moral character and behavior, such as the specific domain of an (im)moral act (e.g., harm versus purity; Chakroff et al., 2017; Masicampo et al., 2014) and an agent's emotions surrounding a behavior (e.g., Ames & Johar, 2009; Barasch et al., 2014; Berman et al., 2015; Critcher et al., 2013, 2020; Pizarro et al., 2003).

Importantly, however, these past theories and research do not make clear predictions about how people expect moral character to change over time. That is, although these models predict that a person who behaves immorally will be seen as having a generally more negative moral character, the degree

of this negativity is not expected to change over time. In other words, according to these past accounts, behavioral information simply *reveals* a person's stable underlying character. Thus, a person who commits immoral *Action A* should be perceived as *generally* more likely to commit immoral *Action B* (given that the person is generally of worse moral character), but the likelihood of committing B should not be any higher *following* the commission of A than preceding it. Conversely, our slippery slope perspective on moral judgment predicts that the commission of A should increase the perceived likelihood that an agent would commit subsequent (relative to past) immoral acts. According to this slippery slope account, behavioral information not only reveals a person's character but also shapes perceptions of the direction in which future character will change.

In the current research, we tested four broad hypotheses derived from the slippery slope theoretical framework outlined above.

Hypothesis 1: Observers will predict that an agent who commits an immoral act will be more likely to commit other immoral acts following (relative to before) the commission of that immoral act. In other words, we expect that observers will see this act as signaling that the agent is proceeding down a "slippery slope" into further future immoral behavior.

Hypothesis 2: Perceptions and expectations of (im)moral change will specifically depend on an agent *committing* an immoral act. In other words, immoral behavior specifically—over and above factors such as immoral thoughts and/or intentions—will elicit these slippery slope judgments. Furthermore, it is not simply the case that perceivers expect that everyone is more likely to commit immoral acts in the future relative to the past but that this effect will hold only for people who are known to have acted immorally.

Hypothesis 3: Drawing on past research showing that moral character judgments are sensitive to different "domains" of (im)moral behavior (e.g., Chakroff et al., 2017; Chakroff & Young, 2015; Corner et al., 2011; Everett et al., 2016), we hypothesized that slippery slope judgments would be sensitive to the perceived relation between the initial immoral act and the predicted future acts. That is, observers would not simply assume that an immoral agent would behave more immorally in general for all types of immoral behaviors. Instead, we expected that perceivers' estimates of the likelihood that an immoral agent would subsequently commit a given immoral act would depend on the similarity between that act and the initial immoral act, with more similar acts (e.g., those closer in severity) being judged as more likely. We focus on the severity of the act as one potential dimension by which acts are similar to each other. We hypothesized that relatively minor infractions will be seen as more likely than relatively major infractions after the initial

infraction. However, consistent with the theory outlined earlier, all infractions should be seen as becoming more likely.

Hypothesis 4: We predicted that one psychological mechanism underlying these slippery slope perceptions would be inferences of attenuated guilt on the part of the agent. We hypothesized that the reason that observers would expect an agent to perform subsequent immoral behaviors following an immoral act is because—akin to the mechanisms shown to underlie real-world increases in immoral behavior (Garrett et al., 2016; Welsh et al., 2015)—observers would believe that agents undergo a moral “corruption” or “numbing of their conscience.” Because performing an immoral act should lead agents to experience less guilt in response to subsequent immoral behavior, this should make it easier for them to commit future immoral acts. Accordingly, directly manipulating perceptions of this change in conscience should influence the predictions that observers make regarding an agent’s future behavior. We contrast this “corrupted character” explanation with a simple cost/benefit account by which individuals judge that agents commit subsequent bad acts because of the net positive outcomes they experience from committing an immoral act (e.g., financial or reputational benefits that outweigh the punishment they receive).

The Current Research

We conducted four studies to test the above hypotheses regarding slippery slope perceptions in moral character judgments. In Study 1, we provide an initial test of whether people’s judgments of others’ moral character exhibit a slippery slope pattern. In Study 2, we examine the effect of actually committing (vs. attempting but not committing) an immoral act. If perceptions of corrupted moral character are the psychological mechanism underlying the slippery slope effect, then the commission of an immoral act should increase expectations of future negative behavior, *even when controlling for intentions to engage in this behavior*. In Study 3, we examine the scope of the slippery slope effect by testing whether observers’ predictions about an agent’s future immoral behavior are shaped by the severity of future immoral behaviors. Study 3 also provides evidence for our predicted mechanism, whereby an immoral agent is judged as becoming less prone to guilt after immoral behavior. Finally, in Study 4, we test whether future transgressions will be expected to be of increasing severity, and we experimentally test two potential psychological mechanisms—moral corruption versus utilitarian consequences—for the slippery slope effect in moral judgments.

Throughout these studies, we test and verify the generalizability of the slippery slope effect by using a variety of experimental stimuli depicting different moral agents and various types of immoral actions. In all studies, we report all manipulations, measures, and exclusions. The sample size

and stopping rules for each study were determined before data collection, and we analyzed data only after all data collection was completed. We conducted a pilot study (reported in our OSF link) to inform our power analyses. For this pilot study, we conducted a power analysis (using G*Power version 3.1; Faul et al., 2009) for 80% power to detect an effect size of $d = .2$ (our estimated effect size based on a review of the literature). This resulted in a recommended sample size of 199, which we set as the target sample size for the pilot study. For subsequent studies, we based our power analyses on 80% power to detect an effect of the size observed in the pilot study. We collected this minimum target sample size and exceeded it when possible to maximize power. We also conducted four Supplemental Studies (reported in our OSF link) to further examine various aspects of the slippery slope effect. All materials, data, analysis scripts, reporting on the pilot study and supplemental studies, and preregistration documentation are available at <https://osf.io/m9qwp>.

Study 1

Study 1 served as an initial demonstration of Hypothesis 1—that an agent will be expected to become more immoral after the commission of an immoral act. We sought to compare judgments made of someone who behaves immorally to a control agent who is faced with a similar situation but does not behave immorally (Hypothesis 2). This allowed us to test and rule out one potential explanation—that it is having a negative experience that *provoked* the immoral act (in this case, having an inconsiderate person park too close to one’s own car) that lead to the subsequent changes in perceived moral character and behavior, or the possibility that people may simply believe that other people in general tend to become more immoral over time, regardless of their prior behavior. Consistent with the predictions outlined earlier, we hypothesized that observers would specifically perceive these changes in moral character and behavior for an agent who actually commits an immoral act (compared with someone in a similar situation who does not commit that immoral act).

Method

Participants. Based on the results of our pilot study, we conducted a power analysis for 80% power to detect an effect size of Cohen’s $d = .35$, the effect size on the composite morality measure from the pilot study. This analysis recommended a sample size of 52 participants per condition. In this and all subsequent studies, we ensure this minimum sample size. To further increase statistical power in this study, we recruited 265 participants from Amazon’s Mechanical Turk (more than $N = 131$ per condition; Buhrmester et al., 2011). As specified in our preregistration, we excluded 11 participants for failing the attention check, leaving a final sample size of 254 participants (56% female, $M_{age} = 38.7$).

