The Tipping Point of Perceived Change: Asymmetric Thresholds in Diagnosing Improvement Versus Decline

Ed O'Brien and Nadav Klein
University of Chicago

Change often emerges from a series of small doses. For example, a person may conclude that a happy relationship has eroded not from 1 obvious fight but from smaller unhappy signs that at some point "add up." Everyday fluctuations therefore create ambiguity about when they reflect substantive shifts versus mere noise. Ten studies reveal an asymmetry in this first point when people conclude "official" change: people demand less evidence to diagnose lasting decline than lasting improvement, despite similar evidential quality. This effect was pervasive and replicated across many domains and parameters. For example, a handful of poor grades, bad games, and gained pounds led participants to diagnose intellect, athleticism, and health as "officially" changed; yet corresponding positive signs were dismissed as fickle flukes (Studies 1a, 1b, and 1c). This further manifested in real-time reactions: participants interpreted the same graphs of change in the economy and public health as more meaningful if framed as depicting decline versus improvement (Study 2), and were more likely to gamble actual money on continued bad versus good luck (Study 3). Why? Effects held across self/other change, added/subtracted change, and intended/unintended change (Studies 4a, 4b, and 4c), suggesting a generalized negativity bias. Teasing this apart, we highlight a novel "entropy" component beyond standard accounts like risk aversion: good things seem more truly capable of losing their positive qualities than bad things seem capable of gaining them, rendering signs of decline to appear more immediately diagnostic (Studies 5 and 6). An *asymmetric tipping point* raises theoretical and practical implications for how people might inequitably react to smaller signs of change.

Keywords: change perception, entropy, negativity bias, tipping point

Things change. From private fluctuations in one's own health, habits, and moods, to broader cycles in the climate, economy, and society, people must continually respond to the possibility of something being different now than it was before.

Responding effectively, however, requires that people first come to the conclusion that an entity has indeed transformed in some substantive manner. In the current paper, we refer to this moment as the perception of the tipping point—the point at which people begin to perceive noise as signal. Understanding how people make it to this point matters. Decisions to uphold or end a marriage, for example, may depend on the moment at which a sequence of fun dates or nasty fights finally adds up, when those experiences no longer seem circumstantial and are deemed irrevocably diagnostic of more enduring conditions. From changes in the self to changes in the world, people's decisions about how and when to act likely follow from their initial conclusions that change itself has "officially" begun.

Ten studies (N = 2,586) converge across various methodologies, measures, and parameters to suggest a pervasive asymmetry across the psychology of tipping points: things change, but they are perceived as doing so at different rates and in different ways depending on whether people think about change for the better or change for the worse.

Perceiving Change

Perhaps because change is so inescapably common, scholars from diverse fields have long sought to study it. Heraclitus argued around 500 BCE that change was the sole fundamental element of the universe, observing "On those stepping into rivers staying the same, other and other waters flow" ("You cannot step into the same river twice": Heraclitus, *Flux*, 3.1, B12). In the 18th century, Newton's infinitesimal calculus allowed engineers and mathematicians to pinpoint an object's change in space across distant locations (Baron, 1969). And of course still today, historians and economists essentially seek to identify principles that organize change of the past and forecast that of the future (Staley, 2002).

These pursuits allude to the rich tradition of efforts to study objective changes out in external environments. The current paper focuses on how people subjectively perceive such changes. Several

Ed O'Brien, Booth School of Business, University of Chicago; Nadav Klein, Harris School of Public Policy, University of Chicago.

This research was supported by the Willard Graham Faculty Research Award at the University of Chicago Booth School of Business, awarded to Ed O'Brien. Parts of this research were presented at the annual conferences for the Society for Judgment and Decision Making and the Association for Psychological Science. Jaewon Yoon, Miguel Ortega, Ellen Roney, and Alex Kristal helped collect data. Linda Hagen, Anuj Shah, Rick Larrick, Leaf Van Boven, Dan Kahan, George Wu, Yaacov Trope, Chris Hsee, Linda Skitka, Dan Cervone, Amit Kumar, and the Epley lab at Booth shared informative comments and discussion.

Correspondence concerning this article should be addressed to Ed O'Brien, Booth School of Business, University of Chicago, 5807 South Woodlawn Avenue, Chicago, IL 60637. E-mail: eob@chicagobooth.edu

 http://dx.doi.org/10.1037/pspa0000070

research areas elucidate our understanding of change perception, though they highlight distinct psychological features. We review 3 prominent approaches.

By far the most investigated and best understood among these features is the role of attention. People can respond only to changes that they in fact notice, but noticing can be hard. Some changes are missed because they occur too gradually (Campbell, O'Brien, Van Boven, Schwarz, & Ubel, 2014; Simons, Franconeri, & Reimer, 2000). A frog, as the metaphor warns, goes from boiling to boiled not because he likes it warm but because he struggles to notice the point at which warm becomes hot. Other changes are missed simply because most environments comprise an overwhelming presence of stimuli that make competing claims on our attention (Grimes, 1996; Simons & Ambinder, 2005; Wilken & Ma, 2004). Subtle distinctions aside—for example, people sometimes better notice additive versus subtractive change (Agostinelli, Sherman, Fazio, & Hearst, 1986), and in some ways better notice auditory versus visual change (Demany, Trost, Serman, & Semal, 2008)—change perception in this literature is largely viewed as an attentional process (Beck, Rees, Frith, & Lavie, 2001; Pashler, 1988; Rensink, 2002).

A separate literature considers change perception less in terms of tracking change in real time and instead in terms of people's lay theories of how change occurs. From this perspective, change perception reflects an evaluative process grounded in basic judgment principles. For example, when gauging how much a target has changed or will change over time, people typically rely on implicit assumptions about how things *should* change (Dweck, 1999; Eibach, Libby, & Gilovich, 2003; O'Brien, 2013, 2015a; Robinson & Clore, 2002; Ross, 1989; Withey, 1954). In one experiment, participants who completed a study skills course inferred that their test scores must have changed by a sizably wider margin than participants who did not take the course, regardless of how much they actually learned by completing it (Conway & Ross, 1984). These sorts of lay assumptions tend to lack key information, which fosters notoriously miscalibrated beliefs about the trajectory of change (Fischhoff & Beyth, 1975; Gilbert & Wilson, 2000; Kahneman & Snell, 1992). Beyond the goal of accuracy, people have also been shown to actively create the pretense of change, such as by exaggerating the degree to which they have matured as a means of reinforcing positive self-views (Albert, 1977; O'Brien, 2015b; O'Brien & Kardas, 2016; Quoidbach, Gilbert, & Wilson, 2013; Wilson & Ross, 2001). A handful of distinctions aside (e.g., cultural differences in appraising personal growth: Ji, Nisbett, & Su, 2001), principles of change perception in terms of lay theories are well studied and quite robust.

Still another literature focuses on basic learning principles that appear relevant for change perception. People readily form impressions of the world, often in mere moments based on fragments of available information (Albright, Kenny, & Malloy, 1988; Ambady, Bernieri, & Richeson, 2000; Grill-Spector & Kanwisher, 2005; Thorpe, Fize, & Marlot, 1996). Given this rapidity and its implications for stereotyping, researchers have been keen on studying impression "updating"—the extent to which providing new information about a target subsequently shifts one's spontaneous assessment (Chen & Chaiken, 1999; Cone & Ferguson, 2015; Gawronski & Bodenhausen, 2006; Gregg, Seibt, & Banaji, 2006; Griffin & Tversky, 1992). For example, upon seeing a photograph of a disfigured face at Time 1, participants rated this person negatively via an IAT test; but when told the person's scars came from saving children from a house fire, implicit attitudes at Time 2 became positive (Mann & Ferguson, 2015). Studies on impression updating offer general insight into whether impressions will or will not shift in response to new information.

When Changes Emerge

These literatures reveal impressive insights into change perception, but may not fully account for how people specifically assess tipping points—those budding moments at which people come to view the tendencies of an entity as reflecting lasting differences.

Studies in the attention literature are typically restricted to austere procedures (e.g., tracking whether people notice a deleted object across two photographs, or subtle sensory differences in light and sound: Agostinelli et al., 1986; Levine & Shefner, 1981). This affords little understanding of the perception of more ambiguous and socially richer changes whereby "objective" detection is difficult to quantify, such as the dying spark of a romantic relationship or the glimmer of a new friendship. Furthermore, these paradigms cannot account for how people perceive change when attention is held constant (e.g., if a person is asked to consider a full set of dates to ascertain the state of one's relationship).

The literature on lay theories comes closer to a richer range of change domains, though these studies focus almost entirely on self-change (e.g., evaluating one's own personality) with much less emphasis on how we assess changes in other people, social dynamics, and societal trends (for one exception, see Eibach et al., 2003). People often monitor changes in these complex peripheral experiences in addition to their own change. Moreover, these studies near-exclusively measure judgments of surface-level magnitude; participants are shown to evaluate something as different across multiple points in time (e.g., "I'm quite extraverted these days, but I was not last year") with little insight into the dynamic nature of *when* such differences were perceived to have unfolded and emerged.

