Social Cognition, Vol. 34, No. 2, 2016, pp. 149–166

THE TIPPING POINT OF MORAL CHANGE: WHEN DO GOOD AND BAD ACTS MAKE GOOD AND BAD ACTORS?

Nadav Klein and Ed O'Brien
University of Chicago

Moral and immoral behaviors often come in small doses. A person might donate just a few dollars to charity or cheat on just one exam question. Small actions create ambiguity about when they might reflect a permanent change in an actor's moral character versus simply a passing trend. At what sum of good or bad behaviors do observers believe that others have transformed for better or worse, when their actions begin to reflect "them"? Five experiments reveal that this *moral tipping point* is asymmetric. People require more evidence to perceive improvement than decline; it is apparently easier to become a sinner than a saint, despite exhibiting equivalent evidence for change. This asymmetry emerges more strongly when targets commit new actions (e.g., begin treating others well or poorly) than when targets cease existing actions (stop treating others well or poorly). This asymmetry in moral judgment fosters inequitable thresholds for reward and punishment.

Keywords: moral judgment, change perception, valence, reward, punishment, tipping point

People evaluate others' character by observing what others do (Gilbert & Malone, 1995). Some behaviors are easy to evaluate, such as when Tom Crist gave his entire lottery winnings of $40 million to charity (ABC News, 2013) or when Bernie Madoff perpetrated the largest accounting fraud in American history (Washington Post, 2008). In everyday life, however, others' actions are not so dramatic, creating more ambiguity in how to evaluate them; character is revealed gradually rather than suddenly. For example, a person might donate a small amount to charity every few years (rather than one large sum), or cheat in school every now and then (rather than in every class, every semester). The current article explores the

Address correspondence to Nadav Klein, University of Chicago, Harris School of Public Policy, 1155 E. 60th Street, Chicago, IL 60637; E-mail: nklein@chicagobooth.edu.

question of *how many* of these smaller instances must be observed before people come to view others as morally virtuous or morally corrupt. At what sum of such behaviors do we think others have appreciably changed for better or for worse, that their good or bad actions do not merely seem like "flukes" but like stable indicators of who they are?

The question of "how many" observations are required to perceive moral change pervades popular interest. All major religions include tenets that prescribe steps to absolve bad behavior and achieve redemption (e.g., Colossians 1:14). The folk-tale of a villain-turned-saint is an attractive one, featured in well-known stories such as Dickens's *A Christmas Carol* (1834) and movies such as *Groundhog Day* (Albert & Ramis, 1993). And acclaimed television shows like *The Sopranos, Breaking Bad*, and *Mad Men* all hinge on the question of how many bad acts a person must commit before they seemingly lose their humanity (Harris, 2014).

Existing research does not provide a clear answer. Traditional studies of change perception focus mostly on the role of attention, and how attentional limitations can undermine one's ability to detect objective changes to a visual scene or object (e.g., Agostinelli, Sherman, Fazio, & Heart, 1986; Simons & Levin, 1997). However, insights from these studies do not seem to bear on the current question. We hold attention to others' behaviors constant and focus on the quantity of behavioral observations that people believe warrants a shift in abstract impressions of moral character, wherein change perceptions are a matter of judgment rather than a matter of accurately discerning objective reality. Some more recent research comes closer to this notion of perception, examining the extent to which people think that the world or their own personalities have grown different over time (e.g., Eibach, Libby, Gilovich, 2003; O'Brien, 2015; Quoidbach, Gilbert, & Wilson, 2013; Wilson & Ross, 2001). However, these studies do not capture the process by which such differences are evaluated; merely perceiving others as different provides little insight into the dynamic nature of a tipping point and the psychology of "when" change is thought to emerge, particularly in terms of morality.

The current article seeks to shed light on these issues. Where do people draw the moral tipping point in evaluating others? That is, how many acts must a person commit or cease before she seems to have substantively transformed in moral character? Because the number itself is irrelevant without context, we explored two broader principles of this judgment process.

First, does the tipping point depend on *valence*? If people simply require a set number of observations before inferring a systematic pattern (Burgers, 1963; Falk & Konold, 1997), then the number of good behaviors that make a saint should be identical to the number of equivalent bad behaviors that make a sinner. This possibility is in line with traditional models of weighted averaging, which tend to focus on the impact of the absolute number of behaviors we observe in shaping our impressions as compared to the impact of any one specific type of behavior (for reviews see Anderson, 1981; Uleman & Kressel, 2013). On the other hand, a robust principle of evaluative judgment suggests that people tend to weight negative information more heavily than positive information (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Kahneman & Tverksy, 1979; Rozin & Royzman, 2001),

and especially so when evaluating the bad and good actions of others (e.g., a person who steals five dollars is perceived as more intensely immoral than a person who donates five dollars seems intensely moral: Fiske, 1980; Reeder & Brewer, 1979; Reeder & Coovert, 1986; Skowronski & Carlston, 1987; Wojciszke, Brycz, & Borkenau, 1993). These findings suggest a lower threshold for believing others have morally declined versus morally improved. In other words, people may be faster to perceive meaningful change for the worse than change for the better in others' character.

Second, does the tipping point depend on *exertion*? Within valence, others could seem to change by actively committing novel behaviors (e.g., "becoming bad" via the number of times a person under-tips at a restaurant) as well as by terminating existing behaviors (e.g., "becoming bad" via the number of times an over-tipper ceases to over-tip). We therefore also tested whether the tipping point depends on the action-oriented properties of others' behaviors. Changes that involve addition tend to be easier to notice than those that involve deletion (e.g., adding versus deleting the leg of a table across two images: Agostinelli et al., 1986). Likewise, people may weight committed behaviors more heavily than ceased ones, reflected in a lower threshold for perceiving change. However, if the tipping point truly depends on the valence of change as posited above, then people should more readily believe others have morally declined as opposed to morally improved regardless of *how* these changes are expressed.