However, none of our conclusions are substantively altered if these participants are included in analyses.

Design. We randomly assigned participants to one of two conditions. All participants read a vignette describing a man returning to his car in a mall parking lot to find another car parked too close to his own. We randomized the name of the man between participants from a bank of common U.S. male names. Participants in the *immoral act* condition read that the man became angry and used a screwdriver he had in the trunk of his car to punch a hole in the tires of the other car before driving away. Participants in the *no immoral act* condition read a modified version of this vignette in which the man—although very angry that someone could be so inconsiderate—simply got in his car and drove away.

After reading the vignette, we asked participants two sets of questions about Henry, five questions referring to Henry in the past and five questions referring to Henry in the future. For both the past and future, participants were asked to rate Henry's moral character (how good or bad of a person is Henry; from 1 = *extremely bad* to 9 = *extremely good*), to rate the likelihood that Henry would do something unethical or illegal (from 1 = *Not likely at all* to 9 = *Very likely*), and to judge the likelihood that Henry would commit five specific moral infractions (damage someone else's car; breaker someone else's property; stealing something from someone; assaulting someone; and driving while intoxicated; from 1 = *not likely at all* to 9 = *very likely*). Items were presented in random order. We also randomized the order in which participants answered the past and future questions.

After making judgments of Henry's past and future moral character and behavior, participants completed an attention check in which they were asked to recall what happened in the story by selecting one of several options. Our inclusion criterion for the attention check was selecting the correct answer to this question (i.e., that the man punctured the other car's tire in the *immoral act* condition or that the man drove away without doing anything in the *no immoral act* condition).

Results and Discussion

Per our preregistration, we averaged all moral judgment items together to create a composite moral evaluation score (reverse-scoring the moral character item) for the past ($\alpha = .96$) and the future ($\alpha = .96$). We found that participants in the *immoral act* condition exhibited a slippery slope effect in their judgments, rating the agent as being more immoral in the future ($M = 6.43, SD = 1.37$) versus past ($M = 6.13, SD = 1.43$), $t(132) = 4.89, p < .001$, Cohen's $d = .42$.

We next deconstructed this composite score to look at each individual component: the moral character item, the item assessing the general likelihood of immoral behavior, and the composite likelihood judgments of the five specific immoral behaviors ($\alpha_{\text{past}} = .94, \alpha_{\text{future}} = .95$). Consistent with our predictions, there was a significant difference in the

item assessing the man's perceived general likelihood of committing immoral behavior such that participants saw him as more likely to behave immorally in the future ($M = 7.37, SD = 1.69$) versus past ($M = 7.11, SD = 1.79$), $t(132) = 2.46, p = .02$, Cohen's $d = .21$.

Similarly, there was also a significant difference in judgments of the specific immoral acts, such that participants perceived a greater likelihood that the man would commit these various immoral acts in the future ($M = 6.26, SD = 1.50$) versus past ($M = 5.89, SD = 1.58$), $t(132) = 5.04, p < .001$, Cohen's $d = .44$. Unexpectedly, we found no significant difference in judgments of moral character in the future ($M = 3.60, SD = 1.38$) versus past ($M = 3.70, SD = 1.35$), $t(131) = 1.30, p = .19$, Cohen's $d = .11$, although the difference was directionally consistent with the behavior judgments.

To determine whether the observed slippery slope effect was truly the result of the agent having committed an immoral act, we next compared judgments of the man who actually committed the immoral act (i.e., the *immoral act* condition) with judgments of the man who had an identical experience and became very angry, but ultimately did not commit the immoral act (i.e., the *no immoral act* condition). To test this question, we conducted a mixed-design analysis of variance (ANOVA) on the composite moral evaluation score (as specified in our preregistration), with condition (*immoral act* versus *no immoral act*) entered as a between-subjects factor and time (past versus future) entered as a within-subjects factor. We found a significant main effect of condition on moral evaluations, such that the *immoral act* agent was rated as being more immoral than the *no immoral act* agent, $F(1, 252) = 593.95, p < .001, \eta_p^2 = .70$. Critically, however, we also found that this main effect of condition was qualified by a significant condition \times time interaction, $F(1, 252) = 11.38, p = .001, \eta_p^2 = .04$.

As described earlier, there was a significant difference between past ($M = 5.85, SD = 1.45$) and future ($M = 6.16, SD = 1.38$) judgments in the *immoral act* condition, $t(132) = 4.89, p < .001$, Cohen's $d = .42$; however, there was no such difference between past ($M = 2.21, SD = 1.16$) and future ($M = 2.21, SD = 1.19$) judgments in the *no immoral act* condition, $t(120) = 0.09, p = .93, d < .001$. Thus, it is not the case that the slippery slope effect was due to factors such as being the victim of a transgression or becoming angry, nor is it the case that observers think that everyone will behave more immorally in the future than in the past. Rather, our results suggest that it was the commission of an immoral act that led to these slippery slope perceptions. Consistent with Hypotheses 1 and 2, single acts of immorality appear to change the anticipated trajectory of an agent's moral character and behavior, leading them to be more likely to behave immorally in the future.

Study 2

In Study 2, we replicate and extend Study 1 by separately examining the roles of *attempting* and *committing* an immoral

act (Hypothesis 2). That is, in Study 1, the *immoral act* agent and the *no immoral act* agent differed not only in whether they committed an immoral act but also in whether they attempted to commit that act. The agent who commits the unethical action both *intentionally attempts* that action (i.e., has the desire and foreknowledge to perform it; Malle & Knobe, 1997) and then *successfully commits* that action. Conversely, the *no immoral act* agent does neither. Thus, it is unclear whether and to what degree actually committing an immoral act—versus attempting to commit an immoral act—gives rise to the slippery slope effect.

Previous research has shown that observers are sensitive to both the intentions of an agent and the consequences of an act (e.g., Cushman, 2008; Martin & Cushman, 2016; Nelson, 1980; Vaish et al., 2010) and that both can independently contribute to judgments of (im)morality. Accordingly, in this study, we compared predictions of three agents: a morally neutral agent who makes no attempt at an immoral act, an agent who intentionally attempts an immoral act but is ultimately unable to complete the act, and an agent who intentionally attempts an immoral act and is able to complete the act. We predicted that an individual who attempted (but did not complete) an immoral act would be expected to undergo a larger change in future moral character and behavior than a morally neutral agent. However, we also predicted that committing an immoral act—even over and above attempting to commit that act—would lead to greater expectations of future immoral character and behavior. Although this latter hypothesis was not a critical prediction of our theoretical framework, we viewed it as providing a particularly stringent test of the idea that immoral acts themselves are perceived as indelibly “corrupting” an agent’s moral character (Hypothesis 4).

Method

Participants. We recruited 342 participants from MTurk. We excluded 40 participants for failing the attention check, leaving 302 participants (52% female, $M_{\text{age}} = 38.50$) in our final sample. However, none of our conclusions are substantively altered if all participants are included in the analyses.