Studies on impression updating similarly examine a wider assortment of social judgments, but tend to utilize a "one-shot" learning paradigm. Participants are assessed just twice: once in a state of ignorance (e.g., only seeing a scarred face) and again after learning a single bit of information (e.g., the source of the scar). This design is useful for isolating the updating process, but lacks resemblance to real-world change perception in which perceivers observe multiple smaller changes over time. Also, nearly all of these studies are about moral change, such as describing that a person has committed a clearly heroic or abhorrent act and testing the extent to which one's impression comes undone (Fiske, 1980; Mann & Ferguson, 2015; Reeder & Coovert, 1986; Skowronski & Carlston, 1987). Again, change perception across a wider range of domains and experiences, plus more direct insight into *when* such changes are viewed as first emerging, remains unclear.

The Current Research

The current paper addresses these issues. How much evidence of possible change do people need to observe before concluding that a given entity has become "officially" different? Because a literal number is meaningless without context, we sought to explore and establish one broader principle of this process: the role of change direction.

That is, the changes that people must navigate in daily life often signal more than just the existence of differences. Such differences typically indicate if things are changing for the better (improving)

or for the worse (declining). A child's weight, an athlete's performance, and an economy's rank do not merely ebb and flow but meaningfully rise and fall; children become healthier or unhealthier, athletes get hotter or colder, and economies grow richer or poorer. Thus, one intriguing but yet-untested question is how people diagnose tipping points across these thresholds: how much positive evidence must be observed before people conclude "official" improvement, and how does this compare to the amount of negative evidence needed to convince people of "official" decline?

All else being equal, it seems reasonable to assume that the quantity of positive evidence needed to diagnose "official" improvement should be identical to the quantity of negative evidence needed to diagnose "official" decline. For example, if 5 good games in a row signal that a bad player is out of a slump and "officially" improving, then 5 bad games in a row should signal that a good player is in a slump and "officially" declining. This is consistent with theories of weighted averaging, which stress the power of absolute quantities of evidence (regardless of content per se) in shaping global perceptions (see Anderson, 1981; Burgers, 1963; Falk & Konold, 1997; Uleman & Kressel, 2013).

Alternatively, we hypothesize that there may be a robust *asymmetry* in tipping points, such that people may be quicker to diagnose decline (e.g., just a few bad games could quickly signal a diagnostic turn for the worse) versus similar improvement (e.g., just a few good games could be dismissed as premature and still potentially a fluke).

Why? To begin, note that the nature of judging tipping points is sorting fact from fluke. As people track small pieces of compounding evidence, we can therefore assume the following: if the evidence seems flukish (e.g., a rare or unlikely fluctuation), people will wait to see more before concluding a lasting trend; but if the evidence already seems true (e.g., an expected or likely fluctuation), people need to see less. The burden of proof is higher for implausible change than for plausible change. With this in mind, to discern "how much" evidence people demand before tipping for improvement and decline, one might ask how people view the *plausibility* of these changes. There is reason to believe that early signs of a good thing getting worse may seem more plausible than early signs of bad thing getting better, leading people to more readily conclude "official" decline.

Nothing Gold Can Stay: An "Entropy" Hypothesis

We refer to this as an *entropy* account of tracking potential change. Entropy is a term from thermodynamics describing a law of physics by which closed systems must evolve from order to disorder, due to the fact that far more states of disorder are statistically possible (Lieb & Yngvason, 1999). Crudely put, a coffee mug can exist in many broken states but a single unbroken state. Entropy thus reflects a "one-way street" of change: without influence from outside forces, a closed system (e.g., our universe) will grow less organized and eventually deteriorate rather than naturally progress in the other direction, because there are far more ways it can find itself "broken" than "unbroken."

We propose that a conceptually parallel principle might guide the psychology of tipping points. Our central premise is this: on average, people may view good entities as more *capable* of changing for the worse than bad entities are *capable* of changing for the better. When a good entity starts to show signs of possible decline, people may infer that it is now truly breaking down, because it may seem that most good things *can* fall apart, be corrupted, or lose their positive qualities. But people may be more skeptical when a bad entity starts to show the same signs of possible improvement, because it may seem like fewer bad things can just as easily or readily recover. Goodness may seem fickle and finite whereas badness may seem to stick. Thus, people may tip in the face of just a few early signs of decline because good things changing for the worse may strike people as immediately plausible. But people may wait to amass more evidence before tipping for improvement, because true change in this direction may seem less immediately plausible. Below we bolster this hypothesis with observational, empirical, and conceptual support.

Indeed, across many contexts, positive trajectories require various conditions to be met. To become a good student, for example, a bad student needs to display at least some ability, plus work hard and expend effort, plus have resources, plus many other factors; achieving this status requires maintaining many potentially fragile conditions over and over again, which observers may find unlikely for the fate of a typical case. Decline, however, requires nothing and manifests in diverse forms; *everyone* can fail as a student if they so desired, but to thrive requires active, consistent force. In other words, many things need to go right, at the right times, and in the right ways to rise in the ranks; yet just a handful of things needs to go wrong, at any number of times, and in any number of ways to fall. This is further reflected in the fact that, in many domains, trajectories of decline (but not improvement) apply to the majority of individual cases; people can know for sure that even the best athletes will ultimately lose their abilities, but cannot know as surely which struggling athlete will morph into a star. Anything can decline (floors are universal) but not anything can just as easily improve (ceilings are selective). This natural dynamic may lead people to more readily believe initial signs of bad versus good change.

In a pilot study,[1] we asked people about how changes in life typically unfold and confirmed that they view trajectories of decline as much more common, more likely, and generally more plausible than trajectories of improvement. Further, this entropy

[1] We asked 100 participants on Amazon's Mechanical Turk (M_{age} = 34.81, SD_{age} = 11.53; 48.00% Female; 79.00% Caucasian; $0.25 payment) to indicate their beliefs about change. First, they read: "Things in life change. Sometimes things change for the better (improve) and sometimes things change for the worse (decline). In general, when thinking about how change typically unfolds. . . ." Then, they were randomly assigned to either *decline* or *improvement* conditions (*n*s = 50) and responded to 6 questions on a scale from 1 (*not at all*) to 10 (*extremely*): "How easy do you think it is for declines [improvements] to occur?"; "How likely do you think it is for declines [improvements] to occur, generally speaking?"; "How common do you think it is for declines [improvements] to occur, generally speaking?"; "How much do things have to really 'come together' in order for declines [improvements] to occur, generally speaking?"; "How suddenly do you think declines [improvements] can occur, generally speaking?"; and "How much do you agree with the following statement? All good things [bad things] must come to an end." Ratings were combined into a scale with the "come together" item reverse-coded, such that higher scores reflect greater perceived plausibility (α = .70). In line with our theorizing, the results of an independent-samples *t* test showed that people perceive decline to be significantly more plausible (M = 6.56, SD = 1.28) than improvement (M = 5.78, SD = 1.39), $t(98)$ = 2.90, p = .005, d = .58, 95% $CI_{difference}$ [.25, 1.31]. This holds when controlling participant age, sex, and ethnicity via Univariate GLM analyses: condition remained significant, $F(1, 95)$ = 7.57, p = .007, η_p^2 = .07, 95% $CI_{difference}$ [.21, 1.32], and there were no effects of demographics, Fs < .37, ps > .545, ηs_p^2 < .004 (demographics had no meaningful effects in any study, so they are not discussed further). Our main studies unpack these broad descriptions of decline and improvement at greater length and include more clearly defined information. These findings simply provide further validation for our theorizing about entropy perceptions.

account is conceptually supported by Reeder and Brewer's (1979) classic relational model of how people reason about and predict others' range of possible behaviors from learning about others' traits. The most relevant proposition of this extensive model is that the lower another person falls on some trait's continuum, the more that observers will predict a restricted range of the person's possible behaviors in relation to the floor. For example, people might intuitively predict that a bad basketball player can only play poorly and cannot play well; the predicted range of a bad player's possible behaviors is narrow. Conversely, people may predict that someone with high basketball-playing abilities can play well *as well as* poorly (e.g., s/he can easily play badly if motivated to do so, or if s/he is undermined by one of many possible situational factors); the predicted range here is wide. These implicit rules of inference have been well supported since the model was put forth (Devine, Hirt, & Gehrke, 1990; Kim, Dirks, Cooper, & Ferrin, 2006; Reeder & Fulks, 1980; Skowronski & Carlston, 1989), and in fact have been found to specifically emerge for how people judge others at the lower end of "socially desirable capacities" such as having poor competitive abilities or poor social skills (Gawronski, 2004; Reeder, 1993; Reeder, Pryor, & Wojciszke, 1992; Ybarra, 2002). This suggests the continuum is effectively about bad-to-good rather than "low" or "high" standings per se,[2] in support of our current reasoning. Although this model largely addresses one-shot judgments of extreme magnitude, all in the interpersonal domain (e.g., judging a very shy person's capacity for acting in an very extraverted way), we can extrapolate to tipping points and the extent to which people may be persuaded by smaller compounding signs of change. Because, a priori, people believe good entities are capable of displaying and developing bad features but not vice versa, people's post hoc reactions to features that actually materialize should follow in kind. In line with an entropy account, a diagnosis of decline may emerge rather quickly, because positive entities should seem quite capable of also taking negative forms (after all, most individual cases in most domains can and ultimately may begin some form of descent). A diagnosis of improvement may emerge relatively slower, because negative entities should seem generally less capable of taking positive forms (fewer just as easily ascend to the top). In short, initial signs of positive change should invite greater skepticism than initial signs of negative change, as reflected in people demanding more evidence of possible change for the better than for the worse before being convinced of its instantiation (even if the nature and quality of evidence is otherwise similar).

