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

Rajen A. Anderson

Cornell University

Benjamin C. Ruisch

Ohio State University

David A. Pizarro

Cornell University

Author note: Correspondence concerning this manuscript can be addressed to Rajen A. Anderson, raa255@cornell.edu, Uris Hall, Cornell University, Ithaca, NY 14850

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

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/?view_only=513d17cf0e93445b8e4066f0535103cb.

Abstract

Across eight studies (total N = 2,989), we find support for the hypothesis that people exhibit "slippery slope" thinking in their judgments of moral character, such that committing a single immoral act is seen to (a) lead to a lasting negative shift in a person's character and (b) increase a person's likelihood of committing additional immoral acts in the future. In Studies 1-3b we document the slippery slope effect, finding that a person who commits an immoral act is judged as subsequently more likely to commit additional immoral acts, and as generally having worse moral character in the future than in the past. In Studies 4a-4b, we also find that it is the *commission* of an immoral act specifically—rather than merely intending or attempting to commit an immoral act—that gives rise to this slippery slope effect on judgments. In Study 5, we find that the slippery slope effect is most likely to occur for subsequent immoral acts that are perceived as similar to the initial act. Finally, in Study 6 we test two potential psychological mechanisms underlying this effect: (1) that an immoral act is seen to "corrupt" a peron's moral character and (2) that the rewards and punishments that result from committing an immoral act shape future moral behavior. We find that only the former mediates the slippery slope effect, indicating that this effect does not stem from the possible positive consequences of the immoral act for the agent (e.g., beliefs that "crime pays"), but instead from a perceived corrupting of moral character. In contrast to "consistency-in-character" models of moral judgment, we find that people judge immoral behavior as indelibly corrupting the character of a moral agent.

Keywords: moral judgment, slippery slope, intentionality, character, punishment

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

"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 character follows a trajectory or 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 "sliding down a slippery slope," "falling from grace," or being on a "highway to hell." Similarly, a good person who chooses *not* to commit a misdeed is often described as "sticking to the straight and narrow" or "choosing the high road." Central to these lay conceptualizations of character is the importance of individual events and (im)moral behaviors in shaping the future trajectory of a person's character: positive moral behaviors propel a person toward future moral behavior, while immoral behaviors push a person toward committing subsequent immoral acts. Although these ideas are common in metaphors and lay discussion surrounding moral behavior, 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 program, we examine the question of how an agent's immoral behavior shapes people's perceptions of the trajectory of that agent's future behavior and moral character. Specifically, we test for slippery slope thinking in people's judgments and 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 towards

committing other immoral acts in the future. Existing theories on character attribution are relatively agnostic regarding 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 moral character and behavior can change as a function of the actions taken by an individual.

The Slippery Slope in Past Theory and Research

The slippery slope has primarily been discussed in the domains of philosophy and law in the context of argumentation. Slippery slope arguments are typically used to 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, slippery slope arguments 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, these arguments seek to make the initial small change seem potentially dangerous or immoral, therefore discouraging others from enacting it.[1] The persuasive strength of a slippery slope argument depends on the perceived similarity between the current action or behavior (*Innocuous Action A*) and the posited end state (*More Negative Effect B*; Corner, Hahn, & Oaksford, 2011; Volokh, 2003).

In the present research, we use the metaphor of the "slippery slope" as a tool 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 literatures, 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*.

Although slippery slope thinking has not been examined in the context of moral judgments, there is some preliminary evidence that real-world immoral behavior may in fact follow a pattern of increasing likelihood over time. That is, behaving immorally *does* seem to relate to an increased likelihood of behaving immorally in the future, at least under some circumstances (e.g., Baack, Fogliasso, & Harris, 2000; *BBC Ethics Guide*, 2012; Jennings, 2011; Tenbrunsel & Messick, 2004). For example, some empirical research has found that smaller, initial dishonest acts like misreporting performance on a test to earn more money can predict larger dishonest acts later (Welsh, Ordóñez, Snyder, & Christian, 2014). Individuals also readily justify and rationalize small indiscretions, potentially paving the way for increasingly immoral future acts (Mazar, Amir, & Ariely, 2008). Furthermore, neurological activity in the amygdala has been shown to decrease with repeated dishonest acts, suggesting adaptation to and normalization of the behavior. This normalization, in turn, can then lead people to subsequently perform more egregious acts of dishonesty (Garrett, Lazzaro, Ariely, & Sharot, 2016).

To the degree that people are aware of these real-world patterns of behavior, this provides additional reason to predict that people's lay judgments of character may exhibit a slippery slope effect, such that committing *immoral behavior A* will increase the perceived likelihood of committing *immoral behavior B*. However, 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*. That is, we make no claims as to the (in)accuracy of slippery slope

judgments, but are instead interested in deepening our understanding of the nature and process of moral evaluation.

Moral Character and Predicting Future 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 Helzer & Critcher, 2018; Pizarro & Tannenbaum, 2011; Uhlmann, Pizarro, & Diermeier, 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, Piazza, & Rozin, 2014) – and these judgments of moral character play a central role in person perception. For example, judgments of a target's morality more strongly predict liking and respect for that target than do judgments of that target's competence and sociability (Hartley et al., 2016). In addition, observers rate morality and moral traits as more central to a person's identity than non-moral traits (Newman, Freitas, & Knobe, 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, 2012). 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 (such that observers update

their impressions of a person based on that person's moral actions; e.g., Cone & Ferguson, 2015; Gawronski & Bodenhausen, 2006; Mann & Ferguson, 2015; Rydell & McConnell, 2006; Siegel, Mathys, Rutledge, & Crockett, 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. 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 how people will predict change in others' moral character over time.

We formulated two competing hypotheses regarding how observers may predict future immoral behavior and character change. 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. That is, observers may simply believe that bad people perform bad acts, but that the degree of a person's (im)morality does not 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, Griffin, & Ross, 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, Gilbert, & Wilson, 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 of worse moral character, and as being more likely to perform immoral acts in the future, relative to the past.

This question—how people perceive and interpret *change* in an individual's moral character and behavior over time—has received little attention in the literature. Even so, research shows that people employ slippery slope rhetoric for a variety of arguments (Corner et al., 2011), suggesting that such logic is common in everyday life. In addition, a growing body of literature on diagnosing change provides some reason to predict that observers may interpret an immoral act as signaling a negative change in moral character and behavior (for a review, see O'Brien, 2020). Specifically, recent work on perceptions of social change (e.g., whether the economy is declining versus improving) suggests that people need relatively little information to diagnose change, particularly when asked to diagnose negative change (i.e., change for the worse, versus change for the better; Klein & O'Brien, 2016, 2017). However, this past research does not address predictions of future change – i.e., not just diagnosing whether something or someone has already changed, but how they will change in the future. Thus, the questions of whether and to what degree this same propensity toward diagnosing change (especially negative change) may generalize to perceptions of moral character remain unanswered.

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.

Contrasting Slippery Slope Predictions with Past Theory and Research on Moral Character

In sum, past research on 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, Russell, Piazza, & Young, 2017; Critcher Inbar, & Pizarro, 2013; Klein & O'Brien, 2016; Reeder & Brewer, 1979; Sripada, 2012; Uhlmann et al., 2015). This research has also revealed that more nuanced considerations can also shape people's judgments of others' moral character, such as the specific domain of an (im)moral act (e.g., harm versus purity; Chakroff et al., 2017; Masicampo, Barth, & Ambady, 2014).