Design. We randomly assigned participants to one of three conditions. In all conditions, participants read a story about a high school student nervous about an upcoming exam who considers cheating (something he has never done before). In the *no-attempt* condition, the student considers cheating but decides against doing so and proceeds to take the exam to the best of his ability. In the *immoral attempt* condition, the student decides to cheat on the exam and writes notes in the palm of his hand. However, the ink becomes smeared from sweat on his palms, and he is unable to read the notes and thus unable to cheat. The student is then said to have no choice but to take the test without cheating. The *immoral act* condition was identical to the *immoral attempt* condition,

with the following critical difference: although the notes become somewhat smeared from the sweat on his palms, he can still read the notes and thus cheats on the exam.

Participants then answered questions about the student, assessing judgments of his moral character and behavior in both the future (one year after the event) and the past (one year before the event). For both past and future, participants indicated how good or bad of a person the student was (from 1 = *extremely bad* to 9 = *extremely good*) and how likely it was that the student would do something unethical or illegal (from 1 = *not likely at all* to 9 = *very likely*). We also asked participants to answer six questions about how likely it was that the student would perform a variety of specific acts (change the rules of a game halfway through to win; not tell a romantic partner he tested positive for an STI; plagiarize someone else’s work; lie about something; cheat on a homework assignment; and take a shortcut in a race; from 1 = *not likely at all* to 9 = *very likely*).

After making predictions about the agent, participants completed an attention check asking them to recall the behavior of the agent in the story (i.e., whether he cheated on the test, wanted to cheat but could not, did not attempt to cheat, or story did not say). Our inclusion criteria for the attention check was if a participant selected the correct answer for their assigned condition.

Results and Discussion

To test for the effects of condition on evaluations of the agent’s past versus future character, we conducted a 3 (act: *no attempt*, *immoral attempt*, *immoral act*) \times 2 (time: *past versus future*) mixed-model ANOVA. Consistent with our previous studies, we tested this model on each of our three primary dependent measures: (a) global judgments of moral character (i.e., how good or bad a person the target is), (b) the target’s general likelihood of doing something unethical or illegal, and (c) the averaged perceived likelihood that the agent would commit the six specific immoral behaviors ($\alpha_{\text{past}} = .93$, $\alpha_{\text{future}} = .94$).

In keeping with our predictions, there was a significant condition \times time interaction on all three dependent measures: judgments of how good or bad a person the target was, $F(2, 295) = 7.16$, $p = .001$, $\eta_p^2 = .05$, likelihood of doing something unethical or illegal, $F(2, 299) = 16.51$, $p < .001$, $\eta_p^2 = .10$, and the average likelihood of performing the specific immoral acts, $F(2, 299) = 22.05$, $p < .001$, $\eta_p^2 = .13$.

We first compared judgments of the agent who did not attempt an immoral act with judgments of the agents who committed and/or attempted to commit an immoral act. There were significantly larger changes between past and future evaluations of the target in the *immoral act* and *immoral attempt* conditions relative to the *no attempt* condition (Table 1). These findings show that perceivers view committing an immoral act—or even attempting to commit one—as signaling a change in the trajectory of an agent’s future moral character and behavior.

Table 1. Descriptive Statistics for Study 2.

Dependent Measures	Condition	Past <i>M</i> (<i>SD</i>)	Future <i>M</i> (<i>SD</i>)	95% CI of Mean Difference
Good/Bad Person (1 = Extremely Bad, 9 = Extremely Good)	No Attempt	7.05 (0.86)	7.10 (0.96)	[-.10, .20]
	Immoral Attempt	6.07 (1.51)	5.63 (1.44)	[-.68, -.20]
Likelihood of unethical behavior (1 = Very Unlikely, 9 = Very Likely)	Immoral Act	5.85 (1.40)	5.24 (1.38)	[-.89, -.34]
	No Attempt	2.67 (1.37)	2.51 (1.22)	[-.38, .05]
Average likelihood of performing specific unethical behaviors (1 = Very Unlikely, 9 = Very Likely)	Immoral Attempt	3.81 (2.03)	4.77 (2.08)	[.59, 1.33]
	Immoral Act	4.64 (2.10)	5.82 (1.71)	[.83, 1.53]
	No Attempt	2.78 (1.36)	2.64 (1.22)	[-.29, .02]
	Immoral Attempt	4.16 (1.92)	4.92 (1.92)	[.45, 1.08]
	Immoral Act	4.69 (1.83)	5.94 (1.57)	[.95, 1.55]

Note. Participants rated the agents who successfully committed and who attempted but failed to commit an immoral act as being worse in moral character and more likely to commit other unethical behaviors (both in general and as an average of specific acts) in the future than in the past. Conversely, there was no significant change in ratings of the agent who did not attempt the immoral act.

To determine whether there was a unique contribution of committing (versus only attempting) an immoral act in giving rise to the slippery slope effect, we next conducted a follow-up 2×2 mixed-model ANOVA comparing the past-future difference between the *immoral act* and the *immoral attempt* conditions only (excluding the *no attempt* condition). We did not observe a significant effect for either judgments of how good or bad a person the target was, $F(1, 216) = 0.90, p = .34, \eta_p^2 = .004$, or for general judgments of whether the agent would do something unethical or illegal, $F(1, 217) = 0.72, p = .40, \eta_p^2 = .003$. However, we did observe a significant effect on perceptions of whether the target would perform the six concrete immoral acts, such that observer's viewed the agent who actually committed the immoral act as more likely to behave immorally in the future, $F(1, 217) = 4.79, p = .03, \eta_p^2 = .02$.

Together, these results provide additional evidence for the slippery slope effect in moral judgment and show that even attempting an immoral act—even if that act is ultimately not committed—is sufficient to elicit these expectations of future moral decline. Furthermore, these findings also provide some tentative support for the possibility that committing the immoral act (over and above intentionally attempting one) may also independently play a role in eliciting the slippery slope effect, although these effects may emerge more strongly for judgments of future behavior than global character judgments (see Supplemental Study 2 on our OSF link for an additional test of the effect of commission on judgments of future behavior).

Study 3

In Study 3, we aimed to answer three questions regarding the nature of slippery slope moral judgments. First, how do observer judgments differ between different *future* time points? Both Studies 1 and 2 employed a “past versus future”

design, comparing how observers thought the agent was in the past before the act and how observers thought the agent would be in the future after the act. Study 3 employs a new design where participants make judgments of the agent at two points in the future. We predicted that observers would make different judgments for an immoral agent (vs. a control agent) a shorter time into the future and a longer time into the future after an immoral act.

Second, what kind of future acts do observers predict? In Study 3, we aimed to address one aspect, that of severity (Hypothesis 3). If an agent steals something, would observers distinguish between the agent then stealing something of minor value versus something of great value? Would these judgments change when considering different points of time in the future? Turning to the colloquial notion of the slippery slope, the construction of the argument is not just that additional negative events will occur in the future but that those events will be, on average, *worse* than what has come before (Lode, 1999; Schauer, 1985; Volokh, 2003). By examining participants' judgments of the likelihood of committing acts of different severities, we can examine with greater fidelity how the slippery slope effect works. For examination of another dimension, the domain of the acts, see Supplemental Study 3 on our OSF link.