An Additional Account? The Role of "Alarm"

To review: Assessing tipping points involves separating fact from fluke. Thus, evidence that seems plausible from the start should facilitate a quick tipping point, whereas implausible evidence should be met with greater skepticism and postpone a tipping point. We have outlined various reasons to believe that evidence of decline will seem more immediately plausible, resulting in the hypothesized asymmetry. Early signs of decline may be overweighted in tipping point judgments because they have higher "truth" value.

And yet, there are many other ways in which negative evidence could be overweighted simply by virtue of being bad (for reviews of valence asymmetries and the power of negativity, see Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001). Most relevant for tipping points, note that although people may well acknowledge that declines are more common, easier to achieve, and more plausible than improvements—shown by our pilot results and review of support—few of us presumably *want* these declines to occur. Signs of decline are not merely signs of truth or falsehood; they are also, by definition, problematic. In turn, research on risk aversion and related phenomena suggests that negative events can elicit more extreme reactions than similar positive events in any number of ways (e.g., absorb more attention, get processed in richer detail, stir stronger emotions), reflecting an adaptive aid for identifying threats and problems; people tend to be hypersensitive to costs (Gonzalez & Wu, 1999; Loewenstein, Weber, Hsee, & Welch, 2001; Nesse, 2005; Tversky & Kahneman, 1992). Through such mechanisms, perceivers may come to view negative evidence through a more intense or exaggerated lens, leading them to tip more readily for decline. In short: early signs of decline may be overweighted in tipping point judgments *also* because they have higher "alarm" value (e.g., extreme reactions may lead people to infer that something serious must be underway, or motivate people to form this conclusion "just in case"). If people tip quickly, it may be unclear whether this is because change for the worse seems more immediately credible and therefore lowers people's burden of proof (as in an "entropy" account) or because change for the worse seems more immediately consequential and therefore heightens people's sensitivity to the evidence (as in an "alarm" account).

We intend not to dismiss this latter possibility; note that it provides even more support for the asymmetry and simply highlights another explanation for why it might occur. These factors may well work together (e.g., it is quite possible to view signs of change as both plausible and alarming) and amplify the asymmetry in some of our studies and in everyday life. As we will discuss and show throughout the paper, we attempt to at least partly dissociate these factors so to shed light on a novel and yet-untested "entropy" effect in contributing to change perception *beyond* "alarm"-related reasons alone.

Overview of Studies

We have proposed an investigation of the psychology of tipping points, defined as those first points at which observed noise becomes perceived signal—when deviations in an experience can no longer be dismissed as passing flukes and people come to conclude that lasting change has "officially" begun.

[2] One exception is morality: highly moral others are the ones who are viewed as more behaviorally restricted (i.e., as capable of only acting morally), whereas immoral others are viewed as capable of both bad and good behaviors (Reeder & Brewer, 1979; Reeder et al., 1992). Although the underlying rule of inference differs, the predicted effect remains consistent with our framework: people overweight immoral actions relative to moral actions (see Klein & O'Brien, 2016). However, this is thought to be a unique case of pure hedging, because the costs of errors in moral assessment can be especially extreme (e.g., mistaking aggressive others for peaceful others, regardless of the evidence). Given this special status, along with a large literature on the many other unique nuances of moral domains of judgment versus other domains (for one review see Greene & Haidt, 2002), the current paper mainly focuses on a wide variety of other (yet-untested) domains of change.

Many important changes unfold in such a fashion. Rarely, for example, does a single obvious event ignite or extinguish the spark of a relationship, or mark the start or end of ability. Rather, people observe many smaller moments that at some point (in one's mind) "add up" to warrant a diagnostic shift.

The current paper has 3 goals. First, we seek to simply document and establish how people diagnose these tipping points, specifically across the direction of change. Second, we seek to test our hypothesis that these tipping points may be asymmetric, even when the kind of evidence on either side is roughly constant: people may tip more readily for decline than for similar improvement. Third, we seek to explore whether an *entropy effect* (a general perception that, on average, good things are truly more capable of decline than bad things are capable of improvement) may represent an added, unique contributor.

We begin by documenting the basic effect via assessing the amount of evidence that people demand before hitting a tipping point (Studies 1a, 1b, and 1c) and via their overt judgments of change over time (Study 2). Next, we replicate and expand this effect to downstream behavior as reflected in real bets with real money on trend continuation into the future (Study 3). We then more directly explore mechanisms (in addition to utilizing some methods and measures throughout Studies 1–3 that speak against pure "alarm"). We did this by first ruling out various exogenous differences that could account for the effect (Study 4a, 4b, and 4c), and then by directly testing competing accounts (Studies 5–6). Study 5 tested whether people still readily tip for decline when the costs for tipping quickly are made high; if so, this goes against a pure "alarm" explanation and suggests people may also tip quickly because they feel confident that signs of decline denote true change (even when it may be very costly to form this diagnosis too soon), akin to an entropy account. Most persuasive, Study 6 tested whether people no longer readily tip when decline in a domain is uncommon, not common; if so, this suggests people indeed draw on the *plausibility* of otherwise alarming signs when trying to assess "official" decline. Together, these studies suggest that an entropy effect may wield its own important influence on tipping points.

Studies 1–3: Establishing the Tipping Point of Change

First, we sought to establish the basic effect. Participants either imagined or actually experienced shifts from a bad or good baseline across many different domains, and indicated the first point at which they felt that these fluctuations reflected lasting change. We hypothesized they would tip more quickly when evaluating compounding evidence for decline than when evaluating similar evidence for improvement.

In this and all studies, we predetermined sample sizes of at least 50 participants per condition, as a general rule of thumb and expanding on similar research (e.g., Cone & Ferguson, 2015, about 40–50 per cell; Eibach et al., 2003, about 20–30 per cell; Wilson & Ross, 2001, about 20–30 per cell). This aligns with the results of a power analysis (Faul, Erdfelder, Lang, & Buchner, 2007) of our pilot study, which recommends 48 per cell for 80% power. We report all manipulations and measures. No participant was excluded. Our file drawer is small and consistent.[3] Data and materials are readily available upon request.

Study 1

A Pervasive Asymmetry

Participants read about the progression of different experiences and reported the amount of evidence they demanded before inferring "official" change. We sought to widely replicate and generalize this tipping point process. We assessed many targets of judgment, in many domains, on various scales of change, among multiple populations.

Study 1a: Changes in frequency. Participants were recruited from 2 different sources as a means of internal replication and for enhancing external validity, comprising a total $N = 239$. In this study, change was scaled in terms of frequency (i.e., the absolute number of Times X needs to be observed before a person hits the tipping point).

First, we recruited 101 participants from Amazon's Mechanical Turk ($M_{age} = 36.93$, $SD_{age} = 13.35$; 44.60% Female; 64.34% Caucasian) to complete a study about judgment in exchange for $0.40. They were randomly assigned to either *decline* ($n = 51$) or *improvement* ($n = 50$) conditions. Each participant read and evaluated 8 scenarios in randomized order, one at a time (see Appendix A.I). The 8 different scenarios pertained to changes in athletic performance, academic performance, physical health, mood, luck, habits, friendship, or personality. For each, "Decline" participants were given a positive starting point and were asked to indicate how many observations of the next 10 must be bad for them to be convinced in lasting change for the worse. For example, in the athlete scenario, participants indicated how many games of the next 10 must a *good* athlete play *poorly* to signal the start of official decline. "Improvement" participants simply rated the opposite (e.g., how many games of the next 10 must a *bad* athlete play *well* to signal the start of official improvement). After, all participants reported demographic information.

Second, we recruited 139 adult passersby ($M_{age} = 33.17$, $SD_{age} = 12.59$; 57.60% Female; 59.00% Caucasian) at the Museum of Science and Industry (MSI) in Chicago, to participate in exchange for candy. Again, participants were randomly assigned to *decline* ($n = 70$) or *improvement* ($n = 69$) conditions, and completed the study in the same way.