Finally, for both valence and exertion, we sought to generalize any potential patterns across a wide range of parameters and domains. Five experiments were designed to first establish the basic effects (Experiments 1a–1c) and explore the dynamics of tipping point judgments (Experiment 2), and then to test downstream consequences for how people reward versus punish others who are seemingly changing (Experiment 3).

EXPERIMENTS 1A–1C: PERCEIVING MORAL CHANGE

We first conducted three studies to measure the moral tipping point—the number of behaviors that people must observe before coming to believe that others have morally transformed, and whether this number varies as a function of valence (observing moral versus immoral actions) and exertion (others who commit versus cease these actions).

METHOD

All three studies followed a similar design and recruited participants from Amazon's Mechanical Turk in exchange for nominal pay. Participants read about the behaviors of a fictional target in professional settings (Experiment 1a), social settings (Experiment 1b), and academic settings (Experiment 1c). Participants were randomly assigned to 2 (valence: moral versus immoral behavior) × 2 (exertion:

committing versus ceasing the behaviors) fully between-subjects designs. In all studies reported in this article, the sample size was set to at least 40 participants per experimental cell.

Experiment 1a. Participants (N = 201, 41.8% women) read about "Barbara" who works in an American office. All participants were told that Barbara acts like a "typical person" in the way she treats her coworkers during a given week. Most of the time her behavior is neutral, keeping to her work and not affecting others. But sometimes she behaves especially nicely (e.g., holds the door, gives compliments) and other times she behaves especially nastily (e.g., cuts in line, spreads gossip). Participants then imagined they notice a change in Barbara's behavior: more and more consistently over the weeks, Barbara seems to be doing "more of the good things" (*committing good*) versus "more of the bad things" (*committing bad*), or "less of the good things" (*ceasing good*) versus "less of the bad things" (*ceasing bad*). For our main dependent variable, participants reported how many consecutive weeks (from *1 week* to *16 weeks*) of this new behavior would convince them of a substantive change in Barbara's moral character, that her change in behavior was not merely a fluke.

Experiment 1b. Participants (N = 201, 40.8% women) read about "Marti" who enjoys eating at restaurants. All participants were told that Marti usually tips the wait staff, but not always at the same rate. Mostly he tips a standard rate (15–20% of the bill), but sometimes he tips generously (up to 40% of the bill) and sometimes he stiffs the wait staff and tips poorly (as low as nothing). Participants then imagined they notice a change in Marti's behavior: at "more and more meals" he is leaving generous tips (*committing good*) versus poor tips (*committing bad*), or at "fewer and fewer meals" he is leaving generous tips (*ceasing good*) versus poor tips (*ceasing bad*). As in the previous study, participants reported how many consecutive meals (from *1 meal* to *40 meals*) with this new behavior would convince them that Marti's moral character has transformed.

Experiment 1c. Participants (N = 161, 39.7% women) read about "Robert" who signed up for a multi-part longitudinal project in a psychology laboratory. The project allegedly measured Robert's generosity via a series of behavioral studies over time. As described to participants, each of Robert's studies was structured as a modified dictator game, whereby Robert and a partner both received $10 from the experimenters. Robert could behave neutrally (i.e., keep his own share and also let the partner keep their share), selflessly (i.e., give his share to the partner), or selfishly (i.e., take the partner's share). Participants read that over the course of a few months of such studies, Robert's behavior has been relatively balanced. He sometimes behaves neutrally, sometimes selflessly, and sometimes selfishly. However, they then imagined they notice a change in Robert's behavior: in "more and more studies" he is behaving selflessly (*committing good*) versus selfishly (*committing bad*), or in "fewer and fewer studies" he is behaving selflessly (*ceasing good*) versus selfishly (*ceasing bad*). Again, participants reported in how many subsequent studies (from *0 studies* to *20 studies*) this new behavior would need to occur in order to convince them that Robert's moral character has transformed.

Additional Variables. In all experiments, participants completed a manipulation check for valence regarding how ethical they thought the target's new behaviors were (from -3 = *very bad/offensive*, to +3 = *very good/impressive*). We also tested

whether good and bad behaviors may have differed in other ways that affected evaluations of the moral tipping point. Beyond ethicality, there may have been perceived differences in (i) how realistic the change in behavior seemed (rated on a scale from -3 = *very unrealistic*, to +3 = *very realistic*), (ii) how typical the change in behavior seemed (from -3 = *very rare*, to +3 = *very common*), and (iii) how intentional the change in behavior seemed (from -3 = *not planned/intentional at all*, to +3 = *very planned/intentional*; this item was included only in Experiments 1b–1c). Finally, all participants read descriptions of each condition and indicated which they had read in their session (memory check), and they completed a bogus question to which the correct response was to type the word "point" in a text box rather than choose a scale-based item (attention check).

RESULTS

In no study did more than 1.9% of participants fail the attention check or more than 16.0% fail the memory check. Excluding these participants does not affect the results, so the entire samples were retained for all analyses.

Experiment 1a. First, the manipulation worked. An ANOVA testing the effects of valence and exertion on the perceived ethicality of Barbara's new behaviors revealed no main effects, Fs < .41, ps > .37, and the expected interaction, $F(1, 197) = 563.71$, $p < .001$, $\eta_p^2 = .74$. Unsurprisingly and in line with the manipulation, actively doing nice things around the office seemed more ethical ($M = 5.94$, $SD = .99$) than actively doing nasty things around the office ($M = 2.58$, $SD = .67$), $t(97) = 19.81$, $p < .001$, $d = 4.02$, 95% $CI_{difference}$ [3.02, 3.70], just as ceasing nasty behaviors around the office seemed more ethical ($M = 5.70$, $SD = 1.23$) than ceasing nice behaviors around the office ($M = 2.67$, $SD = .83$), $t(100) = 14.57$, $p < .001$, $d = 2.91$, 95% $CI_{difference}$ [2.61, 3.44].