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. For example, a person who commits an immoral act on August 2nd should not, according to these theories, be seen as any more likely to commit another immoral act on August 3rd than on August 1st. Rather, these accounts hold that perceptions of this person's moral

character should be globally and uniformly affected by the information about her immoral behavior. 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 her 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. Further, 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; Volokh, 2003), we hypothesized that slippery slope judgments would be sensitive to the perceived relation between the initial immoral act and the predicted future acts. Specifically, 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 within the same general category of behavior) being judged as 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. That is, 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 over time (Garret et al., 2016; Welsh et al., 2014) – 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 a series of eight studies (total N = 2,989) to test the above four 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, such that they anticipate that an agent who commits an immoral act will be more likely to commit subsequent immoral acts in the future. In Studies 2 and 3a-3b we replicate and build on Study 1 by contrasting judgments of an agent who commits an immoral act with judgments of an agent who does *not* commit an immoral act (e.g., one who simply considers acting immorally). In doing so, we isolate and test the effect of acting immorally on predictions of future moral character and behavior. In Studies 4a and 4b 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 5, 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 similarity between a given immoral action and the initial immoral act. Finally, in Study 6, we test whether future transgressions will be expected to be of increasing severity, and we test two potential psychological mechanisms – moral corruption versus utilitarian positive consequences – for the slippery slope effect in moral judgments.

Statistical Power and Open Science Practices

Throughout these studies, we test and verify the generalizability of the slippery slope effect by using a wide variety of experimental stimuli depicting different moral agents and various types of immoral actions. In all studies, we report all manipulations and dependent measures. 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. To determine the sample size for Study 1, we conducted a power analysis (using G*Power version 3.1; Faul,

Erdfelder, Buchner, & Lang, 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 Study 1. For subsequent studies, we based our power analyses on 80% power to detect an effect of the size observed in Study 1. We collected this minimum target sample size and exceeded it when possible to maximize power. Following recent best practices recommendations (e.g., McShane & Böckenholt, 2017), we also conducted an internal meta-analysis of our studies to further increase statistical power and better estimate the true size of any observed effects (Braver, Thoemmes, & Rosenthal, 2014). All materials, data, and preregistration documentation are available at https://osf.io/m9qwp/?view_only=513d17cf0e93445b8e4066f0535103cb.

Study 1

Our primary aim in Study 1 was to provide an initial test of whether people exhibit slippery slope thinking in their judgments of moral character. Specifically, we tested Hypothesis 1, that observers would judge an agent as having worse moral character and being more likely to behave immorally after (vs. before) committing an immoral act. By assessing perceptions of both the agent's past and future (im)morality, we were able to specifically assess whether observers perceived that the immoral act signaled a change in the trajectory of the agent's moral character and future behavior, versus simply affecting global character judgments, as predicted by existing theories of moral judgment.

Methods

Participants

Participants were recruited from Amazon's Mechanical Turk (MTurk; for a discussion of this platform as a research tool, see Buhrmester, Kwang, & Gosling, 2011). We chose this

participant pool because past research has shown that MTurk samples tend to be more diverse than the typical college student samples – in particular, to be older, less educated, and generally less "WEIRD" (i.e., Western, Educated, Industrialized, Rich, and Democratic; Henrich, Heine, & Norenzayan, 2010) than college students (Paolacci & Chandler, 2014) – considerations that are particularly important when examining moral judgment (e.g., Haidt, Koller, & Dias, 1993; Shweder, Mahapatra, & Miller, 1987).

Based on the power analysis detailed above, we recruited 199 U.S. participants. As specified in our preregistration, we excluded 11 participants for failing the attention check, leaving a final sample size of 188 (54% female, 46% male, M_{age} = 36.42). However, none of our conclusions are substantively altered if all participants are included in analyses.

Design

Participants read the following vignette[2]:

> *"Henry was walking back to his car after leaving the shopping mall. When he got back to his car, he found that the car next to his had parked really close and didn't leave much room for him to get inside. Henry was really angry that someone could be so inconsiderate and wanted to get back at the person. After thinking about it for a few minutes, Henry decided to use the screwdriver from the toolbox in his trunk to punch a hole in one of the tires. Henry went around to his trunk and pulled out the screwdriver. Looking around to make sure no one was looking, Henry walked around to the other car, crouched down, and used the screwdriver to punch a hole in one of the tires, causing it to deflate. Henry squeezed into his car and drove off. No one saw what Henry did."*

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 three specific moral infractions (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 preregistered inclusion criteria for the attention check was selecting the correct response to this question (i.e., that Henry used a screwdriver to puncture the other car's tire).

Results and Discussion

We first created composite measures of the target's perceived morality in the past and future. To do so, we averaged all five items for both the past (α_{past} = .84) and future (α_{future} = .86), such that higher scores indicated more negative moral character and a greater likelihood of committing immoral acts. To test for a slippery slope effect in participants' judgments, we conducted a series of paired-samples *t*-tests, comparing past and future ratings for both the composite morality measure and the five items individually (see Table 1).

Overall, the measures showed the predicted pattern of results, such that observers rated Henry as more immoral in the future than in the past, $t(187) = 4.74$, $p < .001$, $d = .35$. Looking at the individual components of the composite score, we found that participants judged Henry as having worse moral character ($t(187) = 2.64$, $p = .009$, $d = .19$), as generally being more likely to commit other immoral acts in the future ($t(187) = 2.56$, $p = .01$, $d = .19$), and as being more likely to commit the three specific immoral acts ($t(187) = 4.20$, $p < .001$, $d = .31$). Looking at the pattern of results on the specific immoral acts, we found significant differences on 2 of the 3

items (only the stealing item was not statistically significant, although it was directionally consistent with the other items).

	Past *M* (SD)	Future *M* (SD)	*t*(187)	*p*	Cohen's D
Good/Bad Person (1 = Extremely Bad, 9 = Extremely Good)	3.74 (1.29)	3.52 (1.29)	2.64	.009	.19
Likelihood of Unethical Behavior (1 = Very Unlikely, 9 = Very Likely)	7.10 (1.80)	7.35 (1.57)	2.56	.01	.19
Likelihood of Stealing (1 = Very Unlikely, 9 = Very Likely)	5.46 (1.92)	5.57 (1.99)	1.13	.26	.08
Likelihood of Assault (1 = Very Unlikely, 9 = Very Likely)	5.39 (1.83)	5.86 (1.87)	5.84	< .001	.43
Likelihood of Driving While Intoxicated (1 = Very Unlikely, 9 = Very Likely)	5.31 (1.92)	5.56 (2.00)	2.96	.004	.22
General Immoral Evaluation (average of all items, with Good/Bad Person reverse-scored)	5.91 (1.39)	6.16 (1.41)	4.74	< .001	.35

Table 1. Summary of results for Study 1. Participants rated the agent as generally worse in moral character and more likely to commit a variety of immoral behaviors (with the exception of stealing) in the future after the initially described immoral act than in the past before the immoral act.

Overall, the pattern of results we observed in Study 1 supported our predictions and provide evidence for Hypothesis 1. In contrast to consistency models of moral judgment (e.g.,

Baxter & Goldberg, 1987; Buehler et al., 1994; Kelley, 1967), we found that it is not simply the case that an individual who commits an immoral act is judged as having generally worse moral character; rather, observers judge that the agent will *become* more immoral in the future following the commission of that immoral act. In other words, just as in lay metaphors of moral character, the commission of an immoral act is seen as signaling a change in one's "moral path," both for an individual's overall moral character and expected likelihood of committing other immoral acts in the future.