[3] All studies reported in this paper were conducted just once as they are presented. Throughout the review process, we dropped 5 other studies that showed consistent effects (3 studies similar to Study 4 and 2 studies similar to Study 5). Early on, we also conducted 2 unreported studies that showed null effects. Their designs were unique from all others. In the first, we created an animation in which grey circles pop out one-by-one and form into a moving assembly line across the screen. Participants were told to imagine that each circle represented time passing in a typical life, and grey circles were "neutral" events (no other descriptors). Eventually a red circle popped out, which represented either a "good" or "bad" event depending on condition (no other descriptors). More and more red circles kept popping out, until eventually all were red. The second study was similar, except we used a static matrix of grey circles. Piece by piece, one of the circles would turn red, until eventually the entire matrix was red. In both studies, participants had to stop the animation at the first point they felt a meaningful pattern was emerging. Participants did not click any faster when the red circles were framed as bad versus good. It is unclear whether these null results are meaningful failures to replicate versus something unusual about these designs. For example, the task was highly abstract, participants essentially knew a priori that a pattern would form, and the changes were not randomized (when creating the stimuli, we manually chose which ones would turn red in a way that simply felt to us like a growing pattern). This represents our entire file drawer. Please contact us for any additional information about any of these studies.

Study 1b: Changes in duration. This study followed similar procedures except we sought to extend the patterns to a different scale of change. Building on frequency, here we tested duration; whereas Study 1a assesses the question of *how many* instances must be observed to elicit a perceived tipping point, Study 1b assesses *how long* an observed shift must be maintained. The same 8 domains were used, except they were adapted and framed around duration (see Appendix A.II). For example, in the adapted athlete scenario, "Decline" participants read that a good athlete was beginning to show signs of playing worse, and were asked to rate how long this shift in performance must persist to signal the start of official decline (each scenario was rated on a scale from 1 = *just a short time*, to 10 = *a long time*). "Improvement" participants read and rated the converse scenarios in the same way (e.g., rating how long a bad athlete's positive signs must persist to signal the start of official improvement).

We recruited a total $N = 228$, again from 2 different sources: 100 participants from Amazon's Mechanical Turk ($M_{age} = 36.29$, $SD_{age} = 13.01$; 50.00% Female; 82.00% Caucasian) in exchange for \$0.40 (*decline*, $n = 50$; *improvement*, $n = 50$); and 128 adult passersby at the MSI ($M_{age} = 37.01$, $SD_{age} = 15.29$; 51.60% Female; 68.80% Caucasian) in exchange for candy (*decline*, $n = 64$; *improvement*, $n = 64$).

Study 1c: Changes in magnitude. Whereas Study 1a assesses *how many* and Study 1b assesses *how long*, Study 1c assesses change in terms of magnitude: *how much* of a change must be observed from one period to another. The same 8 domains were used, but they were adapted and framed around magnitude (see Appendix A.III). For example, in the adapted athlete scenario, "Decline" participants indicated how much of a drop-off from a previous good season must a player make to signal the start of official decline (each scenario was rated on a scale from 1 = *10% worse*, to 10 = *100% worse*, with each scale point increasing by 10 percentage points). "Improvement" participants completed converse procedures (e.g., rating how much of a spike from a previous bad season must a player make to signal the start of official improvement, from 10% better to 100% better).

We recruited a total $N = 225$, again from 2 different sources: 101 participants from Amazon's Mechanical Turk ($M_{age} = 31.89$, $SD_{age} = 9.93$; 37.60% Female; 67.30% Caucasian) in exchange for \$0.40 (*decline*, $n = 51$; *improvement*, $n = 50$); and 124 adult passersby at the MSI ($M_{age} = 39.27$, $SD_{age} = 15.35$; 62.10% Female; 62.90% Caucasian) in exchange for candy (*decline*, $n = 60$; *improvement*, $n = 64$).

Results and Discussion

Study 1a (Frequency). Of central interest, we first report the overall effect with the 8 domains collapsed into a scale ($\alpha = .78$) as a way to most clearly depict the effect. Data were submitted to Univariate GLM analyses, with population and condition as fixed factors and this tipping point as the dependent variable. As predicted, there was no main effect of population, $F(1, 236) = .63$, $p = .43$, $\eta_p^2 = .003$, no interaction, $F(1, 236) = .01$, $p = .90$, $\eta_p^2 < .001$, and only a main effect of condition: participants needed to observe fewer negative instances out of 10 before they perceived official decline ($M = 5.27$, $SD = 1.37$) than the number of equivalent positive instances to perceive official improvement ($M = 6.53$, $SD = 1.23$), $F(1, 236) = 54.61$, $p < .001$, $\eta_p^2 = .19$, 95% $CI_{difference}$ [.93, 1.60].

Unpacking this finding across individual domains via Multivariate GLM analyses robustly replicates this overall effect (see Table 1, Column 1). Again, for each scenario, there were no main effects of population, Fs < 3.33, ps $> .07$, $\eta s_p^2 < .014$, nor were there any interactions, Fs < 1.16, ps $> .28$, $\eta s_p^2 < .005$; we observed only the predicted main effects of condition, Fs > 7.08, ps $< .008$, $\eta s_p^2 > .03$. In other words, across *every* domain and regardless of subject population, people expressed asymmetric tipping points in their reactions to compounding evidence of change: they became convinced of lasting decline significantly more quickly than of lasting improvement.

Study 1b (Duration). Data were analyzed in the same way. Again, first in terms of overall effects with the 8 domains collapsed into a scale ($\alpha = .73$), all results robustly replicate. There was no main effect of population, $F(1, 224) = .30$, $p = .56$, $\eta_p^2 = .003$, no interaction, $F(1, 224) = .13$, $p = .72$, $\eta_p^2 = .001$, and only the predicted main effect of condition: as hypothesized, bad signs needed to persist for a significantly shorter length of time ($M = 5.61$, $SD = 1.19$) before participants perceived official decline compared with the length of time that good signs needed to persist ($M = 6.82$, $SD = 1.21$) to perceive official improvement, $F(1, 224) = 57.42$, $p < .001$, $\eta_p^2 = .20$, 95% $CI_{difference}$ [.90, 1.54]. This overall effect again robustly held across individual domains (see

Table 1
Study 1: Tipping Points for Each Life Domain Across Conditions

	Study 1a: *Frequency* of change required to elicit tipping point		Study 1b: *Duration* of change required to elicit tipping point		Study 1c: *Magnitude* of change required to elicit tipping point	
Life domain	Decline	Improvement	Decline	Improvement	Decline	Improvement
1. Athletic	5.37_a (1.88)	6.08_a (2.04)	5.93 (2.13)	6.38 (2.07)	4.82_m (2.13)	5.60_m (2.27)
2. Academic	5.29_b (2.23)	6.27_b (2.06)	5.78_A (2.26)	6.33_A (2.20)	4.08_n (1.92)	6.25_n (2.07)
3. Health	5.44_c (2.16)	6.64_c (2.06)	5.34_i (2.17)	7.02_i (2.20)	4.80_o (2.03)	5.58_o (2.40)
4. Mood	5.73_d (2.44)	6.55_d (2.33)	5.94_B (2.22)	6.57_B (2.08)	5.32_D (2.13)	5.95_D (2.16)
5. Luck	5.03_e (2.43)	6.68_e (2.24)	4.81_j (2.62)	7.30_j (2.68)	4.79_p (2.44)	6.01_p (2.62)
6. Habits	6.30_f (2.43)	7.21_f (1.97)	5.82_C (2.05)	6.51_C (2.31)	4.52_q (1.91)	5.64_q (1.97)
7. Friendship	4.13_g (1.89)	6.19_g (2.29)	4.75_k (2.26)	6.51_k (2.24)	4.73_r (1.88)	6.00_r (2.28)
8. Personality	4.87_h (2.31)	6.66_h (2.63)	6.49_l (2.22)	7.92_l (1.86)	5.06_s (1.84)	5.85_s (2.36)

Note. Means and standard deviations (in parentheses) are presented. Means sharing lower-case subscripts differ at $p < .01$; means sharing upper-case subscripts differ between $p = .01$ and .05. Regardless of domain or scale of change, people broadly require less evidence of possible change before diagnosing the start of "official" decline, versus the amount of equivalent evidence that they require before diagnosing the start of "official" improvement.

Table 1, Column 2): the main effect of condition was significant for a full 7 of the 8 scenarios, Fs > 4.07, ps $< .045$, $\eta s_p^2 > .02$ (athletics was marginal in the predicted direction, $F(1, 224) = 3.21$, $p = .07$, $\eta_p^2 = .01$). The same asymmetric tipping point emerged, here in terms of duration.