Our primary analysis concerns Barbara's moral tipping point—the number of observations required for participants to believe that she has morally transformed. An ANOVA testing the effects of valence and exertion on this reported number revealed no main effects, Fs < .025, ps > .87, but a significant interaction, $F(1, 197) = 18.83$, $p < .001$, $\eta_p^2 = .087$ (see Figure 1). When actively committing behaviors, Barbara had to do nice things for more consecutive weeks to "become good" ($M = 6.29$, $SD = 3.43$) than the number of weeks of equivalent nasty things required to "become bad" ($M = 4.34$, $SD = 2.52$), $t(97) = 3.22$, $p = .002$, $d = .65$, 95% $CI_{difference}$ [.75, 3.15]. Likewise, when ceasing behaviors, Barbara had to cease nasty actions for more consecutive weeks to lose her badness ($M = 6.24$, $SD = 3.47$) than the number of weeks of ceased nice actions required to lose her goodness ($M = 4.48$, $SD = 2.57$), $t(100) = 2.92$, $p = .004$, $d = .58$, 95% $CI_{difference}$ [.56, 2.96]. Put another way, these results suggest an asymmetry in the moral tipping point that *truly* depends on valence: it takes relatively few bad actions to be seen as "changed for the worse" (whether actively doing bad or ceasing to do good), but relatively many good actions to be seen as "changed for the better" (whether actively doing good or ceasing to do bad).

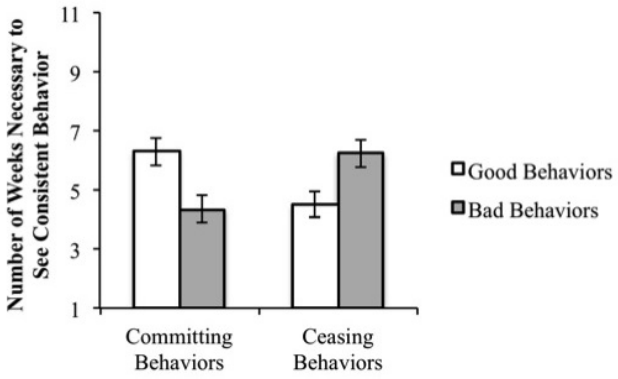


FIGURE 1. Judgments of moral change in Experiment 1a. *Note*. Error bars represent standard errors.

Experiment 1b. Again, the manipulation worked. An ANOVA testing the effects of valence and exertion on the perceived ethicality or Marti's behaviors at restaurants revealed an incidental main effect of valence, $F(1, 197) = 5.05$, $p = .026$, $\eta_p^2 = .025$, no main effect of exertion, $F(1, 197) = .16$, $p = .69$, and the expected interaction, $F(1, 197) = 294.78$, $p < .001$, $\eta_p^2 = .60$. Actively leaving generous tips seemed more ethical ($M = 5.92$, $SD = 1.10$) than actively leaving poor tips ($M = 2.73$, $SD = 1.11$), $t(99) = 14.47$, $p < .001$, $d = 2.91$, 95% $CI_{difference}$ [2.76, 3.63], just as ceasing poor tipping behavior seemed more ethical ($M = 5.48$, $SD = 1.23$) than ceasing generous tipping behavior ($M = 3.00$, $SD = 1.23$), $t(98) = 10.07$, $p < .001$, $d = 2.03$, 95% $CI_{difference}$ [1.99, 2.97].

Our primary analysis concerns Marti's moral tipping point, and here we replicated the same patterns. An ANOVA testing the effects of valence and exertion on the number of required observations revealed an incidental main effect of valence, $F(1, 197) = 4.59$, $p = .03$, $\eta_p^2 = .02$, as well as an incidental main effect of exertion, $F(1, 197) = 10.84$, $p = .001$, $\eta_p^2 = .05$, but also the same critical interaction, $F(1, 197) = 11.48$, $p < .001$, $\eta_p^2 = .055$. When actively committing behaviors, Marti had to tip generously many more times to "become good" ($M = 11.46$, $SD = 6.23$) than the number of poor tips required to "become bad" ($M = 7.63$, $SD = 5.49$), $t(99) = 3.28$, $p = .001$, $d = .66$, 95% $CI_{difference}$ [1.52, 6.15]. Similarly, when ceasing behaviors, Marti had to cease his poor tipping behavior more times to lose his badness ($M = 10.12$, $SD = 6.36$) than the number of ceased generous tips to lose his goodness ($M = 8.36$, $SD = 5.24$), $t(98) = 1.51$, $p = .13$, $d = .31$, 95% $CI_{difference}$ [-.55, 4.07], although this is only directional. In general, the same asymmetry emerged: fewer bad behaviors were required to believe a person has "changed for the worse" than the number of good behaviors required to believe a person has "changed for the better," with little regard for action-oriented nature of this change.

Experiment 1c. All previous results were replicated. The manipulation worked. An ANOVA testing the effects of valence and exertion on perceived ethicality revealed an incidental main effect of valence, $F(1, 197) = 3.97$, $p = .048$, $\eta_p^2 = .025$, no main effect of exertion, $F(1, 197) = 2.17$, $p = .14$, and the key interaction, $F(1, 197)$

TABLE 1. Additional Variables Measured in Experiments 1a–1c

	Exp. 1a: Professional Settings	Exp. 1b: Social Settings	Exp. 1c: Academic Settings
Behavior change is *typical*			
Committing good	4.45 (1.40)$_a$	3.64 (1.61)$_a$	3.95 (1.43)$_a$
Committing bad	4.70 (1.23)$_a$	4.35 (1.19)$_b$	5.03 (1.40)$_b$
Ceasing good	4.15 (1.41)$_a$	4.30 (1.54)$_b$	5.57 (1.27)$_b$
Ceasing bad	4.28 (1.21)$_a$	4.42 (1.37)$_b$	3.88 (1.54)$_a$
Behavior change is *realistic*			
Committing good	5.31 (1.19)$_a$	4.60 (1.68)$_a$	4.73 (1.50)$_a$
Committing bad	5.08 (.97)$_a$	5.00 (1.65)$_a$	5.21 (1.38)$_a$
Ceasing good	4.88 (1.31)$_a$	4.80 (1.60)$_a$	5.52 (1.31)$_a$
Ceasing bad	5.26 (1.10)$_a$	5.18 (1.21)$_a$	4.47 (1.63)$_a$
Behavior change is *intentional*			
Committing good	—	5.04 (1.24)$_a$	4.73 (1.40)$_a$
Committing bad	—	4.78 (1.55)$_a$	4.89 (1.64)$_a$
Ceasing good	—	4.30 (1.79)$_a$	4.21 (1.73)$_a$
Ceasing bad	—	4.84 (1.35)$_a$	4.95 (1.58)$_a$

Note. Standard deviations are in parentheses. Different subscripts denote means that differ at $p < .05$.