Study 2

We had two primary aims in Study 2. First, we wanted to replicate the slippery slope effect from Study 1 – that an agent will be expected to become more immoral after the commission of an immoral act (Hypothesis 1). Second, we wanted 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 alternative explanations for the results of Study 1 – for example, that it was the negative experience that *provoked* the immoral act (in this case, having an inconsiderate person park too close to one's own car) that led 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 above, we hypothesized that observers would specifically perceive these changes in moral character and behavior for an agent who actually commits an immoral act (compared to someone in a similar situation who does not commit that immoral act).

We made two further changes in this study. First, we added two additional questions to our dependent measures to include a wider variety of specific immoral acts. Second, to ensure

that the results of our first study could not be attributable to any specific inferences that participants may have made about the individual based on his name (e.g., regarding age, race, or socioeconomic status), in this study we randomized the name of the target individual by drawing from a bank of common U.S. male names.

Methods

Participants

Based on the results of Study 1, 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 Study 1. This analysis recommended a sample size of 52 participants per condition. In this and all subsequent studies we ensure this minimum sample size. When possible, we increased this sample size to increase statistical power based on available resources. To further increase statistical power in this study, we recruited 265 participants from Amazon's Mechanical Turk (more than N = 131 per condition). As specified in our preregistration, and consistent with Study 1, 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. As in Study 1, 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 the same vignette from Study 1. 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, participants answered the past and future morality questions from Study 1 (moral character judgment, general likelihood of behaving immorally, and likelihood of committing three specific immoral acts), as well as rated the man's likelihood of committing two additional immoral behaviors. Specifically, in addition to the immoral acts from Study 1, participants were asked to judge the likelihood that the man would damage someone else's car and the likelihood that he would break someone else's property (from 1 = *not likely at all* to 9 = *very likely*). After making these judgments, participants completed an attention check in which they were asked to recall what happened in the story. 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 (α = .96) and the future (α = .96). Replicating Study 1, 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 general likelihood of immoral behavior, and the composite likelihood judgments of the five specific immoral behaviors (α_{past} = .94, α_{future} = .95). Consistent with our predictions and replicating Study 1, there was a significant difference

for 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 for 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 (a meta-analysis, detailed below, revealed that the size of this effect, although not statistically significant, did not significantly differ from that of our other studies. We therefore judged that these results are likely simply due to random variation between samples, rather than to any meaningful difference between the measures or studies; Lakens & Etz, 2017).

To determine whether this slippery slope effect was truly the result of the agent having committed an immoral act (versus, e.g., being the victim of a transgression and/or being in an anger-provoking situation), 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 ANOVA on the composite moral evaluation score (as specified in our preregistration), with condition (*immoral act* versus *no immoral act*) specified as a between-subjects factor and time (past versus future) specified as a within-subjects factor. Consistent with past work, 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 X time interaction, F (1, 252) = 11.38, $p = .001$, $\eta_p^2 = .04$. As described above, 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 that we documented previously 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 specifically the commission of an immoral act that led to these slippery slope perceptions. In other words, as hypothesized, 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 3a

We had two primary aims in Studies 3a and 3b. First, we wanted to replicate our previous results using a different vignette and a different immoral behavior. Second, we aimed to isolate and examine the relative influence of immoral *action* versus immoral *intention*, and to understand the role of each factor in giving rise to the slippery slope effect. This allowed us to test and rule out possible confounds from our previous studies. In particular, the *immoral act* and *no immoral act* conditions in Study 2 differed not only in the behavior that was performed, but also in the described thoughts and intentions of the agent leading up to that behavior. Specifically, the agent in the *immoral act* condition is explicitly described as considering and

deciding to engage in the immoral act, whereas in the *no immoral act* condition, the agent is not described as having considered the immoral act. This raises the possibility that the effects that we observed may be due to the *consideration* of the immoral act (or having "immoral thoughts" more generally), rather than performing the immoral act itself. For example, it may be the case that observers inferred that simply having considered this immoral act made it more likely that the agent would behave more immorally in the future.

To isolate the role of actually committing the act in giving rise to the slippery slope effect, in Studies 3a and 3b we modified the vignette of the *no immoral act* condition so that the agent is explicitly described as considering the immoral act, but ultimately decides against performing it. As noted above (Hypothesis 2), we hypothesize that the slippery slope effect is specifically due to the commission of the immoral act itself, rather than factors such as considering committing such an act. Accordingly, we predicted a similar pattern of results to that of Study 2 – that observers would anticipate a greater change in future moral character and behavior for an agent that actually commits an immoral act, versus one who considers the immoral act but does not perform it. This pattern of results would be consistent with our hypothesized mechanism for the slippery slope effect: that observers see the act itself as corrupting the agent's moral character (Hypothesis 4).

Methods

Participants

We recruited 230 participants (49% female, M_{age} = 36.9) from MTurk. As in our previous studies, we excluded participants who failed the attention check (n = 8), leaving a final sample size of 222 (49% female, M_{age} = 37.05). However, none of our conclusions are substantively altered if these participants are included in the analyses.

Design

We randomly assigned participants to one of two conditions. In both conditions, participants read a story describing a high school student worried about an upcoming exam. In both versions of the vignette, the student is explicitly described as considering cheating on the exam. In the *immoral act* condition, the student proceeds to cheat on the test, while in the *no immoral act* condition the student decides against cheating and simply takes the test.

As in our previous studies, after reading the vignette, participants answered questions about the agent's past and future moral character and behavior. Participants were asked to evaluate how good or bad of a person the agent was (from 1 = e*xtremely bad* to 9 = *extremely good*) and how likely the agent was to do something unethical or illegal (from 1 = *not likely at all* to 9 = *very likely*), both in the past and in the future. (Participants in this study did not judge the agent's likelihood of committing specific immoral acts – a question that we address in our next study, Study 3b.) We randomized the presentation order of the past and future questions between participants. After making judgments of the agent's moral character and behavior, participants completed an attention check in which they were asked to recall whether the student cheated on the test, didn't cheat on the test, or whether this information was not specified. Consistent with our previous studies, our inclusion criteria for the attention check was selecting the correct answer for one's assigned condition (i.e., reporting that the agent cheated on the test in the *immoral act* condition, or did not cheat on the test in the *no immoral act* condition).

Results and Discussion

Replication of the Slippery Slope Effect

We first tested whether we replicated our previous effects by examining participants' judgments of the *immoral act* agent. As predicted, and in keeping with our previous studies, we

found a significant slippery slope effect in participants' moral judgments, such that observers believed that the agent in the future would be of generally worse moral character, $t(111) = 6.12$, $p < .001$, $d = .58$, and would be more likely to commit other immoral acts, $t(111) = 9.17$, $p < .001$, $d = .87$. (In this and all subsequent studies we replicate this basic slippery slope effect across all of our dependent measures. Effect sizes and confidence intervals can be found in Figure 2).

The Effect of Committing an Immoral Act

We then tested our primary hypothesis for Study 3a, that an agent who commits an immoral act, relative to an agent who considers but does not commit that act, will be judged as having worse moral character and being more likely to commit immoral acts in the future than in the past (Hypothesis 2). To test this hypothesis, we conducted a 2 X 2 mixed-design ANOVA, with condition (*immoral act* versus *no immoral act*) as the between-subjects factor and time (past versus future) as the within-subjects factor (see Table 2 for descriptive statistics of Study 3a).