Incidentally, population did have a single main effect, such that Turk participants waited longer than MSI participants to form conclusions about *luck*, $F(1, 224) = 5.74$, $p = .02$, $\eta_p^2 = .03$, 95% $CI_{\text{difference}}$ [.15, 1.58]. However, there were no other main effects of population, Fs > 2.57, ps $< .11$, $\eta s_p^2 > .01$, and no interactions, Fs < 2.46, ps $> .12$, $\eta s_p^2 < .01$. As such, these results do not bear on our central findings in any meaningful way.

Study 1c (Magnitude). Data were analyzed in the same way, and all findings again replicate. When collapsing across domains ($\alpha = .82$), there was no main effect of population, $F(1, 221) = 1.61$, $p = .21$, $\eta_p^2 = .007$, no interaction, $F(1, 221) = 2.21$, $p = .14$, $\eta_p^2 = .01$, and again only the hypothesized main effect of condition: goodness needed to drop down by significantly less ($M = 4.77$, $SD = 1.32$) for participants to perceive official decline compared with the margin that badness needed to spike up ($M = 5.86$, $SD = 1.45$) for participants to perceive official improvement, $F(1, 221) = 37.02$, $p < .001$, $\eta_p^2 = .14$, 95% $CI_{\text{difference}}$ [.76, 1.49]. This held for each domain (see Table 1, Column 3): the main effect of condition was highly significant for every scenario, Fs > 4.79, ps $< .03$, $\eta s_p^2 < .02$. We found the same robust asymmetry in tipping points, here in terms of magnitude.

Otherwise in Study 1c, we observed no main effects of population for any scenario (as expected), Fs < 2.54, ps $> .11$, $\eta s_p^2 > .01$. However, although there were no interactions for 7 of the 8 scenarios (also as expected), Fs < 1.87, ps $> .17$, $\eta s_p^2 < .008$, we did observe a significant interaction for *health* such that Turk participants showed the asymmetry but MSI participants did not, $F(1, 221) = 5.00$, $p = .03$, $\eta_p^2 = .02$. Nonetheless, this sole departure stands in stark contrast to the consistent effect throughout the studies.

Together, Study 1 reveals a pervasive asymmetry in the amount of evidence that people demand before they infer lasting change. For example, a handful of failed exams, bad games, gained pounds, and lost bets lead people to judge academic performance, athletic ability, health, and luck as having "officially" changed for the worse; and yet corresponding positive developments are more likely to be dismissed as flukes rather than seen as substantive change for the better. As hypothesized, people are quick to diagnose decline but slow to diagnose improvement, despite facing similar evidence for change. This asymmetric tipping point emerged regardless of whether people judged various different domains and experiences, and across scales and subject populations.

Study 2

Framing the Same Exact Evidence

One limitation of Study 1 is that the phrasing in some scenarios may have created unequal evidence to begin with (e.g., participants may have inferred a more intense event from "sad day" vs. "happy day"). This scaling issue pervades research on valence asymmetries given that negative entities often lack positive counterparts (e.g., germs) and objective quantities can apply differently to different individuals (e.g., losing or gaining $1.00 can feel different despite numerical equivalency: see Rozin & Royzman, 2001).

Testing for the asymmetry across many domains and contexts (as in the current paper) helps rule out the idiosyncratic effects of any one phrasing, inference, or stimulus. Nonetheless, Study 2 sought to more directly address this issue, serving as an additional way to establish the basic effect. We moved beyond hypothetical scenarios to reactions to change "in real time." Participants across conditions viewed the same graph depicting the same ambiguous trajectory of data, but this objectively identical trend was framed as either suggesting change for the worse or change for the better. We hypothesized that the same noisy information would be interpreted as more reflective of "real" change simply when framed as depicting potential decline than as depicting potential improvement.

Participants. We recruited 500 participants from Amazon's Mechanical Turk ($M_{\text{age}} = 31.76$, $SD_{\text{age}} = 9.97$; 41.20% Female; 74.40% Caucasian) to complete a study about information processing in exchange for $0.25.

Procedure. Participants perused a static image of a graph, allegedly depicting real fluctuations in the United States economy from the 1950s through 2000s. They were told the data were from the "Economic Volume Index," allegedly capturing one specific and novel subset of economic quality. We used fictional data (unbeknownst to participants) to avoid reliance on outside knowledge of actual change in other metrics of the economy.

The depicted trend line was very slightly and ambiguously sloped downward over time, and participants' task was to interpret it (see Appendix B). To ostensibly aid in their interpretation, we gave some additional details: participants were told either that lower values suggest things are getting worse on this particular index (*decline* condition), or that lower values suggest things are getting better (*improvement* condition). Objectively, note that all participants saw the same exact information. They rated this specific trend—not the economy in general but the given data about this particular subset of the economy. Specifically, they rated the extent to which they interpreted the given graph as "showing a clear trend rather than just noise"; how much it was cause for "alarm" or "celebration" (phrase dependent on condition); and how much people in general should feel "worried" or "relieved" about it (phrase dependent on condition). Each of these 3 items was rated on a scale from 1 (*definitely no*) to 10 (*definitely yes*), comprising our dependent measures.

Our full design included 2 additional between-subjects manipulations beyond this improvement/decline factor. In other words, participants were randomly assigned into 1 of 6 conditions, with the above description representing just 2 of them (all ns ≥ 59). First, other participants followed the same procedures except saw mirror images of the graphs, such that trend lines sloped identically *upward* and they were told that *higher* values are bad or good. This rules out incidental effects of confusion when slopes violate traditional associations for decline as "down" and improvement as "up." Second, we manipulated the domain of the graph: other participants followed all procedures except their graphs depicted the "Health Volume Index," a fictional subset of public health. This again addresses incidental effects of outside knowledge, and more broadly adds to our goal of testing for the generalizability of the basic tipping point effect across many experiences. Finally, participants reported demographic information and responded to 2

manipulation checks, one regarding the meaning of the slope they were shown (*Improvement*; *Decline*) and the other regarding the general domain of the index (*Economy*; *Health*).

Results and Discussion

Only 6.60% of participants (33 of 500) failed the manipulation check for slope meaning, and only 1.6% of participants (8 of 500) failed the manipulation check for domain. Eliminating them does not affect any result, so they are retained.

The 3 ratings were collapsed into a scale ($\alpha = .86$), with higher scores reflecting greater interpreted "realness" of the change. Data were submitted to Univariate GLM analyses, with slope meaning, slope direction, and domain as fixed factors, and this scale as the dependent variable. Again, our hypothesis predicts *only* a main effect of slope meaning, such that participants should interpret decline as more "real" than equivalent improvement regardless of the direction of the depicted slope and the domain of change.

This is indeed what we found. There was no main effect of slope direction, $F(1, 492) = 2.00$, $p = .16$, $\eta_p^2 = .004$; no main effect of domain, $F(1, 492) = .51$, $p = .48$, $\eta_p^2 = .001$; no 2-way interactions (Fs $< .57$, ps $> .45$, η_p^2s $< .001$); and no 3-way interaction, $F(1, 492) = 1.25$, $p = .26$, $\eta_p^2 = .003$. As predicted, we only observed a main effect of slope meaning: graphs were indeed interpreted as more "real" when they were framed as depicting decline ($M = 6.57$, $SD = 2.20$) than when the same exact graphs were framed as depicting improvement ($M = 5.84$, $SD = 1.92$), $F(1, 492) = 16.10$, $p < .001$, $\eta_p^2 = .03$, 95% $CI_{\text{difference}}$ [.38, 1.11]. This asymmetric pattern robustly held within each condition, as reflected via pairwise comparison analyses (see Figure 1): When thinking about the economy, participants were more likely to infer "real" change when the same trend was framed as decline versus improvement regardless of whether the trend was downward-sloping ($F(1, 492) = 5.60$, $p = .018$, $\eta_p^2 = .011$) or upward-sloping, although the upward pattern was not statistically significant ($F(1, 492) = .87$, $p = .35$, $\eta_p^2 = .002$). Likewise for health, participants were more likely to infer "real" change from the same trend framed as decline versus improvement regardless of a downward slope ($F(1, 492) = 3.89$, $p = .049$, $\eta_p^2 = .008$) or an upward slope ($F(1, 492) = 7.59$, $p = .006$, $\eta_p^2 = .015$).

Some nuances aside, the same basic asymmetry emerged. People more readily tip in diagnosing an "official" lasting trend when trying to interpret possible change for the worse versus change for the better—even when staring at objectively identical evidence.

Study 3

Betting on Bad Luck

Study 3 again involves real-time change, utilizing yet another method to further establish the effect. This method again rules out incidental effects that render the stimuli objectively different; like Study 2, the evidence in Study 3 will not depend on phrasing.