Third, why might observers judge that an agent will be more likely to commit immoral acts? Our hypothesized mechanism—that of a “corrupted” moral character—is consistent with our existing studies, but those designs did not directly examine it. In this study, we aimed to examine the evidence for the hypothesis that at least one reason observers judge that an agent will be more likely to commit other immoral acts in the future is due to changes in the agent's underlying moral and emotional reactions. Such a pattern would be in keeping with findings suggesting that real-world immoral behavior can increase over time because people become desensitized to committing immoral acts (e.g., cheating; Garrett et al., 2016;

Welsh et al., 2015). We predicted that the immoral agent (vs. a control agent) will be seen as experiencing less guilt, but that this effect would be even stronger later into the future (Hypothesis 4).

Method

Participants. We recruited 402 U.S. participants through Prolific.co, an online data collection service (Palan & Schitter, 2018). We excluded nine participants for failing an attention check, leaving a final sample of 393 participants (47% female, $M_{\text{age}} = 33.10$) in our final sample. However, none of our conclusions are substantively altered if all participants are included in the analyses. Based on the observed effects from previous studies, this sample size provides power over 0.95.

Design. We randomly assigned participants to one of two conditions (action: *immoral act*, *no immoral act*) between-subjects design. Participants in both conditions read a vignette about a person named Paige riding on a bus who notices a fellow passenger sleeping with money visible from her purse. In the *immoral act* condition, Paige steals the money from a fellow passenger; in the *no immoral act* condition, Paige considers doing so but does not steal the money. In both conditions, the vignette ends with Paige exiting the bus and no one noticing what Paige has done.

After reading the vignette, participants proceeded to a new screen, where we asked them think about how Paige would be in the future after the event described: “Specifically, we want you to think about what Paige will be like both one week after this event as well as one year after this event.” Participants then completed two sets of questions, one set about Paige 1 week after the event and one set about Paige 1 year after the event. The presentation order of the sets was counterbalanced between participants. For each set, we asked participants three questions about Paige’s likely future behavior, differing in severity: if she was in the right situation, how likely is it that she would steal something of (a) “minor value (up to \$50),” (b) “moderate value (between \$50 and \$300),” and (c) “great value (more than \$300)” 1 week/year after this event (from 1 = *not likely at all* to 9 = *very likely*)? We also asked two questions about Paige’s predicted self-conscious moral emotions, asking both about her guilt and shame if she stole something else 1 week/year after this event (from 1 = *none at all* to 9 = *a great deal*). The two questions assessing guilt and shame were very highly correlated with each other, so we averaged them together to form a composite index of Paige’s predicted self-conscious emotions after stealing again 1 week ($r_{\text{week}} = .89$) and 1 year into the future ($r_{\text{year}} = .91$).

Results and Discussion

Likelihood Judgments. To test for differences in participants’ likelihood judgments of future immoral, we conducted a mixed linear model, with the condition as a between-subjects

factor and time (1 week vs. 1 year) and severity of the future act (minor value vs. moderate value vs. great value) as within-subjects factors (see Table 2 for descriptive statistics). Replicating our past studies, we found a significant main effect of condition, whereby participants made higher likelihood estimates in the *immoral act* condition than the *no immoral act* condition, $F(1, 391) = 216.89, p < .001, \eta_p^2 = .36$. There was also a significant main effect of time, such that participants made higher likelihood estimates for the agent’s behavior 1 year into the future compared with 1 week into the future, $F(1, 391) = 37.64, p < .001, \eta_p^2 = .09$. There was also a main effect of severity, such that participants made highest likelihood estimates for minor immoral acts, followed by moderate immoral acts, and least for great immoral acts, $F(2, 390) = 216.89, p < .001, \eta_p^2 = .36$.

These main effects were qualified by the predicted condition \times time interaction. Estimates of future immoral behavior increased when considering the agent 1 year in the future (vs. 1 week in the future), but more strongly when the agent had already committed an immoral act, $F(1, 391) = 18.97, p < .001, \eta_p^2 = .05$. Using a new design, this effect replicates the results from our previous studies: observers predict that immoral agents will become more likely of committing other immoral acts later in time compared with earlier in time. Consistent with the slippery slope prediction, the immoral agent is judged as becoming *more immoral* over time.

In addition, there was a significant condition \times severity interaction on likelihood judgments, $F(2, 390) = 30.27, p < .001, \eta_p^2 = .13$. For participants in the *no immoral act* condition, there were minimal differences in judged likelihood between stealing something of minor value, moderate value, or great value. For participants in the *immoral act* condition, these differences were much stronger, whereby participants estimated the highest likelihood for stealing something of minor value, then moderate value, and then great value.

There was also a significant time \times severity interaction, $F(2, 390) = 3.34, p = .04, \eta_p^2 = .02$. Given that this interaction was not predicted, and not very strong, we caution against reading too much into it. Finally, there was a nonsignificant condition \times time \times severity interaction on likelihood judgments, $F(2, 390) = 0.02, p = .82, \eta_p^2 < .01$.

Taken together, these results provide strong support for the slippery slope effect. We found that participants think that immoral agents will become more immoral further into the future from the initial immoral event. In addition, while immoral agents are judged as most likely to commit future immoral acts of relatively lower severity, we found that the judged likelihood of acts of greater severity *also* increased further into the future. That is, the immoral agent is judged as not simply repeating the same act over and over again but that acts of greater severity become more likely as well.

Predicted Guilt and Shame. To test for differences in participants’ judgments of future guilt and shame from stealing, we conducted a mixed linear model, with the condition as a

Table 2. Descriptive Statistics for Study 3.

	Immoral act condition			No immoral act condition		
	1 week	1 year	95% CI of Mean Difference	1 week	1 year	95% CI of Mean Difference
Likelihood of minor value theft	5.76 (2.30)	6.43 (2.06)	[.38, .98]	2.89 (1.87)	2.96 (1.87)	[-.10, .25]
Likelihood of moderate value theft	4.97 (2.27)	5.84 (2.30)	[.59, 1.15]	2.69 (1.69)	2.90 (1.75)	[.02, .39]
Likelihood of great value theft	4.22 (2.32)	5.06 (2.41)	[.56, 1.11]	2.61 (1.81)	2.73 (1.83)	[-.06, .32]
Guilt/Regret	5.08 (1.96)	4.17 (1.99)	[-1.16, -.67]	6.97 (1.75)	6.80 (1.81)	[-.36, .01]

Note. Participants rated the agents who stole, compared to the control agent who did not steal, as being more likely of committing future acts of different magnitude, and this difference increased further into the future. This was associated with a corresponding change in the agent's predicted guilt and regret for immoral behavior.

between-subjects factor and time (1 week vs. 1 year) as a within-subjects factor (see Table 2 for descriptive statistics). We found a significant main effect of condition, $F(1, 391) = 171.25, p < .001, \eta_p^2 = .30$, and a significant main effect of time, $F(1, 391) = 47.81, p < .001, \eta_p^2 = .11$. These main effects were qualified by the predicted condition \times time interaction, $F(1, 391) = 21.96, p < .001, \eta_p^2 = .05$. When evaluating the control agent who considered but did not commit the immoral act, there was relatively little difference in predictions of her guilt and shame from stealing something 1 week versus 1 year into the future. However, when considering the agent who did commit the immoral act, participants anticipated that she would feel less guilt and shame if she stole something 1 year into the future compared with 1 week into the future.