	Condition	Past *M* (SD)	Future *M* (SD)
Good/Bad Person (1 = Extremely Bad, 9 = Extremely Good)	Neutral	7.08 (1.15)	7.13 (1.18)
	Immoral	6.10 (1.38)	5.33 (1.42)
Likelihood of unethical behavior (1 = Very Unlikely, 9 = Very Likely)	Neutral	3.01 (1.67)	3.03 (1.58)
	Immoral	4.16 (2.02)	5.75 (1.84)

Table 2. Descriptive statistics for Study 3a. Participants rated the agent who committed the immoral act as worse in moral character and more likely to commit other unethical behaviors in the future than in the past. Conversely, there was no significant change in ratings of the agent who did not commit the immoral act.

Consistent with past work, there was a significant main effect of condition on judgments of character, such that participants rated the agent as more immoral in the *immoral act* condition than in the *no immoral act* condition, $F(1, 220) = 79.75$, $p < .001$, $\eta_p^2 = .27$. Critically, however, this effect of condition was again qualified by a significant condition X time interaction, such that the difference in perceived character between the *immoral act* agent and the *no immoral act* agent was larger for judgments of the future relative to the past, $F(1, 220) = 29.40$, $p < .001$, $\eta_p^2 = .12$. When the agent did not commit the immoral act, judgments of character changed little from the past to the future. Conversely, consistent with our slippery slope account, when the agent did commit the immoral act, participants judged that agent as having worse moral character in the future than in the past.

We next examined participants' judgments of the agent's likelihood of engaging in other immoral behavior, using the same 2 X 2 mixed-model analysis. As with participants' judgments of the agent's character, there was a significant main effect of condition on perceived likelihood of unethical behavior: participants judged the *immoral act* agent as generally more likely to commit unethical acts than the *no immoral act* agent, $F(1, 220) = 85.84$, $p < .001$, $\eta_p^2 = .28$. As above, however, this main effect of condition was again qualified by a significant condition X time interaction, $F(1, 220) = 44.25$, $p < .001$, $\eta_p^2 = .17$. Specifically, participants predicted little change in the *no immoral act* agent's behavior, but, as hypothesized, they predicted a significant increase in the *immoral act* agent's likelihood of committing other unethical acts in the future.

In sum, consistent with past work, in this study we found a main effect of committing an immoral act, such that the agent who committed an immoral act (versus one who only considered that act) was judged to be of generally worse moral character. (We consistently replicate this effect across all of our studies.) This effect is in line with "consistency models" of moral

character and suggests that people judge an immoral agent as generally being of worse moral character (even before the event described in the vignette). Importantly, however, we also found that judgments of the agent who committed an immoral act – but not one who merely contemplated that act – changed from past to future, suggesting that people perceive the act of transgressing as precipitating a change within the agent. These results conceptually replicate those of our previous studies, while specifically identifying the commission of an immoral act as necessary for the slippery slope effect to emerge.

Study 3b

In Study 3b, we aimed to replicate and extend the results of Study 3a by having participants make judgments of the agent's likelihood of committing specific immoral behaviors (similar to Studies 1 and 2). Specifically, we examined whether participants judged an agent who commits an immoral act (versus one who simply contemplates that immoral act) as more likely to commit a range of specific immoral behaviors in the future.

Methods

Participants

We recruited 232 participants from MTurk. Consistent with our previous studies, we excluded 21 participants for failing the attention check, leaving a final sample size of 211 (55% female, M_{age} = 39.20). However, none of our conclusions are substantively altered if all participants are included in analyses.

Design

The design of Study 3b was identical to that of Study 3a, with the addition of questions assessing the agent's perceived likelihood of committing five specific immoral behaviors in both the past and the future. Consistent with our previous studies, we chose behaviors with some

degree of conceptual similarity to the initial immoral act – in this case, behaviors related to fairness and honesty. (We formally test the role of conceptual similarity in Study 5) Specifically, in addition to the general character and behavior judgments from Study 3a, we also asked participants how likely it was that the agent, if given the opportunity, would lie about something, cheat on another test, plagiarize someone else's work, take a shortcut in a race, and cheat on a homework assignment (from 1 = *Not likely at all* to 9 = *Very likely*).

Results and Discussion

Replication of Study 3a

We first tested whether we replicated the results of Study 3a. As above, we began by examining the item assessing general moral character by conducting a 2 (condition: *immoral act* versus *no immoral act*) X 2 (time: past versus future) mixed-model ANOVA. Consistent with past work, we again found a significant main effect of condition such that the agent who committed the immoral act was rated as more immoral than one who simply considered that act[3], F (1, 207) = 64.47, $p < .001$, $\eta_p^2 = .24$. Importantly, however, replicating our previous results, we also found that this effect of condition was qualified by a significant condition X time interaction, such that the difference in perceived character between the *immoral act* agent and the *no immoral act* agent was larger for future judgments than past judgments, F (1, 207) = 46.04, $p < .001$, $\eta_p^2 = .18$.

We next examined participants' judgments of the agent's general likelihood of doing something unethical or illegal, using the same mixed-model ANOVA. We again found a significant main effect of condition, with observers judging the agent who committed the immoral act as more likely to commit other unethical behavior, F (1, 209) = 98.59, $p < .001$, $\eta_p^2 = .32$. However, we once again observed a significant condition X time interaction, such that the

difference in predicted likelihood between the *immoral act* agent and the *no immoral act* agent was larger for future judgments than past judgments, F (1, 209) = 39.20, $p < .001$, $\eta_p^2 = .16$. Again, while observers expected little difference in the past and future behavior of an individual who simply considered an immoral act, the agent who actually committed this immoral act was seen as subsequently more likely to engage in other immoral behaviors. These results replicate our previous studies, showing that people judge that a single immoral act can "corrupt" an agent's character and make them more prone to committing future immoral acts.

Predicting Specific Behaviors

We next examined participants' judgments of the agent's likelihood of committing other specific immoral acts. We first averaged across the individual acts to create a composite score of perceived likelihood that the agent would commit unethical acts in the future ($\alpha = .97$) and in the past ($\alpha = .95$). As with the general judgments of character and behavior, there was again a significant main effect of condition, such that the agent who committed the immoral act (versus one who only considered the act) was perceived as more likely to commit other unethical behaviors, F (1, 209) = 158.36, $p < .001$, $\eta_p^2 = .43$. Consistent with our hypotheses, however, this main effect of condition was again qualified by a significant condition X time interaction, F (1, 209) = 89.90, $p < .001$, $\eta_p^2 = .30$, such that the difference between the *immoral act* agent and the *no immoral act* agent was larger for future judgments ($M_{immoral\ act} = 6.27$ vs $M_{no\ immoral\ act} = 2.75$) than in past judgments ($M_{immoral\ act} = 4.56$ vs $M_{no\ immoral\ act} = 2.95$).

In summary, Study 3b replicated and extended the findings of Studies 1-3a by demonstrating that the commission of a single immoral act – over and above simply contemplating that act – leads observers to anticipate future changes in that agent's character and behavior. Further, these effects emerge not only on general summary measures of moral

character and behavior, but also in participants' judgments of the likelihood that the individual would engage in a range of specific immoral acts.