= 197.26, $p < .001$, $\eta_p^2 = .56$. Actively selfless behavior seemed more ethical (M = 5.68, SD = .93) than actively selfish behavior (M = 2.84, SD = 1.10), $t(77) = 12.38$, $p < .001$, $d = 2.82$, 95% $CI_{difference}$ [2.38, 3.30], just as ceasing selfish behavior seemed more ethical (M = 5.62, SD = 1.27) than ceasing selfless behavior (M = 3.57, SD = 1.09), $t(80) = 7.87$, $p < .001$, $d = 1.76$, 95% $CI_{difference}$ [1.53, 2.57].

More importantly, we found the same basic asymmetry in our primary analysis concerning Robert's moral tipping point. An ANOVA testing the effects of valence and exertion on the number of required observations revealed an incidental main effect of valence, $F(1, 157) = 3.66$, $p = .058$, $\eta_p^2 = .023$, as well as an incidental main effect of exertion, $F(1, 157) = 9.10$, $p = .003$, but again the same critical interaction, $F(1, 157) = 7.59$, $p = .007$, $\eta_p^2 = .046$. When actively committing behaviors, Robert had to make many more selfless decisions to "become good" (M = 7.61, SD = 4.05) than the number of equivalent selfish decisions required to "become bad" (M = 5.92, SD = 2.96), $t(77) = 2.10$, $p = .039$, $d = .48$, 95% $CI_{difference}$ [.09, 3.29]. Similarly, when ceasing behaviors, Robert had to cease many more selfish actions to lose his badness (M = 8.60, SD = 4.85) than the number of ceased selfless decisions required to lose his goodness (M = 6.88, SD = 3.55), $t(80) = 1.84$, $p = .069$, $d = .41$, 95% $CI_{difference}$ [-.14, 3.58], although this is only marginal. Nonetheless, the same basic pattern emerged: thinking that another person has "changed for the worse" required less equivalent evidence than thinking another person has "changed for the better," with little effect of exertion.

Additional Variables. As mentioned, we also tested whether the moral tipping point was affected by other differences in the observed behaviors beyond valence and exertion. Table 1 presents summary statistics of how typical, realistic, and

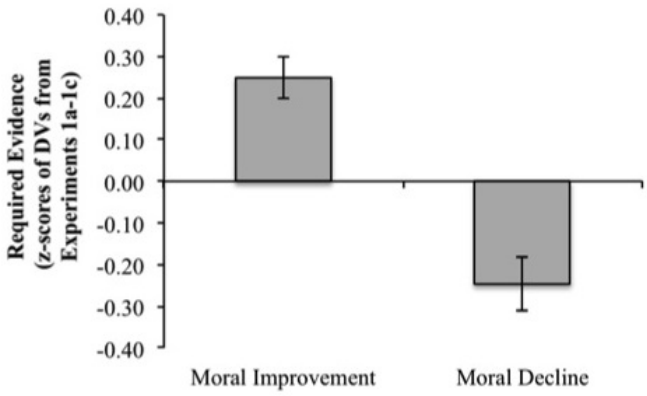


FIGURE 2. Required behavioral evidence for moral improvement versus moral decline in Experiments 1a–1c. *Note*. Error bars represent standard errors.

intentional the new behaviors seemed. No meaningful patterns emerged. There were no differences in perceived realism and intentionality, and incidental differences in perceived typicality such that giving large tips (Experiment 1b), actively making selfless dictator decisions, and ceasing to make selfish dictator decisions (Experiment 1c) seemed least typical, ts > 2.08, ps < .04. Furthermore, entering these items as covariates in ANCOVA did not affect the moral tipping point in any way. In all studies, the key interaction between valence and exertion remained significant when controlling the items (Fs > 6.24, ps < .014). The items did not produce any other significant effects themselves (Fs < 2.36, ps >.12), except for an incidental effect for how realistic the behaviors seemed in Experiment 1b, $F(1, 194) = 3.90$, $p = .05$. In other words, these other variables that may differ across the perception of good versus bad behaviors do not explain our central moral tipping point finding.

In sum, Experiments 1a–1c provide consistent evidence for an asymmetry in perceiving moral change in others. Whether thinking about coworkers at an office, diners at a restaurant, or players in an economic game, people need to observe just a few immoral actions to believe others have "become bad" yet many moral actions to believe they have "become good"; it is easier to become a sinner than a saint. Critically, this effect replicates across different domains and methodological nuances, as well as across type of exertion. Bad has a steeper tipping point than good regardless of whether others actively commit bad versus good behaviors, or cease existing behaviors that are good versus bad (see Figure 2).

EXPERIMENT 2: TIPPING IN REAL TIME

Experiment 2 sought to more dynamically capture the notion of a tipping point by providing a series of identical instances of behavior to participants and assessing their judgments across multiple observations. Participants learned about an actor who began to engage in a new behavior, manipulated along valence and exer-

tion as in previous experiments. After learning about each instance of this new behavior, participants stated whether at that point they were certain that the actor's moral character had changed, or that they needed more information before determining that a tipping point had been crossed. In this way, we could explore our tipping point hypothesis more dynamically, as new information is learned and incorporated into impressions.