This provides evidence that part of the mechanism behind the slippery slope effect is due to changes in mental states related to moral characters, like guilt and shame from committing an immoral act. Overall, the immoral agent was judged as experiencing less guilt and shame than the control agent. However, this difference increased when thinking about how the agent would feel further into the future. This suggests that observers believe that immoral agents become morally “corrupted” over time, no longer experiencing self-conscious emotions to the same degree after committing an immoral act.

Study 4

In our final study, we had two primary aims. Our first aim was to provide a more nuanced examination of observers' predictions regarding how an immoral agent's behavior would continue to change in the future. Specifically, we were interested in testing the prediction that an agent who commits an immoral act would specifically be expected to subsequently commit other immoral acts that would *increase in severity over time* (versus, e.g., committing immoral acts of the same degree of severity, or immoral acts that vary randomly/unpredictably in their severity). This would replicate and extend the results from Study 3 using a different methodology.

Our second aim was to test two possible psychological mechanisms that may underlie the slippery slope effect in moral judgment. The first potential mechanism we wished to test was that agents are perceived as likely to commit increasingly immoral acts over time because they are positively reinforced for their immoral behavior. That is, people may hold the lay belief that immoral behavior brings about rewards (e.g., resource or reputational benefits) that make the agent more likely to commit future acts for further gain. If so, then eliminating or diminishing the overall positive outcomes for the agent—for example, by punishing the agent for the immoral act—could attenuate the slippery slope effect. If the behavior is no longer “profitable” for the agent (i.e., the potential rewards no longer exceed the potential costs), observers may then predict that the agent will no longer engage in the behavior.

The second potential mechanism that we wished to examine was the “corruption hypothesis” that we outlined above (Hypothesis 4): that agents are perceived as increasingly likely to commit future immoral behavior because transgressing becomes affectively easier for them. If true, then undercutting this perception of corrupted character—for example, by indicating that the agent experienced regret or guilt after the immoral act—could attenuate the slippery slope effect. While Study 3 provided some initial evidence, with Study 4 we wanted to experimentally manipulate the agent's mental state after committing the unethical act.

Method

Participants. We recruited 807 participants (55% female, $M_{\text{age}} = 38.5$) from MTurk. We did not include an attention check in this study and did not exclude any participants from the analyses.

Design. Participants were first told that they would read a short story and answer some questions about it. Participants were randomly assigned to one of four conditions, based on a 2 (punishment: *punishment, no punishment*) \times 2 (regret: *regret, no regret*) design. In all conditions, participants were given a general description of an immoral

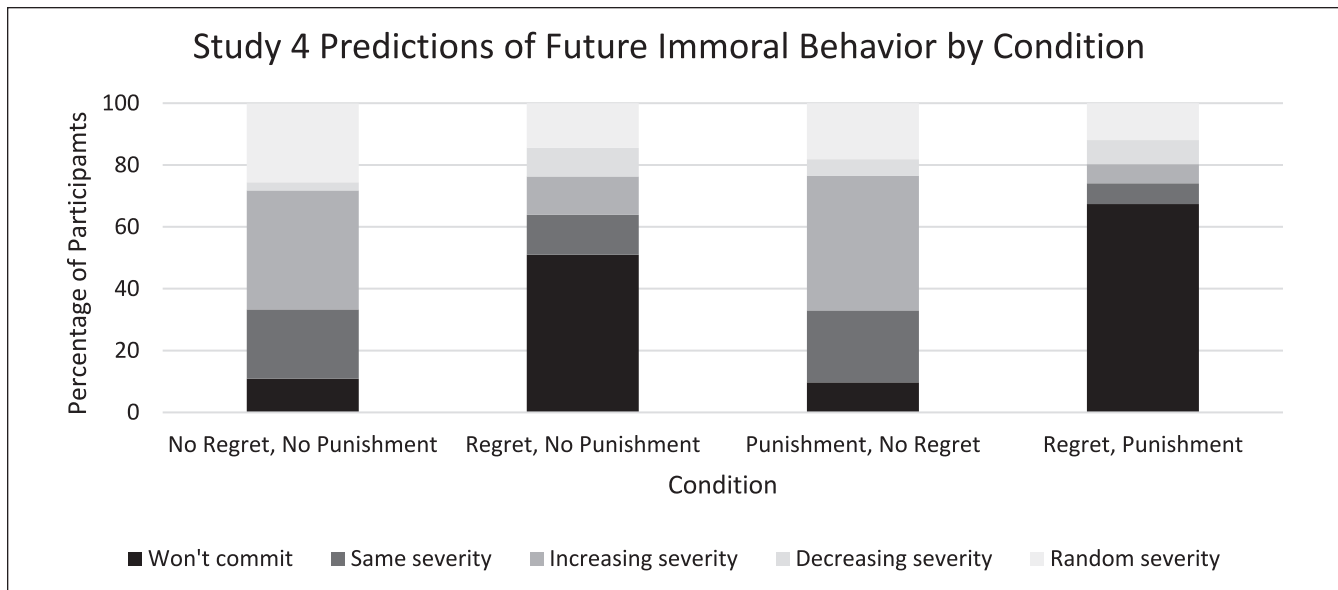


Figure 1. Study 4 Results, Separated by Condition and Predicted Outcome.

Note. When an immoral agent experienced no regret, participants most often predicted that the agent would commit immoral acts of increasing severity, even if they had been punished for their immoral behavior. When the agent did experience regret, participants most often predicted that the agent would cease committing immoral acts.

actor, asking them to evaluate a person who does something morally questionable, such as committing minor theft. The ending of the vignette differed by condition, with the agent either being caught and punished for this behavior (*punishment* condition) or experiencing no punishment (*no punishment* condition). The vignettes also differed as to whether the person who committed the immoral act felt regret and guilt about performing this behavior (*regret* condition) or experienced no regret or guilt (*no regret* condition).¹

After reading the vignette, participants were asked to make judgments about the agent's future (im)moral behavior. They were provided with a list of possible future behavioral trajectories and were asked to choose the one they viewed as most likely: "This person won't commit any other immoral acts, or will only do so very infrequently"; "This person will commit other immoral acts of the same severity over time"; "This person will commit other immoral acts with increasing severity over time"; "This person will commit other immoral acts of decreasing severity over time"; "This person will commit other immoral acts, but the severity of the acts will be random."