Study 4a

In Studies 4a and 4b, we replicated and extended our previous results by separately examining the roles of *attempting* and *committing* an immoral act. That is, in our previous studies 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 (i.e., the intended outcome actually occurs and is caused by the agent). 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, Carpenter, Tomasello, 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).

Methods

Participants

We recruited 342 participants from MTurk. As in our previous studies, we excluded 40 participants for failing the attention check, leaving 302 participants (52% female, M_{age} = 38.50) in our final sample. However, none of our conclusions are substantively altered if all participants are included in 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. In the *no attempt* condition, the student considers cheating on the exam 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 on 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 *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 in order 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 the story did not say). As in our previous studies, 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*) X 2 (time: past versus future) mixed-model ANOVA. Consistent with our previous studies, we conducted this model on each of our three primary dependent measures (1) global judgments of moral character (i.e., how good or bad a person the target is), (2) the target's general likelihood of doing something unethical or illegal, and (3) the averaged perceived likelihood that the agent would commit the six specific immoral behaviors (α_{past} = .94, α_{future} = .95).

In keeping with our predictions, there was a significant condition X 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, η_p^2 = .05, likelihood of doing something unethical or illegal, F (2, 299) = 16.51,

$p < .001$, $\eta_p^2 = .10$, and 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 to judgments of the agents who committed and/or attempted to commit an immoral act. As seen in Table 3, 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. 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.

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 x 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. Further, 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.

	Condition	Past *M* (SD)	Future *M* (SD)
Good/Bad Person (1 = Extremely Bad, 9 = Extremely Good)	No Immoral Attempt	7.28 (0.86)	7.28 (0.77)
	Immoral Attempt	6.44 (1.47)	5.86 (1.68)
	Immoral Act	6.32 (1.47)	5.44 (1.68)
Likelihood of unethical behavior (1 = Very Unlikely, 9 = Very Likely)	No Immoral Attempt	2.67 (1.37)	2.51 (1.22)
	Immoral Attempt	3.81 (2.03)	4.77 (2.08)
	Immoral Act	4.64 (2.10)	5.82 (1.71)
Average likelihood of performing specific unethical behaviors (1 = Very Unlikely, 9 = Very Likely)	No Immoral Attempt	2.78 (1.36)	2.64 (1.22)
	Immoral Attempt	4.16 (1.92)	4.92 (1.92)
	Immoral Act	4.69 (1.83)	5.94 (1.57)

Table 3. Descriptive statistics for Study 4a. 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.

Study 4b

Study 4a provided mixed evidence as to whether committing (versus merely attempting) an immoral act independently plays a role in eliciting slippery slope judgments. In Study 4b, we therefore conducted a replication of the *immoral attempt* and *immoral act* conditions in order to

assess the robustness of this effect. As noted above, although this prediction is not central to our theoretical framework, we viewed it as providing additional support for our hypothesized psychological mechanism of moral corruption (Hypothesis 4).

Methods

Participants

We recruited 251 participants (51% female, M_{age} = 35.75) from MTurk. As we did not include an attention check in this study, we excluded no participants from analyses.

Design

The design of this experiment was very similar to that of that *immoral attempt* and *immoral act* conditions from Study 4a. Participants read a story about a high school student who attempts to cheat on a test by writing notes on his hand – an immoral act that is either unsuccessful (*immoral attempt*) or successful (*immoral act*). As in our previous studies, after reading the vignette, participants rated the likelihood that the target would perform various unethical or illegal behaviors in the future (start using drugs, shoplift, assault someone, cheat on another test, and drive while intoxicated; from 1 = *not likely at all* to 9 = *very likely*). Based on the results of Study 4a, in this study we assessed only the critical comparison point of participants' future predictions (we did not assess judgments of the agent in the past).

Results and Discussion

We averaged across the five likelihood judgments (α = .84) to form a single index of predicted likelihood of committing future immoral acts. As predicted, we found that participants in the *immoral act* condition (M = 4.82) rated the agent as significantly more likely to commit other unethical acts in the future than did participants in the *immoral attempt* condition (M = 4.38), $t(249) = 2.29$, $p = .02$, Cohen's $d = .29$. These results replicate those of Study 4a and show

that committing an immoral act – over and above having immoral intentions and attempting that act—contributes to the slippery slope effect. This pattern of results is consistent with our hypothesized psychological mechanism for the slippery slope effect – that the commission of an immoral act is viewed as indelibly corrupting an agent's moral character and future behavior.

Study 5

In Study 5, we examined whether and to what degree the slippery slope effect is contingent on the similarity between the initial immoral behavior and the predicted future behaviors. That is, when judging the future behavior of an agent who commits an immoral act, do observers generalize only to other immoral acts within the same "domain" as the initial immoral act (e.g., generalizing from one fairness violation to another, but not from fairness to harm; cf. Haidt & Joseph, 2004)? Or will they also perceive that the agent would be more likely to commit other dissimilar immoral acts as well?

The literature on moral evaluation appears to provide support for both predictions. Research on moral typecasting has demonstrated that more agentic individuals are judged as more likely to perform both moral and immoral acts than less agentic individuals (e.g., Gray & Wegner, 2009, 2011). This work suggests that people may be domain-insensitive in judging an agent's future behavior, simply concluding that this person is an immoral agent and therefore will commit other immoral acts, regardless of the specific nature of those acts.

However, there is also evidence that observers can employ more nuanced considerations in their moral judgments. As discussed above, past work has shown that moral judgments often exhibit domain-sensitivity. For example, observers predict that impure agents will commit both impure and harmful acts, while harmful agents are expected to only commit other harmful acts (Chakroff et al., 2017). Based on this past work, we predicted that the slippery slope effect in

moral judgment would exhibit domain sensitivity. Specifically, we hypothesized that, in predicting an immoral agent's future behavior, observers would expect that agent to be particularly likely to commit future (versus past) violations of the same general class or domain (e.g., within the domains of harm, fairness, or purity), whereas "cross-domain" immoral acts would be judged as relatively less likely (Hypothesis 3). For example, we expected that observers would judge an agent who stole as more likely to commit other fairness violations, such as other acts of theft, fraud, or cheating, than they would be to commit a harm violation, such as physically attacking someone. Conversely, an agent who assaulted someone should be expected to be more likely to commit other acts of harm than they would fairness violations. If so, this would constitute an important moderating factor in determining the "scope" of the slippery slope effect, suggesting that it is particularly likely to occur within (relative to across) moral domains. Similarly, this would also have implications for the exact nature of the moral corruption that we hypothesize underlies the slippery slope effect, suggesting that this numbing of conscience may be specific to the nature of the immoral act.

Method

Participants

We recruited 802 participants (46% female, M_{age} = 36.3) from MTurk. We excluded 48 participants for failing the attention check, leaving 754 participants (46% female, M_{age} = 36.71) in our final sample. However, none of our conclusions are substantively altered if all participants are included in analyses.

Design

We randomly assigned participants to one of four conditions, based on a 2 (domain: *theft*, *harm*) X 2 (action: *immoral act*, *no immoral act*) between-subjects design. Participants in all

conditions read a story in which an agent considers committing an immoral act. Participants read that the person in the story either considered stealing something (*theft* condition) or physically harming someone else (*harm* condition). To further ensure the generalizability of any observed effects, we used two different harm vignettes and two different theft vignettes, depicting different agents in different situations (see OSF page for all vignettes). Participants read one of two possible outcomes for the story: either the person committed the action they considered (*immoral act* condition) or decided against committing the action they considered (*no immoral act* condition).