Our findings so far suggest that people may be more likely to predict a negative trend today will continue into the *future* versus an equivalent positive trend, if negative signs seem more lastingly "real" and positive signs more flukish. For example, hitting a tipping point for perceiving a declining relationship or economy implies people have come to believe the entity will remain bad for some time. This maps onto downstream consequences: people may be overly quick to actually *act* in the face of possible decline (e.g., by dumping a partner or selling stocks), before amassing a fuller range of evidence.

Study 3 explored this possibility via an extremely conservative test. Participants were invited to make bets using real money of their own on a future outcome, given some evidence of a possible trend in the present: either the continuation of a present streak of bad or good outcomes. We made clear that the chance of either continuation was random. Therefore, by any objective standard, people should be equally likely to take the bet. Our observed asymmetry instead suggests that people may be more likely to bet on bad luck.

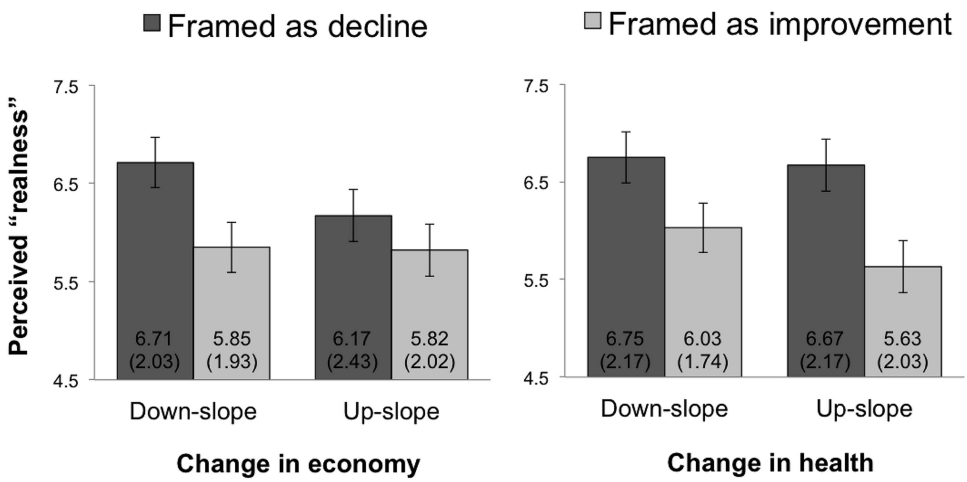


Figure 1. Study 2: Interpretations of the provided charts of change across conditions, with higher scores reflecting a greater perception that the data is more than mere noise. Means and standard deviations (in parentheses) are presented, with error bars signifying ±1 standard error. The same exact trajectory was interpreted as more "real" simply when it was framed as decline, regardless of its slope or the purported domain of change.

Participants. We recruited 200 participants from Amazon's Mechanical Turk (M_{age} = 34.45, SD_{age} = 11.33; 53.00% Female; 76.50% Caucasian) to play a short and simple "Guessing Game" in exchange for $0.20.

Procedure. Upon sign-up, participants were unaware of any gambling component or that they would have a chance to make money beyond their participation payment of $0.20. They were told that we (as researchers) wanted to test out our survey technology, which involved playing a very simple card flipping game. They were instructed that they would make 3 completely independent guesses, each with equally randomized outcomes.

First, all participants were shown a set of 8 separate cards laying face down and were asked to simply pick one. They were told that on the next screen the survey would reveal whether they had picked a card with the phrase "WIN" or with the phrase "LOSE" on its front side, both of which were described as being equally randomized in the set. After selecting a card, participants were randomly assigned to either *improvement* (n = 102) or *decline* (n = 98) conditions. "Improvement" participants were shown to have drawn a winning card whereas "Decline" participants were shown a losing card, which was rigged by the experimenters. Participants proceeded to a new screen displaying a new set of 8 cards, described as independent and unrelated to the prior set but still with equally randomized outcomes. They made their second pick. "Improvement" participants drew another win and "Decline" participants another loss, again rigged by the experimenters. Participants clicked to a new screen to make their final pick, to be made in the same way.

At this point, however, participants were shown a surprise pop-up message before being able to pick the third card. They were offered a wager as a bonus chance to make more money for participating. "Improvement" participants were asked if they wanted to bet that their third pick would be a winning card. They were told if they took the bet and it was a win, we would award them $0.10 on top of their set payment (small in absolute terms but meaningful in the context of the study, a 50% net bonus); but that if they took the bet and it was not a win, we would subtract $0.10 from their set payment (a 50% net loss). We measured the percentage of participants who actually took the bet, serving as our dependent variable. After making their choice, they in fact did not make a third pick and proceeded to the rest of the study (see next paragraph). "Decline" participants saw the same prompt, except they were offered to bet whether their third pick would be a loss. Again, we measured the percentage of participants who took the bet. For all participants, we added a clear disclaimer in the pop-up message explaining that their card picks were not rigged in any way; that we (as researchers) had no vested interests in the outcome of the bet; and that we would implement "a number of additional technological features" before they made their third pick to help "absolutely ensure" they felt it was a fair draw.[4]

Upon proceeding to the rest of the study, all participants were debriefed and were asked to describe their thought process for taking or not taking the bet via an open-ended text box. They then reported demographic information and responded to a manipulation check about their first 2 draws (*Both wins*; *Both losses*). All were awarded a $0.10 bonus.

Results and Discussion

Only 1.50% of participants (3 of 200) failed the manipulation check. Eliminating them does not affect any result, so they are retained.

Data were submitted to binary logistic regression analyses, with condition as the predictor and gambling decision for the third draw (1 = bet; 2 = no bet) as the dependent variable. Again, all else being equal, half of participants within each condition should take the bet, given that the third outcome was presented as completely independent and randomized. Instead, we observed the hypothesized effect of condition: although exactly 50.00% of "Improvement" participants (51 of 102) opted to gamble on a third *win* in a row, a full 73.50% of "Decline" participants (72 of 98) opted to gamble on a third *loss* in a row, B = 1.02, SE = .30, Wald = 11.33, p = .001. The observed odds ratio coefficient was 2.77, 95% $CI_{Exp(B)}$ [1.53, 5.01]—that is, participants were about 3 times more likely to gamble on the continuation of bad luck than on the continuation of identical good luck. One way to interpret this difference is directly in line with our proposed entropy account: a hot streak seems fickle and runs out quickly (i.e., positive signals are perceived as a passing fluke) whereas an identical cold streak seems here to stay (i.e., negative signals are perceived as a lasting trend), so much that people are willing to risk real money on it. This emerged in a setting in which the odds of either outcome were objectively random, providing a highly conservative conceptual replication of our basic tipping point effect.

Open-ended responses provided additional insight. An independent coder combed each response. The most interesting comparison is between participants who opted not to bet on a third win in a row versus did opt to bet on a third loss in a row, to compare the extent to which they mentioned the streak as converse justification. "Improvement" participants who did not bet on continued improvement cited reasons like, "I just felt like I had a lucky streak, and it was bound to end soon," and "I thought since 2 out of the 3 were wins, the next would probably not. Even though I know each time you have a 50/50 chance" (about half wrote along these lines: 52.94%, 27 of 51). Yet "decline" participants used the same justification for doing the opposite and betting on continued decline, citing reasons like, "I was not having any luck so most likely I would lose again," and "Even though it is randomized I thought I would go with my gut. I mean . . . it was a loss the first two times" (about half: 58.30%, 42 of 72). The same past experiences were implicated as more "real" and lasting when hinting at a potential cold streak versus hot streak.

Study 3 again confirms the same basic effect from another perspective, expanding our framework to people's own actual behavior in response to real-time potential change.

Has "Entropy" Contributed to These Results? Reviewing Studies 1–3 and Previewing Studies 4–6

So far, the hypothesized asymmetry has proven robust: whether imagining various kinds of fluctuations, or interpreting charts of generational trends, or reacting to budding streaks, just a few bad

[4] Across conditions, the extent to which participants may have fully believed the study design is unclear (e.g., believing that Amazon would actually allow us to issue penalties). But note how such possibilities make it *less likely* to find any effects of the manipulation (e.g., all participants in all conditions should take the bet if they are generally skeptical), therefore bolstering the conservative nature of this particular test of our hypothesis.

signs led people to readily conclude lasting change for the worse, but people were less convinced by otherwise similar good signs of change for the better.

Why? We have proposed that one novel contributor may reflect an entropy effect: good things may seem more immediately capable of declining than bad things seem capable of improving, thereby lowering the burden of proof for diagnosing a true pattern. That is, initial signs of decline may be overweighted because they come with higher "truth" value, leading to a quicker judgment in sorting what seems real versus flukish. But it may *also* be that initial signs of decline are overweighted because they have higher "alarm" value: negative signs are, by definition, problematic, so people may react in any number of more extreme ways (e.g., ways related to risk aversion) that lead to a quick tip.