Results and Discussion

To test the effects of punishment and regret on predictions of future behavior, we conducted a chi-square test on participants' behavior judgments, based on how many participants in each condition selected each response option. We found that the conditions significantly differed in the degree to which each possible future pattern of behavior was selected,

$\chi^2(12, N = 806) = 274.75, p < .001$ (see Figure 1). Consistent with our previous studies,

When the agent was not said to have experienced punishment or regret, participants most often thought that the agent would commit other immoral acts of increasing severity over time. In other words, observers viewed this immoral act as signaling a change in the trajectory of the agent's future moral character such that the agent would commit immoral acts of increasing severity over time.

We next examined the effects of punishment and regret on slippery slope perceptions. In terms of the most compelling tests of the competing hypotheses, we first focused on each factor separately and then looked at their combined effect. Intriguingly, we found that punishment alone (in the absence of regret) did *not* significantly attenuate the slippery slope effect: There was no significant difference between the control condition and the punishment-only condition, $\chi^2(4, N = 419) = 5.38, p = .25$. However, consistent with our predictions, we found a significant main effect of regret, $\chi^2(4, N = 386) = 94.04, p < .001$, such that the agent who experienced negative emotion after committing an immoral act was viewed as less likely to commit future immoral acts. Interestingly, this effect of regret was further heightened when combined with punishment such that the agent who experienced both regret and punishment was judged as even less likely to commit future immoral acts than the agent who experienced only regret with no punishment, $\chi^2(4, N = 387) = 12.75, p = .01$.

These results conceptually replicate the slippery slope effect using different materials and dependent measures while providing additional nuance to our previous findings.

They reveal that observers specifically anticipate that an agent who commits an immoral act will commit increasingly severe immoral behaviors in the future, rather than, for example, immoral acts of similar or unpredictable severity. Furthermore, these results provide support for our hypothesized corruption mechanism (Hypothesis 4), suggesting that an explicit signal that the agent has not undergone a negative moral change (e.g., experiencing guilt and regret) can counteract the slippery slope effect. Furthermore, these results speak against a rewards-based explanation for the slippery slope effect as the *punishment, no regret* condition was descriptively very similar to the *no punishment, no regret* condition. This suggests that it is not simply the case that observers intuit that “crime pays.” Rather, as hypothesized and in conjunction with the results from Study 3, the commission of an immoral act is perceived as corrupting a person’s moral character and future behavior.

General Discussion

Across four studies, we found robust support for the hypothesized slippery slope effect in moral judgment. We tested and found support for four broad hypotheses: Hypothesis 1: Observers judge immoral agents as of worse moral character and more likely to commit immoral behavior later in time after an immoral act (past vs. future judgments, Studies 1 and 2; future judgments, Studies 3 and 4). Hypothesis 2: Observers do *not* perceive a similar change in morality for targets that do not commit an immoral act (e.g., those who are the victims of a transgression and/or those that simply *consider* committing an immoral act; Studies 1-3). Hypothesis 3: The slippery slope effect exhibits sensitivity to the specific nature of the relationship between the original and the future immoral acts such that observers judge an agent as particularly likely to commit future immoral acts that are similar to the initial transgression (Study 3). Furthermore, an agent’s future immoral acts are specifically predicted to become increasingly severe over time (Studies 3 and 4). Hypothesis 4: The slippery slope effect is driven, at least in part, by perceptions that committing an immoral act indelibly “corrupts” one’s moral character. Supporting this hypothesis, committing an immoral act—even over and above intentionally attempting one—leads to greater predictions of future immoral behavior (Study 2). In addition, observers predict that agents who commit an immoral act will experience less guilt over time (Study 3). Furthermore, explicit signals that counteract perceptions of a corrupted conscience, such as an agent experiencing guilt and regret, interrupt the slippery slope effect. However, simply removing the rewards of an immoral act (e.g., by punishing the agent for the act) do not, on their own, attenuate slippery slope perceptions (Study 4).

Connections to Past Research

Our findings build on and extend the growing body of literature on judgments of moral character, which shows that we

evaluate not just the rightness or wrongness of specific acts, but also what those acts reveal about the underlying dispositions of the people performing those acts (Goodwin et al., 2014; Hartley et al., 2016; Helzer & Critcher, 2018; Pizarro & Tannenbaum, 2011; Uhlmann et al., 2015). These moral character evaluations, in turn, play a central role in person perception and inform the predictions that people make of a target’s future behavior—and therefore how to best interact and engage with that person (e.g., Everett et al., 2016; Jordan et al., 2016). Therefore, given morality’s prominence in social cognition, understanding how people think about change in this domain has broad potential theoretical implications in terms of how people predict another person’s behavior—as well as adjust their own behavior in response to those predictions.

The research we report here shows that observers do not simply make judgments of moral character in a static fashion, but they also make inferences about how moral character will change over time. Many theories in social psychology have emphasized a consistency principle: Individuals are expected to exhibit consistency between their past and future behavior (e.g., Baxter & Goldberg, 1987; Buehler et al., 1994; Helzer & Dunning, 2012; Kelley, 1967; Quoidbach et al., 2013; Vazire & Mehl, 2008). However, the present work demonstrates that, as opposed to a mere consistency effect, observers do not predict that immoral agents are equally immoral at all times. Instead, they exhibit a slippery slope pattern of thinking, expecting individual moral acts to alter the trajectory of an agent’s future character. Our findings highlight a potential gap in the existing literature—observers treat agents not simply as having stable character traits that reveal themselves with new information but instead as character traits and behavioral patterns *that can change* based on information about how that agent has behaved.

Limitations and Future Directions

Given the relative lack of empirical work on the subject, we hope that these findings will inspire future research examining when, why, and how people make predictions about moral change. It will be especially informative to test and understand the conditions under which people do *not* endorse slippery slope predictions. For example, people also frequently anticipate “redemption arcs,” in which an agent stops behaving immorally and becomes increasingly moral over time. Such narrative structures are common in media and storytelling (e.g., *A Christmas Carol*, *The Shawshank Redemption*) and appear to directly conflict with the slippery slope perceptions that we documented here. An intriguing question for future research is under what conditions an agent is expected to redeem themselves rather than fall further down the slippery slope.

A second remaining question concerns the exact progression of unethical behaviors that observers anticipate agents will be most likely to exhibit. For example, in real-life

examples of slippery slope behavior (e.g., Garrett et al., 2016; Welsh et al., 2015), there tends to be a gradual increase in the severity of the unethical acts committed over time. Do observers' predictions exhibit similar nuance? That is, are observers more likely to anticipate gradual (versus steeper) increases in an agent's unethical behavior? Studies 3 and 4 provide some tentative support for this prediction, but future work could explore this possibility more systematically by directly manipulating the rate of change in the severity and frequency of unethical behaviors.