After reading the story, participants responded to two sets of questions, assessing perceptions of the agent in the past (one year prior) and the future (one year after). Participants rated the agent's general moral character (i.e., how good or bad of a person she was, from 1 = *extremely bad* to 9 = e*xtremely good*) and general likelihood of committing another immoral act (from 1 = *not likely at all* to 9 = *very likely*). Participants also made predictions regarding the agent's likelihood of committing various other specific immoral acts (from 1 = *not likely at all* to 9 = *very likely*): four behaviors related to theft (cheating on taxes, stealing something, committing workplace fraud, and committing identity theft) and three behaviors related to physical harm (physically assault someone, sexually assault someone, abuse animals). These behaviors were selected from a pilot study we conducted and were judged as being moderately similar to one another[4]. After answering both sets of questions, participants completed an attention check that asked them to recall what the agent did in the story. As in our previous studies, our inclusion criteria for the attention check was selecting the correct answer for one's assigned condition.

Results and Discussion

We first assessed whether we successfully replicated the results of our previous studies, such that observers judged an agent as undergoing a greater change in (1) moral character and (2) general likelihood of behaving immorally after committing (versus merely contemplating) an immoral act. To do so, we first conducted a 2 (domain: *theft* versus *harm*) X 2 (action: *immoral act* versus *no immoral act*) X 2 (time: past versus future) mixed-model ANOVA on judgments of general moral character and general judgments of future immorality (see Table 4 for descriptive statistics). Replicating our previous findings, we found a significant time X action condition interaction on both judgements of character, F (1,744) = 80.55, $p < .001$, $\eta_p^2 = .10$, and behavior, F (1,750) = 85.24, $p < .001$, $\eta_p^2 = .10$. As in our previous studies, participants viewed the commission of an immoral act as signaling a future change in the target's moral character and behavior.

We then tested our primary hypothesis that act similarity would moderate the slippery slope effect. We first collapsed across the various individual immoral acts to form four composite indices of predicted behavior: past-theft ($\alpha = .92$), future-theft ($\alpha = .92$), past-harm ($\alpha = .90$), and future-harm ($\alpha = .87$). To assess whether the similarity between the initial immoral act and the predicted behaviors influenced participants' judgments, we conducted a mixed-model ANOVA, with time (past versus future) and predicted domain (theft-related versus harm-related) as the within-subjects variables and act (*immoral act* versus *no immoral act*) and transgression domain (*theft* versus *harm*) as the between-subjects variables. As hypothesized, we observed a significant four-way interaction (time X predicted domain X act X act domain) on participants' likelihood predictions, F(1,751) = 31.82, $p < .001$, $\eta_p^2 = .04$. Consistent with our predictions, the pattern of this interaction was such that the past-future difference was larger for similar (harm-

harm; theft-theft) versus dissimilar (harm-theft; theft-harm) acts (see Table 5). However, there was also a significant "cross-domain" slippery slope effect, such that an individual who

Measure	Domain	Act	Past *M* (SD)	Future *M* (SD)
Good/Bad Person (1 = Extremely Bad, 9 = Extremely Good)	*Theft*	*Immoral Act*	5.25 (1.68)	4.24 (1.44)
		No Immoral Act	6.08 (1.60)	6.26 (1.44)
	Harm	*Immoral Act*	5.14 (1.63)	4.39 (1.53)
		No Immoral Act	5.49 (1.67)	5.54 (1.59)
Likelihood of unethical behavior (1 = Very Unlikely, 9 = Very Likely)	*Theft*	*Immoral Act*	4.95 (2.14)	6.72 (1.72)
		No Immoral Act	3.43 (2.00)	3.62 (2.07)
	Harm	*Immoral Act*	4.70 (2.16)	5.75 (1.99)
		No Immoral Act	3.54 (2.00)	3.71 (2.17)

Table 4. Descriptive statistics for Study 5 judgments of moral character and general likelihood of unethical behavior. Important to note is that the largest past-future differences were in the *immoral act* conditions for both sets of judgments.

Condition	Judged likelihood of different behaviors	Past *M* (SD)	Future *M* (SD)
Theft – no immoral act	Theft-related	2.91 (1.71)	3.08 (1.78)
	Assault-related	1.75 (1.26)	1.72 (1.21)
Harm – no immoral act	Theft-related	2.87 (1.68)	2.88 (1.70)
	Assault-related	2.50 (1.55)	2.56 (1.56)
Theft – immoral act	Theft-related	4.43 (1.94)	5.84 (1.66)
	Assault-related	2.39 (1.75)	2.68 (1.74)
Harm – immoral act	Theft-related	3.73 (1.84)	4.42 (1.93)
	Assault-related	3.21 (1.85)	4.03 (1.73)

Table 5. Descriptive statistics for Study 5 showing the judged likelihood of committing a variety of theft-related and assault-related behaviors in the past and the future. When the agent merely considered theft or assault without acting, there was little change between past likelihood and future likelihood for both types of behaviors. However, when the agent committed the immoral act, there was a "within-domain" slippery slope effect, such that the agent was judged more likely to commit same-domain acts than cross-domain acts in the future than in the past.

committed one type of act (theft or harm) was also seen as more likely to subsequently commit immoral acts of the other type in the future as well, $F(1,753) = 45.87$, $p < .001$, $\eta_p^2 = .06$. Thus, although the slippery slope effect is not restricted to individual domains of moral violations, slippery slope perceptions in moral judgment do appear to be sensitive to more nuanced considerations regarding the specific nature of the immoral acts in question. Consistent with our previous studies, the commission of an immoral act led observers to perceive a change

in the trajectory of an agent's future behavior – but, supporting Hypothesis 3, this effect was particularly pronounced for behaviors that were more similar in nature to the original act.

Study 6

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).

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 is 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 reward for the agent – for example by punishing the agent for the immoral act – should attenuate the slippery slope effect.

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. That is, they experience less negative emotion (e.g., regret, guilt, shame) about the immoral acts they commit, and this corruption of conscience, in turn, makes it easier for them to commit future immoral acts. 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., 2014). If true, then undercutting this perception of corrupted character – for example, by indicating that the agent experienced regret or guilt after the immoral act – should attenuate the slippery slope effect.