METHOD

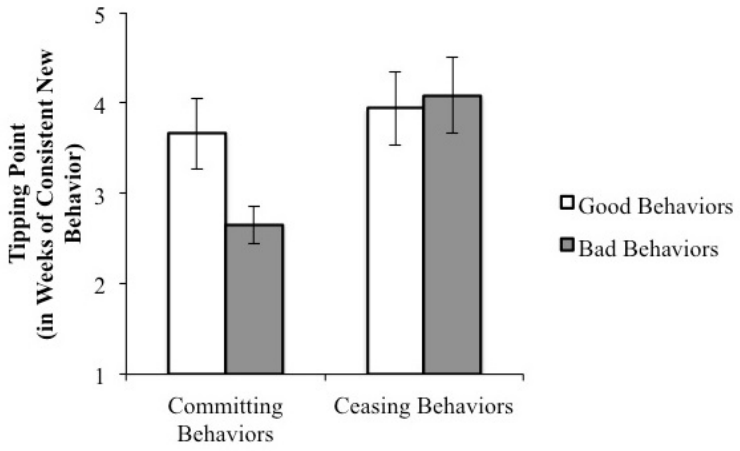
Participants. Participants (N = 202, 38.6% women) were recruited from Amazon's Mechanical Turk in exchange for nominal pay.

Procedure. As in previous experiments, participants were randomly assigned to a 2 (valence: moral versus immoral behavior) × 2 (exertion: committing versus ceasing the behaviors) fully between-subjects design. Borrowing the scenario from Experiment 1a, participants read about "a typical person" who works in an office and behaves mostly neutrally, while at times doing a few especially nice things (holding the door for others, giving compliments, etc.) and a few especially nasty things (cutting in line, spreading gossip, etc.). Again, participants were asked to imagine that this person's behavior started to change. The person began doing "more of the nice things" (*committing good*) versus "more of the nasty things" (*committing bad*), or "less of the nice things" (*ceasing good*) versus "less of the nasty things" (*ceasing bad*).

Unlike previous experiments, information on this behavior change was provided incrementally. Participants began by imagining that they had so far observed this person acting in the new way for a week ("Week 1," the first observation). Participants were asked, "At this point, are you convinced that this person's moral character has 'officially' improved [declined]"? Participants responded by clicking "yes" or "no." If they responded "yes," participation ended. If they responded "no," they proceeded to a new screen. There, they were asked to imagine that they observed this person acting in the new way in the following week too ("Week 2," the second observation). Again, they were asked the yes/no tipping point question, with "yes" terminating participation and "no" continuing on to another week of observing the person's new behavior. This procedure allowed us to more explicitly capture the tipping point itself by recording when participants felt they had seen enough information to warrant a diagnostic shift in their impressions. Participants could have continued to a maximum of 15 weeks of the new behavior before the study ended automatically (this maximum was not breached: responses ranged from 1 week, to 4 participants' judgments of precisely 15 weeks).

RESULTS

For each participant, we extracted the number of weeks needed to reach a tipping point. An ANOVA of tipping point judgments on valence and exertion revealed no main effect for valence, $F(1, 198) = 1.40$, $p = .24$, $\eta_p^2 = .007$, and a main effect for exertion, $F(1, 198) = 5.35$, $p = .02$, $\eta_p^2 = .03$. The interaction did not reach significance,



FFIGURE 3. Dynamic judgments of a tipping point in Experiment 2. *Note*. Error bars represent standard errors.

$F(1, 198) = 2.43$, $p = .12$, $\eta_p^2 = .012$. Closer inspection revealed that when actively committing behaviors, the tipping point comprised more weeks of good behavior ($M = 3.66$, $SD = 2.78$) than weeks of bad behavior ($M = 2.65$, $SD = 1.52$), $t(99) = 2.28$, $p = .025$, $d = .46$, 95% $CI_{difference}$ [.13, 1.90]. This finding replicates our basic effect: participants waited longer (i.e., they demanded more evidence of committing good behavior) before feeling like a person had officially changed for the better, but they were relatively quicker (i.e., they demanded less evidence of committing equivalently bad behavior) before feeling like a person had officially changed for the worse (see Figure 3).

When ceasing behaviors, however, reaching the tipping point did not differ as a function of valence. The number of weeks a person had to cease bad behaviors in order to seemingly improve ($M = 4.08$, $SD = 3.06$) was the same as the number of weeks this person had to cease good behaviors in order to seemingly decline ($M = 3.94$, $SD = 2.85$), $t(99) = .23$, $p = .82$, $d = .05$. This finding does not replicate our basic effect. We believe that this result, taken together with the relatively weak effects between ceasing conditions within two of our three previous experiments (Experiment 1b, $p = .13$; Experiment 1c, $p = .069$), suggests an important distinction in the degree to which committing versus ceasing behaviors elicits a perceived tipping point. We return to this observation in the General Discussion.

EXPERIMENT 3: IMPLICATIONS FOR HOW PEOPLE TREAT OTHERS

An important consequence of the asymmetric moral tipping point may be different thresholds for rewarding versus punishing others. If people require less evidence to accept a change for the worse than for the better in someone else's character, then equivalent behaviors may warrant punishment while not qualifying for

reward, suggesting an inequitable threshold for how people decide to treat others. Experiment 3 tested this possibility.

METHOD

Participants. Participants (N = 161, 41.0% women) were recruited from Amazon's Mechanical Turk in exchange for nominal pay.

Procedure. Participants were randomly assigned into a 2 (valence: moral versus immoral behavior) × 2 (exertion: committing versus ceasing the behaviors) × 5 (within-subjects: instances of behavior) mixed design. Participants read about a high school student named "Bobby." In the *committing good* condition, participants read that Bobby had a history of being an especially bad person (e.g., acting antisocially and being mean to other students). Participants read they were supposed to give him a big reward if he shows signs of "officially" improving and becoming nicer. Participants then read about five instances in which Bobby had an opportunity to be nice to another person and took advantage of this opportunity to be actively nice.