Some aspects of Studies 1–3 suggest that more than *pure* "alarm" contributed to the results. For example, the asymmetry was identical regardless of whether people evaluated changes in themselves versus another person, whereas factors like risk aversion may predict a stronger effect for the self (or other targets) because of varying costs. A wider and wider replication of the same tendency across many targets and domains may suggest that people are simply bringing to bear a general belief about how things typically change. Moreover, the asymmetry remained even when various features were truly equivalent: Studies 2 and 3 help rule out some incidentally objective inflation of our negative stimuli (stimuli and evidence of change were identical), and Study 3 accounts for risks and costs (chances of losing were equally threatening) while speaking directly to entropy beliefs (participants' open-ended responses, and the logic of betting on trend *continuation*). Pure "alarm" may predict an attenuated asymmetry under such conditions since these important features were, at least objectively speaking, equally extreme.

If more than pure "alarm" is at work, we posit that entropy may be an important missing piece. However, although our pilot study showed that people indeed view decline as more plausible than improvement, we ultimately cannot tell the extent to which this contributed to Studies 1–3. Our studies so far have not fully ruled out the various possible effects of the consequential nature of negativity. More direct support is needed for the potentially unique contribution of entropy, both in the current paper and future research.

Studies 4–6 were designed to provide some of this evidence. First, we sought to more directly rule out some stimulus features that could incidentally explain the effect.

Study 4

Ruling out Exogenous Differences

In Study 4, we examined whether the asymmetry is explained by factors outside of our primary framework as proposed. We tested whether the basic effect varies as a function of evaluating changes in oneself versus changes in others (Study 4a); change via addition versus subtraction (Study 4b); and intended versus unintended changes (Study 4c). If the asymmetry replicates across these diverse parameters, this might provide more evidence for a generalized belief about change. Our prior studies suggest that it does (e.g., Studies 1–4 have included both self and others), but here we systematically tested this possibility.

All studies resembled Study 1, except we manipulated the given dimension in the stimulus text. For each, participants were recruited from Amazon's Mechanical Turk to complete a study about judgment in exchange for $0.25.

Study 4a: Self versus other. We assigned 202 participants (M_{age} = 33.14, SD_{age} = 1.53; 42.60% Female; 77.20% Caucasian) into a 2 (valence: *improvement* or *decline*) × 2 (target: *self* or *other*) between-subjects design (all *n*s ≥ 49). Again in the broader spirit of generalizability, each participant evaluated 3 different scenarios (pertaining to changes in work performance, health, or luck; see Appendix C.I). "Decline" participants read about signs of possible decline, and rated how long these signs would need to persist to convince them of official change for the worse (1 = a *very short time*, to 10 = a *very long time*). "Improvement" participants read the converse. Moreover, participants imagined these changes unfolding in themselves (e.g., "Imagine you have incorrectly guessed the outcomes of a few games involving coin flips") or someone else (e.g., "Imagine another person has . . ."). Otherwise the scenarios were identical. After, all participants reported demographic information and completed a manipulation check of the self/other prompt by which they had to recall the target of their scenarios via forced-choice (*I was told to imagine myself in the scenarios*; *I was told to imagine another person in the scenarios*).

Study 4b: Additive versus subtractive. We assigned 222 participants (M_{age} = 32.77, SD_{age} = 10.76; 41.00% Female; 72.07% Caucasian) into a 2 (valence: *improvement* or *decline*) × 2 (change type: *additive* or *subtractive*) between-subjects design (all *n*s ≥ 53). Again maintaining all procedures, participants rated 3 different scenarios (pertaining to changes in community news, work feedback, or emotion (see Appendix C.II). Along with our valence manipulation, each of the changes was described as unfolding by the compounding *presence* of evidence (e.g., improvement via added goodness: "You've started to get more and more great feedback") or *absence* of evidence (improvement via subtracted badness: "You've started to get less and less poor feedback"). Otherwise all materials were identical, and participants made the same tipping point ratings for each. For the manipulation check of the additive/subtractive prompt, participants had to recall the specific nature of the change via forced-choice (*Good things became more and more frequent; Good things became less and less frequent; Bad things became more and more frequent; Bad things became less and less frequent*).

Study 4c: Intended versus unintended. We assigned 213 participants (M_{age} = 33.41, SD_{age} = 11.46; 41.80% Female; 73.70% Caucasian) into a 2 (valence: *improvement* or *decline*) × 2 (change type: *intended* or *unintended*) between-subjects design (all *n*s ≥ 49). Following the same procedures, participants rated 3 scenarios (pertaining to changes in athletic performance, romantic life, or academic performance: see Appendix C.III). Each piece of evidence for change was framed as either being intended (e.g., a sports team that starts winning because they are trying to play well, or starts losing because they are trying to tank) or unintended (e.g., a sports team that happens to win despite trying to tank, or happens to lose despite trying to play well). Otherwise each pair of scenarios was identical. Participants again rated the length of time that the evidence must persist before they would hit a tipping point for "official" improvement or decline. For the manipulation check of the intended/unintended prompt, participants had to recall the specific nature of the change via forced-choice (*The subject was trying/hoping to make things go well, and lo*

and behold, things were going well; The subject was trying/hoping to make things go badly, but despite this, things were going well; The subject was trying/hoping to make things go poorly, and lo and behold, things were going poorly; The subject was trying/hoping to make things go well, but despite this, things were going badly).

Results and Discussion

Study 4a (self/other). Only 6.40% of participants (13 of 202) failed the manipulation check. Eliminating them does not affect any result, so they are retained.

Data were submitted to Multivariate GLM analyses, with valence and target as fixed factors and tipping point judgments for each domain as dependent variables. There was one incidental main effect of target such that participants tipped faster for themselves than for others in the "work" scenario, $F(1, 198) = 3.44$, $p = .065$, $\eta_p^2 = .02$). There were no other main effects of target, Fs < 2.32, ps $> .13$, $\eta s_p^2 < .01$, nor any interactions, Fs < 1.96, ps $> .16$, $\eta s_p^2 < .01$. Of critical interest and replicating the basic effect, we observed only a main effect of valence, for all domains, Fs > 9.85, ps $< .002$, $\eta s_p^2 > .05$.

Pairwise comparisons confirm the same asymmetry regardless of the imagined target of change, across domains: When thinking about the *self*, participants were quicker to conclude work-related decline ($M = 4.85$, $SD = 2.04$) than work-related improvement ($M = 6.24$, $SD = 1.94$), $F(1, 198) = 13.29$, $p < .001$, $\eta_p^2 = .06$, 95% $CI_{\text{difference}}$ [.64, 2.15]; health-related decline ($M = 5.00$, $SD = 2.18$) than health-related improvement ($M = 7.33$, $SD = 1.92$), $F(1, 198) = 37.61$, $p < .001$, $\eta_p^2 = .16$, 95% $CI_{\text{difference}}$ [1.58, 3.08]; and luck-related decline ($M = 6.52$, $SD = 2.72$) than luck-related improvement ($M = 7.82$, $SD = 1.98$), $F(1, 198) = 8.86$, $p = .003$, $\eta_p^2 = .04$, 95% $CI_{\text{difference}}$ [.44, 2.16]. Likewise when imagining change in *others*, work-related decline tipped quicker ($M = 5.16$, $SD = 1.79$) than work-related improvement ($M = 6.94$, $SD = 1.92$), $F(1, 198) = 21.61$, $p < .001$, $\eta_p^2 = .10$, 95% $CI_{\text{difference}}$ [1.03, 2.54]; health-related decline ($M = 5.78$, $SD = 1.72$) than health-related improvement ($M = 7.36$, $SD = 1.76$), $F(1, 198) = 17.26$, $p < .001$, $\eta_p^2 = .08$, 95% $CI_{\text{difference}}$ [.83, 2.32]; and luck-related decline ($M = 6.86$, $SD = 1.82$) than luck-related improvement ($M = 7.50$, $SD = 2.11$), $F(1, 198) = 2.14$, $p = .145$, $\eta_p^2 = .01$, 95% $CI_{\text{difference}}$ [−.22, 1.50] (in the hypothesized direction). Overall, these results add additional evidence for our tipping point asymmetry, suggesting it does not meaningfully depend on target.

Study 4b (additive/subtractive). Only 12.16% of participants (27 of 222) failed the manipulation check. Eliminating them does not affect any result, so they are retained.

Data were again analyzed the same way, and the same general patterns emerged. There were no main effects of change type, Fs < 2.03, ps $> .16$, $\eta s_p^2 < .009$, nor any interactions, Fs $< .88$, ps $> .35$, $\eta s_p^2 < .004$; only the main effect of valence, for news and feedback, Fs > 7.44, ps $< .007$, $\eta s_p^2 > .03$ (emotion was in the hypothesized direction but was not significant, $F(1, 218) = 2.75$, $p = .098$, $\eta_p^2 = .01$).