Methods

Participants

We recruited 807 participants (55% female, M_{age} = 38.5) from MTurk. We did not include an attention check in this study and did not exclude any participants from 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*) X 2 (regret: *regret*, *no regret*) design. In all conditions, participants were given a general description of an immoral 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).[5]

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-squared test on participants' behavior judgments. We found that the conditions significantly differed in the degree to which each possible future pattern of behavior was selected, χ^2 (12, N = 806) = 274.75, $p < .001$ (see Figure 1). Consistent with our previous studies,

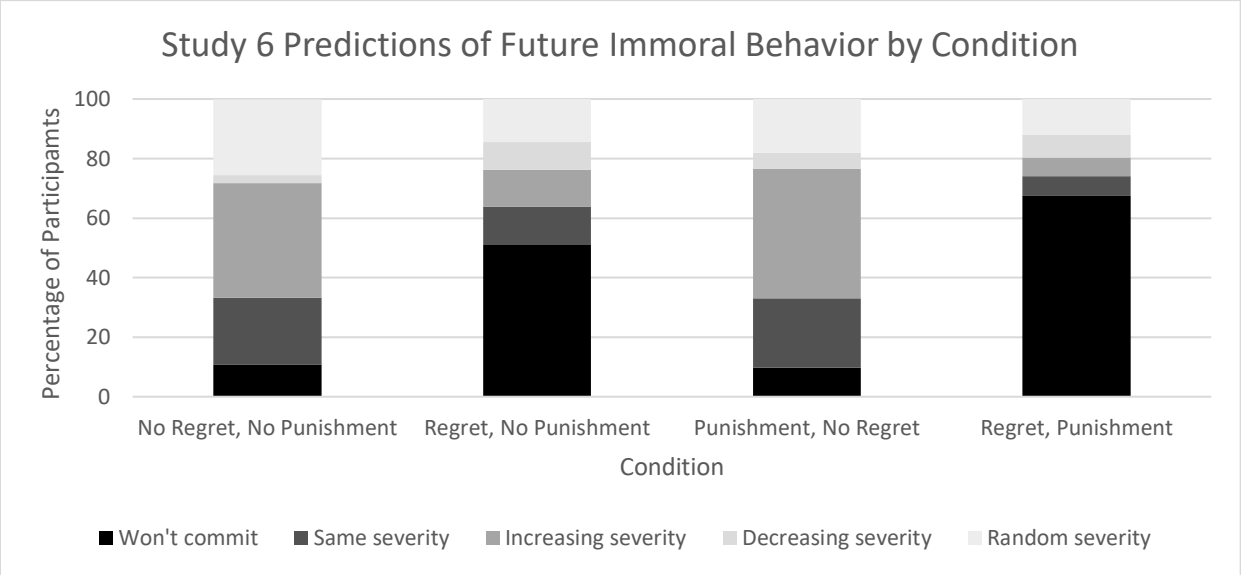

Figure 1. When an immoral agent experienced no regret, participants most often predicted that the agent would commit immoral acts of increasing severity, even if they had been punished for their immoral behavior. When the agent did experience regret, participants most often predicted that the agent would cease committing immoral acts.

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. 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, χ^2 (4, N = 419) = 5.38, p = .25. However, consistent with our predictions, we found a significant main effect of regret (χ^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, χ^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. Further, these results also 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. Further, these results speak against a rewards-based explanation for the slippery slope effect, suggesting that it is not simply the case that observers intuit that "crime

pays." Rather, as hypothesized, the commission of an immoral act is perceived as corrupting a person's moral character and future behavior.

Internal Meta-Analysis

Following recent best-practices recommendations (e.g., McShane & Böckenholt, 2017), we conducted an internal, "within-paper" meta-analysis to determine the average effect size of the slippery slope effect in moral judgment (i.e., the past-future difference for our *immoral act* condition). We used a random-effects model to better extrapolate these effects beyond the current studies to the general population (Hedges & Vevea, 1998). Because we had a nested structure, with measures of (1) moral character, (2) general likelihood of immoral behavior, and (3) likelihood of committing specific immoral acts, we fit a multi-level meta-analysis model (see Konstantopoulos, 2011), specifying nested random effects for study and measure type (moral character, general behavior, specific acts). The average effect size across these studies was Cohen's d = .55 (se = .10, z = 5.31, p < .0001), and the 95% confidence interval for the true effect size was d = .34-.75. We also computed separate average effect sizes for each of our three measures types. All three analyses yielded similar estimates, although the effect sizes were descriptively somewhat larger for judgments of the agent's likelihood of committing specific immoral acts (d = .63, se = .12, z = 5.17, p < .0001) and general likelihood of behaving immorally (d = .56, se = .12, z = 4.54, p < .0001) than for character judgments (d = .42, se = .09, z = 4.36, p < .0001).

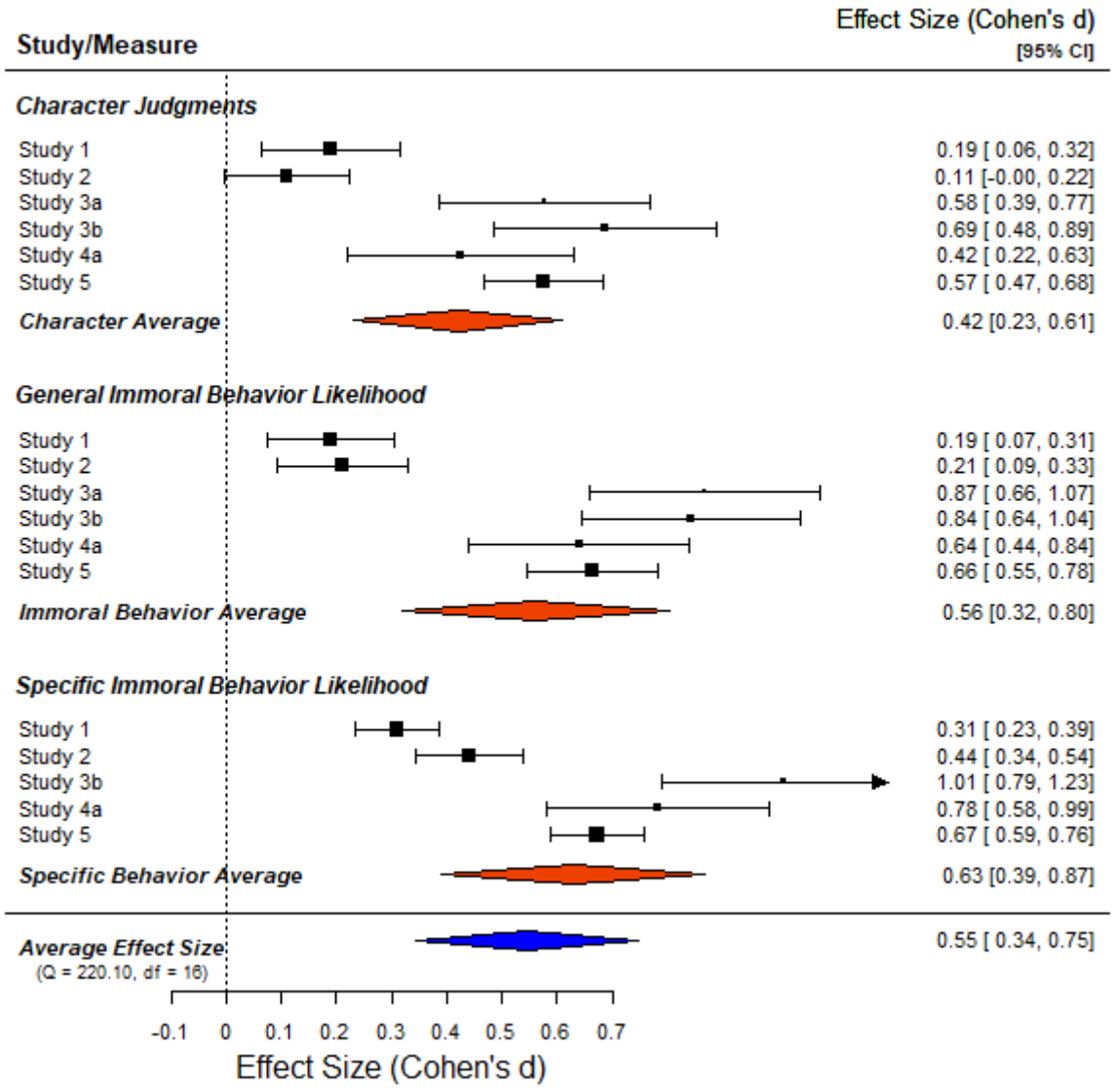

Figure 2. Forest plot of effect sizes of the slippery slope effect. Illustrates the effect size (Cohen's *d*) of the past-future difference for participants in the *immoral act* condition, for each of our three dependent measures. (Studies that did not include past and future judgments are not listed.) Average effect sizes are based on a multi-level meta-analysis with nested random effects for measure type and study.