In the *committing bad* condition, participants read that Bobby had a history for being an especially good person (e.g., acting prosocially and being nice to others), and they were supposed to give him a big punishment if he shows signs of "officially" declining and becoming meaner. Participants then read about five instances in which Bobby had an opportunity to be mean to another person and took advantage of this opportunity to be actively mean.

The *ceasing bad* and *ceasing good* conditions were similar to the *committing good* and *committing bad* conditions, respectively, except that here Bobby was described as having an opportunity to be mean or nice to another person, but unlike his past behavior he did not take this opportunity to be mean or nice.

In all conditions, after learning of each of the five instances of behavior participants were asked how likely they were at that point to give Bobby the punishment or the reward on scales ranging from 1 (*definitely not*) to 7 (*definitely would*). Each participant thus provided five ratings of punishment or reward.

In all conditions, the instructions highlighted the difference between behaviors that represent "a fluke" and behaviors that represent a lasting change in moral character, as well as the objectively equivalent threshold for rewarding and punishing Bobby only if the participant can be sure Bobby's improvement or decline is "real and here to stay."

RESULTS

An ANOVA of reward and punishment judgments on valence, exertion, and instances of behavior with repeated measures on the third factor revealed a main effect for valence, $F(1, 157) = 10.24$, $p = .002$, $\eta_p^2 = .06$, a main effect for exertion, $F(1, 157) = 27.60$, $p < .001$, $\eta_p^2 = .15$, and a main effect for instance of behavior, $F(1, 157) = 275.47$, $p < .001$, $\eta_p^2 = .64$. No interactions were found, Fs < 2.50, ps > .11.

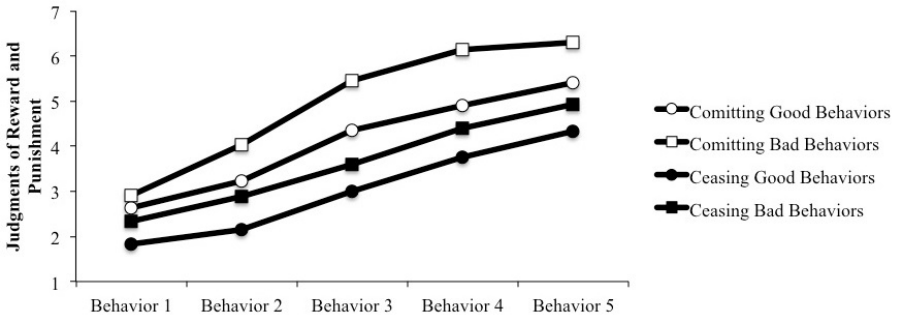


FIGURE 4. Reward and punishment judgments in Experiment 3.

As Figure 4 shows, when actively committing behaviors, participants were more likely to agree to provide punishment for committed bad behaviors than they were to provide a reward for committed good behaviors. This tendency was statistically significant starting from the second instance of behavior, $t(79) = 1.97$, $p = .053$, and thereafter, ts > 2.61, ps $< .011$. This result suggests that reward and punishment judgments mirror the asymmetry in tipping point judgments when actors actively committed behaviors. Moral decline was more readily punished than moral improvement was rewarded.

When ceasing behaviors, reward and punishment judgments generally did not differ between ceasing bad behaviors and ceasing good behaviors. Participants were more likely to reward ceasing bad behaviors than to punish ceasing good behaviors in the second instance of behavior, $t(78) = 2.15$, $p = .035$, and marginally more likely to do the same in the third instance of behavior, $t(78) = 1.88$, $p = .063$, but no other instance of behavior resulted in significant differences, ts < 1.53, ps $> .13$. Thus, when judging ceasing behaviors, participants were directionally more likely to reward moral improvement than to punish moral decline, although this result was noticeably weaker than the opposite asymmetry revealed when judging actively committed behaviors.

Overall, the asymmetry in the moral tipping point extended to decisions of reward and punishment for actively committed behaviors, in line with our proposed framework. Participants were more likely to punish moral decline than to reward identical moral improvement. For ceasing behaviors, slight differences emerged between rewarding moral improvement and punishing moral decline, though these slight differences showed a weak tendency to more readily reward moral improvement than punish moral decline.

GENERAL DISCUSSION

Most people will never donate a lifetime of earnings to others, nor will they ever steal a lifetime of earnings from them. Instead, character is found in the smaller acts of virtue and vice common in daily life. Here we explored "how many" of

these acts must occur before an observer is convinced of an official change in others' character. We refer to this as the *moral tipping point*. Across five experiments and various domains, participants were quicker to diagnose decline but slower to diagnose improvement—despite observing the same amount of evidence. Further, this asymmetry emerged most strongly for actors who actively committed new behaviors. People apparently need to commit just a few bad actions to appear substantively changed for the worse, but need to commit many good actions to appear substantively changed for the better.

THEORETICAL EXTENSIONS

Our findings are conceptually consistent with a large literature on negativity bias in person perception, showing that others' negative actions are weighted more heavily than their positive actions. For example, in a representative article that is most closely related to ours, Reeder and Coovert (1986) described a target person who committed three good behaviors or three bad behaviors (e.g., "stole money from a charity fund"). Participants first rated their impression of the target from "very immoral" to "very moral." Then, participants were presented with one final behavior committed by the target and again rated their impression on the same scale. The authors found a similar asymmetry in terms of participants' latter ratings: one good behavior did little to reduce people's negative impressions leftover from three bad behaviors, whereas one bad behavior indeed reduced people's positive impressions leftover from three good behaviors.

We conceptually replicate these and other similar findings (e.g., Fiske, 1980; Skowronski & Carlston, 1987; Wojciszke et al., 1993), but we believe our framework also helps enrich this classic effect in at least two novel and important ways. First, our findings are the first (to our knowledge) to examine committed versus ceased behaviors. Second, we introduce the effect into the change perception literature, which to date has neither established the judgment process of tipping points nor shown how negativity bias affects dynamic judgments of subjective change in others. Our findings therefore complement existing work: we reveal a negativity bias in the *amount of evidence* necessary to arrive at character judgments (i.e., the point at which neutral targets appear to become good or bad), whereas existing findings highlight the *differential weighting* of different types of evidence people learn about others (i.e., negative information is over-weighted compared to positive information). Together, these sets of findings paint a strikingly complete portrait of negativity bias in impression formation, from our initial characterizations of others to the subsequent resistance in undoing them.