One limitation of our specific methodology concerns the potential interpretations of the findings in our various control conditions. In Studies 1 to 3, the agent in the *no immoral act* conditions considers behaving immorally but ultimately decides not to do so. It is possible that participants interpreted this behavior as indicating not only that the agent did not behave immorally but that they actively resisted the temptation to do so. If so, this may have been seen as a positive virtue (rather than as a neutral act per se), and this agent may have been judged more positively as a result. Importantly, however, such effects could not have explained the slippery slope effect in all studies: For example, in Study 2, we found that successfully carrying out an immoral act increased slippery slope perceptions, even relative to an identical individual who attempted but (for reasons beyond his control) did not successfully execute the act. Nevertheless, future research may examine whether active attempts to resist temptation (or the lack thereof) may be a potential moderator of the slippery slope effect.

Another aspect of our results that may be deserving of further examination concerns the finding from Studies 1 and 2 that the agent described in *immoral act* conditions, compared with the *no immoral act* conditions, was also rated as being more immoral even in the past before the act was committed. In other words, participants appeared to infer that an agent who commits a minor transgression was already more immoral than a neutral agent—even prior to the commission of that act. Broadly speaking, this finding appears consistent with previous research that has demonstrated that even single immoral acts affect observers' judgments of an agent's general moral character (e.g., Cone & Ferguson, 2015). Nevertheless, it is possible that these judgments of the past represent an unexamined aspect of the slippery slope effect itself. Specifically, might slippery slope perceptions extend not only to judgments of future behavior, but also to judgments of past behavior as well? If so, do predictions of past behavior perhaps follow a similar—although inverted—slippery slope structure, with observers anticipating that agents past moral character was more positive in the distant (vs. recent) past? Such a “backward-inference effect” would have potential implications for domains such as legal decision-making and criminal sentencing (e.g., evaluating whether a defendant was likely to have committed a past crime). Particularly given these practical implications, future research should investigate whether and how slippery slope thinking extends to judgments and expectations of past behavior.

It also remains unclear exactly how slippery slope thinking may translate into more complex real-world judgment contexts. Our primary focus and purpose in conducting the present work was to document and precisely understand the slippery slope effect, which required careful control of procedures and measures. However, the real world is often messier than experimentation can accommodate. For example, our studies employed vignette designs where only a small amount of information was provided about the agents, whereas real-world judgments often incorporate a greater range of information (e.g., additional information about past behavior). Understanding how slippery slope predictions play out in specific real-world contexts may have important implications. For example, how might these processes manifest in a courtroom with a sentencing judge predicting a defendant's future behavior? Or a parent trying to decide whether their child's misbehavior reflects a simple momentary lapse of judgment or a trend toward more egregious behavior? If single transgressions are expected to lead to even worse behavior in the future, judges and parents may decide to enact harsh punishment even after only a single offense. Future research may wish to investigate slippery slope thinking in real-world situations to better understand how it influences decisions in situ.

Finally, our research focused exclusively on judgments of immoral acts, such as property damage, cheating, assault, and theft. We restricted our initial investigation of the slippery slope effect to immoral acts because this is the domain of moral judgment that has received the most attention in the literature and because a host of research suggests that immoral acts carry greater psychological weight than moral acts (Baumeister et al., 2001; Fiske, 1980; Goodwin & Darley, 2012; Kahneman & Tversky, 1984; Rozin & Royzman, 2001; Taylor, 1991; Wentura et al., 2000). However, morality encompasses both negative, harmful acts and positive, prosocial acts, and research suggests that observers do not necessarily rely on the same psychological processes to evaluate them (for recent reviews, see Anderson et al., 2018, 2020). Therefore, future research should examine whether observers predict that positive moral acts may change a person's future character and behavior. For example, do observers predict that agents who commit a positive moral act will continue on an “upward” trajectory toward increasingly positive future behavior? Or might the slippery slope effect be limited to negative moral behaviors, as suggested by the lay metaphors previously discussed, and by research suggesting a general asymmetry in people's propensity to diagnose negative versus positive change (Klein & O'Brien, 2016, 2017)?

Conclusion

The present research documents a robust slippery slope effect in moral judgment, showing that people view the commission of an immoral act as changing the expected future course of a

person's moral character and behavior. This work extends existing theories of moral judgment by showing that people both make inferences about an agent's moral character but also how character changes over time. Given the pervasiveness and importance of moral evaluation in social-cognitive processes, we hope that this work will prove informative and generative for future research in understanding how people predict moral behavior.

Acknowledgments

We would like to thank Lance S. Bush and Andres Montealegre for helpful feedback on a previous draft of this manuscript.

Author Contributions

R.A.A. and B.C.R. conceived of the idea. R.A.A., B.C.R., and D.A.P. planned and designed the experiments. R.A.A. and B.C.R. programmed the experiments and analyzed the data. R.A.A. wrote the manuscript. B.C.R. and D.A.P. provided critical feedback on the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Rajen A. Anderson  <https://orcid.org/0000-0002-6285-8358>

Open Practices

All preregistration documentation, materials, data, and analysis scripts for these studies are available on the Open Science Framework at <https://osf.io/m9qwp/>

Supplemental Material

Supplemental material is available online with this article.

Note

1. Along with this study, participants also answered some questions regarding an unrelated hypothesis (see OSF page for complete materials). The order of presentation for this study and the unrelated study was randomized across participants. This randomization did not moderate our effects and is therefore not discussed further.

References

- Ames, D. R., & Johar, G. V. (2009). I'll know what you're like when I see how you feel: How and when affective displays influence behavior-based impressions. *Psychological Science*, *20*(5), 586–593. <https://doi.org/10.1111/j.1467-9280.2009.02330.x>
- Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in Cognitive Science*, *24*(9), 30–39. <https://doi.org/10.1016/j.tics.2020.06.008>
- Anderson, R. A., Kamtekar, R., Nichols, S., & Pizarro, D. A. (2021). “False positive” emotions, responsibility, and moral character. *Cognition*, *214*, 104770. <https://doi.org/10.1016/j.cognition.2021.104770>
- Anderson, R. A., Pizarro, D. A., & Kinzler, K. D. (2018). Reacting to transcendence: The psychology of moral praise. In J. A. Frey & C. Vogler (Eds.), *Self-transcendence and virtue: Perspectives from philosophy, psychology, and theology* (pp. 274–290). Routledge.
- Baack, D., Fogliasso, C., & Harris, J. (2000). The personal impact of ethical decisions: A social penetration theory. *Journal of Business Ethics*, *24*, 39–49. <https://doi.org/10.1023/A:1006016113319>
- Barasch, A., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or selfless? On the signal value of emotion in altruistic behavior. *Journal of Personality and Social Psychology*, *107*(3), 393–413. <https://doi.org/10.1037/a0037207>
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. <https://doi.org/10.1037/1089-2680.5.4.323>
- Baxter, T. L., & Goldberg, L. (1987). Perceived behavioral consistency underlying trait attributions to oneself and another: An extension of the actor-observer effect. *Personality and Social Psychology Bulletin*, *13*, 437–447. <https://doi.org/10.1177/0146167287134001>
- Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2015). The braggart's dilemma: On the social rewards and penalties of advertising prosocial behavior. *Journal of Marketing Research*, *52*(1), 90–104. <https://doi.org/10.1509/jmr.14.0002>
- Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the “planning fallacy”: Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67*(3), 366–381. <https://doi.org/10.1037/0022-3514.67.3.366>
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. <https://doi.org/10.1177/1745691610393980>
- Chakroff, A., Russell, P. S., Piazza, J., & Young, L. (2017). From impure to harmful: Asymmetric expectations about immoral agents. *Journal of Experimental Social Psychology*, *69*, 201–209. <https://doi.org/10.1016/j.jesp.2016.08.001>
- Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, *136*, 30–37. <https://doi.org/10.1016/j.cognition.2014.11.034>
- Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. <https://doi.org/10.1037/pspa0000014>
- Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, *64*(2), 133–152. <https://doi.org/10.1016/j.jml.2010.10.002>
- Critcher, C. R., Helzer, E. G., & Tannenbaum, D. (2020). Moral character evaluation: Testing another's moral-cognitive machinery. *Journal of Experimental Social Psychology*, *87*, 103906. <https://doi.org/10.1016/j.jesp.2019.103906>
- Critcher, C., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, *4*, 308–315. <https://doi.org/10.1177/1948550612457688>

- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380. <https://doi.org/10.1016/j.cognition.2008.03.006>
- Dennett, D. C. (1989). *The intentional stance*. MIT Press.
- Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772–787. <https://doi.org/10.1037/xge0000165>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology, 38*(6), 889–906. <https://doi.org/10.1037/0022-3514.38.6.889>
- Garrett, N., Lazzaro, S. C., Ariely, D., & Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience, 19*(12), 1727–1732. <https://doi.org/10.1038/nn.4426>
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731. <https://doi.org/10.1037/0033-2909.132.5.692>
- Goodwin, G. P., & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology, 48*(1), 250–256. <https://doi.org/10.1016/j.jesp.2011.08.006>
- Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148–168. <https://doi.org/10.1037/a0034726>
- Hartley, A. G., Furr, R. M., Helzer, E. G., Jayawickreme, E., Velasquez, K. R., & Fleenor, W. (2016). Morality's centrality to liking, respecting, and understanding others. *Social Psychological and Personality Science, 7*(7), 648–657. <https://doi.org/10.1177/1948550616665539>
- Hartman, R., Blakey, W., & Gray, K. (2022). Deconstructing moral character judgments. *Current Opinion in Psychology, 43*, 205–212. <https://doi.org/10.1016/j.copsyc.2021.07.008>
- Helzer, E. G., & Critcher, C. R. (2018). What do we evaluate when we evaluate moral character? In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 99–107). Guilford Press.
- Helzer, E. G., & Dunning, D. (2012). Why and when peer prediction is superior to self-prediction: The weight given to future aspiration versus past achievement. *Journal of Personality and Social Psychology, 103*(1), 38–53. <https://doi.org/10.1037/a0028124>
- Jennings, M. M. (2011). *Business ethics: Case studies and selected readings*. South-Western Cengage Learning.
- Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*, 473–476. <https://doi.org/10.1038/nature16981>
- Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*(4), 341–350. <https://doi.org/10.1037/0003-066X.39.4.341>
- Kelley, H. H. (1967). *Attribution theory in social psychology*. Nebraska Symposium on Motivation, 15, 192–238.
- Klein, N., & O'Brien, E. (2016). The tipping point of moral change: When do good and bad acts make good and bad actors? *Social Cognition, 34*(2), 149–166. <https://doi.org/10.1521/soco.2016.34.2.149>
- Klein, N., & O'Brien, E. (2017). The power and limits of personal change: When a bad past does (and does not) inspire in the present. *Journal of Personality and Social Psychology, 113*(2), 210–229. <https://doi.org/10.1037/pspa0000088>
- Lode, E. (1999). Slippery slope arguments and legal reasoning. *California Law Review, 87*, 1469–1543. <https://doi.org/10.2307/3481050>
- Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*(2), 101–121. <https://doi.org/10.1006/jesp.1996.1314>
- Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108*(6), 823–849. <https://doi.org/10.1037/pspa0000021>
- Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition, 147*, 133–143. <https://doi.org/10.1016/j.cognition.2015.11.008>
- Masicampo, E. J., Barth, M., & Ambady, N. (2014). Group-based discrimination in judgments of moral purity-related behaviors: Experimental and archival evidence. *Journal of Experimental Psychology: General, 143*(6), 2135–2152. <https://doi.org/10.1037/a0037831>
- Nelson, S. A. (1980). Factors influencing young children's use of motives and outcomes as moral criteria. *Child Development, 51*(3), 823–829. <https://doi.org/10.2307/1129470>
- Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin, 40*(2), 203–216. <https://doi.org/10.1177/0146167213508791>
- Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. <https://doi.org/10.1016/j.jbef.2017.12.004>
- Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In P. Shaver & M. Mikulincer (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). APA Books.
- Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*(3), 267–272. <https://doi.org/10.1111/1467-9280.03433>
- Quoidbach, J., Gilbert, D. T., & Wilson, T. D. (2013). The end of history illusion. *Science, 339*(6115), 96–98. <https://doi.org/10.1126/science.1229294>
- Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86*, 61–79. <https://doi.org/10.1037/0033-295X.86.1.61>
- Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2
- Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*, 995–1008. <https://doi.org/10.1037/0022-3514.91.6.995>
- Saxe, R. (2009). The happiness of the fish: Evidence for a common theory of one's own and others' actions. In K. D. Markman, W. M. Klein, & J. A. Suhr (Eds.), *The handbook of imagination and mental simulation* (pp. 257–265). Psychology Press.

- Schauer, F. (1985). Slippery slopes. *Harvard Law Review*, *99*(2), 361–383. <https://doi.org/10.2307/1341127>
- Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. Springer Verlag.
- Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750–756. <https://doi.org/10.1038/s41562-018-0425-1>
- Strohinger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin*, *110*(1), 67–85. <https://doi.org/10.1037/0033-2909.110.1.67>
- Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research*, *17*, 223–236. <https://doi.org/10.1023/B:SORE.0000027411.35832.53>
- Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. <https://doi.org/10.1177/1745691614556679>
- Vaish, A., Carpenter, M., & Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. *Child Development*, *81*(6), 1661–1669. <https://doi.org/10.1111/j.1467-8624.2010.01500.x>
- Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology*, *95*(5), 1202–1216. <https://doi.org/10.1037/a0013314>
- Volokh, E. (2003). The mechanisms of the slippery slope. *Harvard Law Review*, *116*(4), 1026–1137. <https://doi.org/10.2139/ssrn.343640>
- Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. Guilford Press.
- Welsh, D. T., Ordóñez, L. D., Snyder, D. G., & Christian, M. S. (2015). The slippery slope: How small ethical transgressions pave the way for larger future transgressions. *Journal of Applied Psychology*, *100*(1), 114–127. <https://doi.org/10.1037/a0036950>
- Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach-and avoidance-related social information. *Journal of Personality and Social Psychology*, *78*(6), 1024–1037. <https://doi.org/10.1037/0022-3514.78.6.1024>