General Discussion

Across eight studies, we found robust support for the hypothesized slippery slope effect in moral judgment, finding that observers perceive a single immoral act as signaling a change in the future trajectory of an agent's moral character and behavior. 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 in the future after an immoral act (versus in the past before that immoral act; Studies 1-6). 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 2-5). Hypothesis 3: The slippery slope effect exhibits sensitivity to the specific nature of the relationship between the original and 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 5). Further, an agent's future immoral acts are specifically predicted to become increasingly severe over time (Study 6). 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 (Studies 4a and 4b). Further, 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 6).

Connections to Past Theory and 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, Hoffman, Bloom, & Rand, 2016).

The research we report here shows that observers do not simply make judgments of moral character in a static or monolithic fashion, but they also make inferences about how moral character will change over time. Many theories in moral 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 before and after an immoral act. Instead, they exhibit a slippery slope pattern of thinking, expecting individual moral acts to alter the trajectory of an agent's future character.

Our results also connect to research and theory in the area of social cognition more broadly, particularly regarding how people revise and update their impressions of others. While past work has examined how people change their impressions of others in light of new information (e.g., Gawronski & Bodenhausen, 2006; Mann & Ferguson, 2015; Rydell &

McConnell, 2006; Siegel et al., 2018), our studies examine people's impressions of how others may change with time. We believe these findings may offer important novel insights into how observers predict the future behavior and character of others—and, in turn, may therefore modulate people's own behavior when interacting with these target others. For example, our findings highlight a potential cognitive obstacle to interpersonal forgiveness: if an agent transgresses, observers not only judge that agent as being immoral, but also predict that s/he will be *even more immoral* in the future.

Limitations and Future Directions

We recognize two primary limitations of the present studies that can be addressed with future research. First, it is unclear from our studies 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, how might these processes manifest in a court room 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? Future research may wish to investigate slippery slope thinking in real-world situations to better understand how it influences decisions *in situ*.

Second, 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 the realm of 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 (and negative information more generally) may carry greater psychological weight than positive moral acts (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Fiske, 1980; Goodwin & Darley, 2012; Kahneman & Tversky, 1984; Rozin & Royzman, 2001; Taylor, 1991; Wentura, Rothermund, & Bak, 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, Crockett, & Pizarro, 2020; Anderson, Pizarro, & Kinzler, 2018). 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 discussed above, 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 a person's moral behavior.

Endnotes

[1] Although slippery slope arguments are often viewed as inherently fallacious, we note that in this work we focus simply on slippery slope *perceptions*. We make no claims regarding the rationality of these beliefs, but simply seek to understand their cognitive structure and process.

[2] We include the full text of the vignette for Study 1. The full text of the vignettes for subsequent studies can be found at the OSF site for this project https://osf.io/m9qwp/?view_only=513d17cf0e93445b8e4066f0535103cb.

[3] In all studies, we consistently find that an agent who commits an immoral act is perceived as more immoral than an agent who does not attempt (or commit) an immoral act, all *p*s > .001. Given that this effect is already well-established in the literature, we no longer discuss it in our remaining studies.

[4] We recruited 130 participants from MTurk and asked them to rate each of the following behaviors according to how similar they were to all of the other behaviors (from 1 Not at all similar to 7 Extremely similar): cheating on a test, assault, using drugs, vandalism, stealing, cheating on a romantic partner, cheating on taxes, corporate fraud, sexual assault, identity theft, animal abuse. We then examined the mean similarity ratings between each of the behaviors to identify behaviors that were rated as similar to assault and stealing. We found that sexual assault (M = 5.75) and animal abuse (M = 5.28) were rated as most similar to assault, and that cheating on taxes (M = 5.11), corporate fraud (M = 5.23), and identity theft (M = 5.48) were rated as most similar to stealing. We thus selected these items to use in the main study.

[5] Along with this study, participants also answered some questions regarding an unrelated hypothesis (see SM for all measures). 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.

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in Cognitive Science*, *24*(9), 30-39.

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*. Routledge: NY.

Baack, D., Fogliasso, C., & Harris, J. (2000). The personal impact of ethical decisions: A social penetration theory. *Journal of Business Ethics*, *24*, 39–49.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323-370.

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.

BBC ethics guide. (2012). Retrieved from http://www.bbc.co.uk/ethics/introduction/slipperyslope.shtml

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.

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.

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.

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, *136*, 30-37.

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.

Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, *64*(2), 133-152.

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

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353-380.

Dennett, D. C. (1989). *The intentional stance*. Cambridge, MA: 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.

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.

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.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731.

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.

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.

Gray, K. & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, 96(3), 505-520.

Gray, K. & Wegner, D. M. (2011). To escape blame, don't be a hero - be a victim. *Journal of Experimental Social Psychology*, 47, 516-519.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55-66.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*, 613–628.

Hartley, A. G., Furr, R. M., Helzer, E. G., Jayawickreme, E., Velasquez, K. R., & Fleeson, W. (2016). Morality's centrality to liking, respecting, and understanding others. *Social Psychological and Personality Science*, *7*(7), 648-657.

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). New York: 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.

Jennings, M. M. (2011). *Business ethics: Case studies and selected readings*. Mason, OH: 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.

Kahneman, D. & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation*. University of Nebraska Press: Lincoln.

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.

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.

Lakens, D., & Etz, A. J. (2017). Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science*, *8*(8), 875-881.

Lode, E. (1999). Slippery slope arguments and legal reasoning. *California Law Review*, *87*, 1469-1543.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101-121.

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.

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

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition*, *147*, 133-143.

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

Nelson, S. A. (1980). Factors influencing young children's use of motives and outcomes as moral criteria. *Child Development*, *51*(3), 823-829.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*(1), 96-125.

O'Brien, E. (2020). When small signs of change add up: The psychology of tipping points. *Current Directions in Psychological Science*, *29*(1), 55-62.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188.

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). New York, NY: APA Books.

Quoidbach, J., Gilbert, D. T., & Wilson, T. D. (2013). The end of history illusion. *Science, 339*(6115), 96-98.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*, 61–79.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296-320.

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.

Saxe, R. (2012). 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*. East Sussex, UK: Psychology Press.

Schauer, F. (1985). Slippery slopes. *Harvard Law Review*, *99*(2), 361-383.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York, NY: Springer Verlag.

Shweder, R. A., Mahapatra, M., & Miller, J. (1987). Culture and moral development. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children* (pp. 1–83). Chicago: University of Chicago Press.

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.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159-171.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin*, *110*(1), 67–85.

Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research*, *17*, 223–236.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72-81.

Vaish, A., Carpenter, M., & Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. *Child Development*, *81*(6), 1661-1669.

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.

Volokh, E. (2003). The mechanisms of the slippery slope. *Harvard Law Review*, *116*(4), 1026-1137.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York, NY: Guilford.

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.