In contrast to actively committing behaviors, tipping point judgments were less conclusive when actors ceased existing behaviors. When directly estimating the tipping point for future behavior (Experiment 1a–1c), participants exhibited an asymmetry in tipping point judgments of both committing and ceasing behaviors. However, when estimating the tipping point dynamically as new information became known, the valence asymmetry remained when actors actively commit-

ted behaviors but disappeared when actors ceased behaviors. The dissimilarity between these results suggests a possible direction for future research, investigating when and why others' actively committed behaviors may influence judgments more strongly than others' ceased behaviors. Intriguingly, our additional results from Experiments 1a–1c (see Table 1) found no incidental differences between committed versus ceased behaviors in terms of how typical, realistic, and intentional they seemed, which suggests a more complex story. One possibility for understanding this difference may lie in the different methodologies used in the present experiments. Seeing each instance of a new behavior may "unpack" these behaviors and enable people to think of them as separate instances (as in Experiments 2–3) rather than a cohesive unit (as in Experiments 1a–1c). "Unpacking" the components of a stimulus has been shown to influence judgments in other research areas, such as estimating of task completion times and forecasting affective reactions to future outcomes (i.e., Kruger & Evans, 2004; Wilson, Wheatley, Meyers, Gilbert, & Axsom, 2000). In similar fashion, explicitly unpacking good or bad behaviors may also affect tipping point judgments, at least for ceasing behaviors.

PRACTICAL IMPLICATIONS

Our finding that the asymmetry emerges even as early as the "tipping point" stage seems uniquely alarming, given that people's conclusion of substantive change in others should (in theory) be the point at which they become most likely to reward or punish others for their observed actions. Experiment 3 bears directly on this implication.

From this perspective, understanding how people perceive moral tipping points might illuminate important asymmetries in public policy and social interactions, especially when the behavior in question involves actively committing actions. In public policy, beliefs about others' ability to change for the better or worse affect laws prescribing sentencing and sentence-commutation guidelines for crimes. At what point does a person who engages in a series of small offenses deserve to be tried for a serious crime? When does a person who exhibits good behavior while serving time in jail earn a shorter sentence? Our findings suggest potential inequity in these common and costly decisions. Similarly in everyday social interaction, having a more stringent threshold for disproving negative than positive first impressions could lead people to refuse to give others a second chance (Fetchenhauer & Dunning, 2010). An asymmetric moral tipping point may even help explain why the reputational costs of social stigmas seem to persist after the socially unacceptable behavior has been corrected (Rodin & Price, 1995).

FUTURE DIRECTIONS

These findings raise a number of new and valuable directions for research. One pertains to expanding the change perception literature. The present research ex-

tends current theories of change by suggesting that people are more likely to detect certain changes than others. While existing research offers insight into people's ability to detect change (e.g., Masuda & Nisbett, 2006; Simons & Levin, 1997), these theories have not suggested precise predictions about the probability that different types of changes will be detected, nor have they necessarily applied to moral domains. Therefore, future work on change perception should consider more socially rich variables (e.g., valence and morality) that might influence people's abilities to detect change, beyond the typical cognitive paradigm (e.g., the addition or deletion of features across two photographs).

Another interesting direction pertains to more specific mechanisms. Negativity bias may not *fully* explain our set of results given the evidence that people are also generally cynical about others' motives (Critcher & Dunning, 2011; Epley, Caruso, & Bazerman, 2006; Fein, 1996; Kruger & Gilovich, 1999; Miller & Ratner, 1998). On the one hand, if people are generally cynical, they may "assume the worst" about others and therefore have a lower burden of proof to perceive changes for the worse than changes for the better. On the other hand, negativity bias also suggests that people may view others' good behaviors as relatively surprising compared to others' bad behaviors. If surprising behaviors are weighted more heavily in judgment (Klayman & Ha, 1987; Stiensmeier-Pelster, Martini, & Reisenzein, 1995), then people should presumably accept a change for the *better* in others based on just a few good behaviors. On this note, none of our control variables (which included items pertaining to expectedness) changed the results in any way. Therefore, while negativity bias is broadly compatible with our findings, fully understanding the psychology of tipping points invites more novel and nuanced insights.

Another future direction relates to the possible connection between tipping point judgments and lay theories of changeability of traits. Research on entity and incremental theories of personality suggests that people differ in the degree to which they believe that moral traits are fixed or changeable (Dweck, Chiu, & Hong, 1995). Believing that a trait can or cannot be changed is likely to affect a variety of judgments (Ross, 1989; Klein, 2015). Future research can better understand whether a global belief that traits are changeable causes people to perceive tipping points more quickly for both moral and immoral behaviors. Alternatively, believing that traits are changeable can also selectively affect tipping point judgments by making them more symmetric across moral and immoral behaviors.

Future research should also explore the generalizability of the effect. For example, do people believe that all immoral behaviors, regardless of domain, "count" similarly toward their tipping point, but that moral behaviors "count" only if they are in the same domain as past moral behaviors? And going beyond morality, might the effect extend to non-moral domains altogether? Given that people are adaptively attuned to social harm and being wronged by others (Cosmides, 1989; Gigerenzer & Hug, 1992), we suspect the asymmetry emerges most robustly in moral (versus non-moral) domains, though this remains an empirical question. Yet another intriguing question regarding generalizability pertains to self judgment: do people also evaluate their own moral and immoral change asymmetrically, or

might self-enhancement motives (Sedikides & Gregg, 2008) flip the moral tipping point when thinking about the self?

Finally, the current findings capture *perceptions* of change. Actual change, to the extent it is quantifiable, may differ from perceptions. Are people who commit a few bad behaviors indeed more likely to commit further bad behaviors than people who commit a few good behaviors likely to commit further good behaviors? Are people who cease a few good behaviors indeed more likely to further cease doing good than people who cease a few bad behaviors likely to cease doing bad? The asymmetric moral tipping may well not accurately reflect reality, suggesting that basic processes of evaluating change might bias people to categorize others as "bad persons" too early—and too unfairly.

REFERENCES

ABC News. (2013). *Tom Crist wins Canadian lottery, donates entire winnings to charity.* Retrieved from www.wjla.com/articles/2013/12/tom-crist-wins-canadian-lottery-donates-entire-winnings-to-charity-98210.html

Agostinelli, G., Sherman, S. J., Fazio, R. H., & Hearst, E. S. (1986). Detecting and identifying change: Additions versus deletions. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 445-454.

Albert, T., & Ramis, H. (1993). *Groundhog Day* [motion picture]. United-States: Columbia Pictures.

Anderson, N. H. (1981). *Foundations of information integration theory*. Boston: Academic Press.

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

Brickman, P., Coates, D., & Janoff-Bulman, R. (1978). Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology, 36*, 917-927.

Burgers, J. M. (1963). On the emergence of pattern of order. *Bulletin of American Mathematics, 63*, 1-25.

Coovert, M. D., & Reeder, G. D. (1990). Negativity effects in impression formation: The role of unit formation and schematic expectations. *Journal of Experimental Social Psychology, 26*, 49-62.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187-276.

Costantini, A. F., & Hoving, K. L. (1973). The effectiveness of reward and punishment contingencies on response inhibition. *Journal of Experimental Child Psychology, 16*, 484-494.

Critcher, C. R., & Dunning, D. (2011). No good deed goes unquestioned: Cynical reconstruals maintain belief in the power of self-interest. *Journal of Experimental Social Psychology, 47*, 1203-1213.

Dickens, C. (1843). *A Christmas carol.* London: Chapman & Hall.

Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A world from two perspectives. *Psychological Inquiry, 6*, 267-285.

Eibach, R. P., Libby, L. K., & Gilovich, T. D. (2003). When change in the self is mistaken for change in the world. *Journal of Personality and Social Psychology, 84*, 917-931.

Epley, N., Caruso, E. M., & Bazerman, M. H. (2006). When perspective taking increases taking: Reactive egoism in social interaction. *Journal of Personality and Social Psychology, 91*, 872-889.

Falk, R., & Konold, C. E. (1997). Making sense of randomness: Implicit encoding as a

basis for judgment. *Psychological Review, 104*, 301-318.

Fein, S. (1996). Effects of suspicion on attributional thinking and correspondence bias. *Journal of Personality and Social Psychology, 70*, 1164-1184.

Fetchenhauer, D., & Dunning, D. (2010). Why so cynical? Asymmetric feedback underlies misguided skepticism regarding the trustworthiness of others. *Psychological Science, 21*, 189-193.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology, 38*, 889-906.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*, 127-171.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21-38.

Harris, M. (2014). The Shonda Rhimes revolution: Finishing what "The Sopranos" started. *Grantland.* Retrieved from http://grantland.com/hollywood-prospectus/shonda-rhimes-scandal-abc/

Kahneman, D., & Tverksy, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-291.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192-238). Lincoln: University of Nebraska Press.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Klein, N. (2015). Insensitivity to gradations in warmth traits constrains beliefs about others' potential for improvement. *Basic and Applied Social Psychology, 37*, 348-361.

Kruger, J., & Evans, M. (2004). If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology, 40*, 586-598.

Kruger, J., & Gilovich, T. (1999). "Naïve cynicism" in everyday theories of responsibility assessment: On biased assumptions of bias. *Journal of Personality and Social Psychology, 76*, 743-753.

Masuda, T., & Nisbett, R. E. (2006). Culture and change blindness. *Cognitive Science, 30*, 381-399.

Miller, D. T., & Ratner, R. K. (1998). The disparity between the actual and assumed power of self-interest. *Journal of Personality and Social Psychology, 74*, 53-62.

O'Brien, E. (2015). Mapping out past and future minds: The perceived trajectory of emotionality versus rationality over time. *Journal of Experimental Psychology: General, 144*, 624-638.

Penny, R. K., & Lupton, A. A. (1961). Children's discrimination learning as a function of reward and punishment. *Journal of Comparative and Physiological Psychology, 54*, 449-451.

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

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

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition, 4*, 1-17.

Rodin, M., & Price, J. (1995). Overcoming stigma: Credit for self-improvement or discredit for needing to improve? *Personality and Social Psychology Bulletin, 21*, 172-181.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96*, 341-357.

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

Schwarz, N., & Bless, H. (1991). Happy and mindless, but sad and smart? The impact of affective states on analytic reasoning. In J. Forgas (Ed.), *Emotion and social judgments* (pp. 55-71). Oxford, England: Pergamon Press.

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 3*, 102-116.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*, 261-267.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal*

of Personality and Social Psychology, 52, 689-699.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*, 131-142.

Stiensmeier-Pelster, J., Martini, A., & Reisenzein, R. (1995). The role of surprise in the attribution process. *Emotion and Cognition, 9*, 5-31.

Uleman, J. S., & Kressel, L. M. (2013). A brief history of theory and research on impression formation. In D. E. Carlston (Ed.), *Oxford handbook of social cognition* (pp. 53-73): New York: Oxford University Press.

Washington Post. (2008). *All just one big lie.* Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2008/12/12/AR2008121203970.html

Wojciszke, B., Brycz, H., & Borkenau, P. (1993). Effects of information content and evaluative extremity on positivity and negativity biases. *Journal of Personality and Social Psychology, 64*, 327-336.

Wilson, A. E., & Ross, M. (2001). From chump to champ: People's appraisals of their earlier and present selves. *Journal of Personality and Social Psychology, 80*, 572-584.

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78*, 821-836.