Pairwise comparisons confirm the same asymmetry: For *additive* change, participants were quicker to conclude news-related decline ($M = 5.84$, $SD = 2.50$) than news-related improvement ($M = 6.59$, $SD = 2.01$), $F(1, 218) = 3.43$, $p = .065$, $\eta_p^2 = .02$, 95% $CI_{\text{difference}}$ [−.05, 1.54]; feedback-related decline ($M = 4.37$, $SD = 2.03$) than feedback-related improvement ($M = 6.39$, $SD = 2.18$), $F(1, 218) = 27.39$, $p < .001$, $\eta_p^2 = .11$, 95% $CI_{\text{difference}}$ [1.26, 2.79]; and emotion-related decline ($M = 5.89$, $SD = 2.26$) than emotion-related improvement ($M = 6.43$, $SD = 1.92$), $F(1, 218) = 1.74$, $p = .189$, $\eta_p^2 = .008$, 95% $CI_{\text{difference}}$ [−.26, 1.33] (consistent with the null main effect, not statistically significant). For *subtractive* change, news-related decline ($M = 6.21$, $SD = 2.11$) tipped quicker than news-related improvement ($M = 7.04$, $SD = 1.89$), $F(1, 218) = 4.02$, $p = .046$, $\eta_p^2 = .02$, 95% $CI_{\text{difference}}$ [.014, 1.63]; feedback-related decline ($M = 4.93$, $SD = 2.08$) than feedback-related improvement ($M = 6.43$, $SD = 1.91$), $F(1, 218) = 14.60$, $p < .001$, $\eta_p^2 = .06$, 95% $CI_{\text{difference}}$ [.73, 2.82]; and emotion-related decline ($M = 5.84$, $SD = 2.25$) than emotion-related improvement ($M = 6.26$, $SD = 2.15$), $F(1, 218) = 1.06$, $p = .30$, $\eta_p^2 = .005$, 95% $CI_{\text{difference}}$ [−.39, 1.24] (again not significant). These results generally confirm that the asymmetry holds regardless of whether change is expressed via the compounding presence or absence of evidence.

Study 4c (intended/unintended). Only 12.70% of participants (27 of 213) failed the manipulation check. Eliminating them does not affect any result, so they are retained.

Data were analyzed the same way, and the asymmetry again emerged across all conditions. There were incidental main effects of change type for both the "athletic" and "romantic" scenarios such that participants tipped faster for intended change, Fs > 5.79, ps $< .017$, $\eta s_p^2 > .03$. There was no main effect of change type for "academics," $F(1, 209) = 1.02$, $p = .31$, $\eta_p^2 = .005$, and most critically, no interactions, Fs < 2.46, ps $> .12$, $\eta s_p^2 < .01$; only the main effect of valence, for all domains, Fs > 27.79, ps $< .001$, $\eta s_p^2 > .12$.

Pairwise comparisons again confirm the basic effect: For *intended* change, participants were quicker to conclude athletic-related decline ($M = 5.29$, $SD = 1.86$) than athletic-related improvement ($M = 6.16$, $SD = 1.55$), $F(1, 209) = 6.75$, $p = .01$, $\eta_p^2 = .03$, 95% $CI_{\text{difference}}$ [.21, 1.54]; romantic-related decline ($M = 4.55$, $SD = 2.02$) than romantic-related improvement ($M = 6.32$, $SD = 1.84$), $F(1, 209) = 19.21$, $p < .001$, $\eta_p^2 = .08$, 95% $CI_{\text{difference}}$ [.97, 2.57]; and academic-related decline ($M = 4.92$, $SD = 2.28$) than academic-related improvement ($M = 6.43$, $SD = 1.58$), $F(1, 209) = 15.75$, $p < .001$, $\eta_p^2 = .07$, 95% $CI_{\text{difference}}$ [.76, 2.26]. For *unintended* change, athletic-related decline tipped quicker ($M = 5.52$, $SD = 1.74$) than athletic-related improvement ($M = 7.13$, $SD = 1.75$), $F(1, 209) = 23.77$, $p < .001$, $\eta_p^2 = .10$, 95% $CI_{\text{difference}}$ [.96, 2.27]; romantic-related decline ($M = 5.18$, $SD = 2.16$) than romantic-related improvement ($M = 7.06$, $SD = 2.23$), $F(1, 209) = 22.32$, $p < .001$, $\eta_p^2 = .10$, 95% $CI_{\text{difference}}$ [1.09, 2.66]; and academic-related decline ($M = 5.23$, $SD = 2.04$) than academic-related improvement ($M = 6.65$, $SD = 1.86$), $F(1, 209) = 14.40$, $p < .001$, $\eta_p^2 = .06$, 95% $CI_{\text{difference}}$ [.68, 2.16]. These results again bolster the asymmetry. Because improvement is typically intended, our past results may have been explained by an *incidental* inference that decline must be especially strong if it overrides such efforts, like a salmon swimming upstream. Going against this possibility, Study 4c reveals that people *still* tipped faster for decline regardless of whether these changes emerged because of one's efforts to make them happen or despite one's efforts against them.

However, Study 4c has an important limitation: it is unclear what "counts" in the context of intentionality given that improvement is effectively defined as fulfilling a planned action and decline as an unplanned roadblock (e.g., see Davidai & Gilovich, 2015); trying to fail and therefore failing may be construed as a

success, and so on. We hesitate to fully rule out intentionality for future research; still other designs may better tap into this dimension (e.g., designs that examine perceived intention by comparing a human actor with a robot or computer).

Overall, Study 4 may nonetheless corroborate the general robustness of the basic effect. The asymmetry replicated beyond many other possible inferences and across many kinds of changes, and so it cannot be understood merely as a function of these exogenous differences. This suggests the presence of a highly generalized negativity bias (see Baumeister et al., 2001; Rozin & Royzman, 2001). Our studies so far reveal that such an asymmetry also emerges in how people track tipping points of change.

If these various features do not account for the asymmetry, what does? Next, we directly manipulated risks and costs so to test against pure "alarm," perhaps hinting at the added contribution of entropy.

Study 5

Tipping Quickly for Decline Even When It's Costly

In Study 5, we used a fictional stepwise paradigm that afforded high control over the manipulation.

Some participants evaluated an unhealthy target changing for the better or a healthy target changing similarly for the worse, and we coded the amount of evidence they demanded before hitting a tipping point. With no other information, this invites an obvious reason to tip quickly for decline: waiting too long to diagnose a healthy target as sick may be seen as especially costly, whereas one might want to definitively confirm that an unhealthy target is better. Here we predicted to replicate the standard asymmetry, which could reasonably reflect pure "alarm."

Critically, other participants were given the *opposite* motivation: they were told that it was more costly to prematurely conclude decline versus more costly to delay concluding improvement. This flips the traditional script for risks and costs. Thus, if people's responses purely reflect "alarm"-related reasons, then these conditions should flip tipping points in kind: people should *delay* their diagnosis of decline and *hasten* their diagnosis of improvement relative to the other conditions. But if the effect also comprises an entropy component as proposed, the asymmetry may hold: if decline seems more immediately "real," then people may simply require little evidence before believing it is truly here (even when mistakenly tipping too soon threatens a high cost).

Participants. We recruited 218 participants from Amazon's Mechanical Turk (M_{age} = 33.84, SD_{age} = 11.79; 45.90% Female; 69.30% Caucasian) to complete a study about mental simulation and role-playing in exchange for $0.35.

Procedure. Participants were randomly assigned into a 2 (valence: *decline* or *improvement*) × 2 (role: *standard hedging* or *reverse hedging*) between-subjects design (all *n*s ≥ 51). First, all participants read the following introduction. Its fictional premise diverges from the reality-based contexts in previous studies, but it allows us to precisely control various parameters that may otherwise render objective thresholds incomparable:

> Imagine a community consisting of "healthy people" and "unhealthy people." They are perfect opposites in every way. For example, suppose that the healthy people score at 4/10 on health and that the unhealthy people score at 6/10. Moreover, there are equal numbers of healthy and unhealthy people here.

This phrasing holds constant various objective attributes so to compare the same actual thresholds. Participants were then assigned to conditions. First, we describe the control versions of the manipulation, which imitate prior studies and serve as a basic replication. "Standard hedging" participants continued on and read the following, across valence:

> Your task is to role-play as 'Overseer.' You pick a person at random to track. They are a healthy [unhealthy] person. From time to time you observe this person at various occasions, where you can track how they might or might show signs of change (you have zero influence on them and simply observe from afar).

They then continued to the next screen, from which point the text differed by condition. "Decline" participants read that, some time later, they checked up on this healthy person (marking "Occasion #1") and ". . . on this particular occasion, the person felt a slight wave of feeling worse that recently came and went" (no other specific details were provided). Participants then saw 1 of 2 options. The first option was: *NO CHANGE: At this point, they haven't "officially" changed in my eyes; what I've seen could be a fluke. I'd need to see more "evidence."* The second option was: *YES CHANGE: At this point, they have "officially" changed in my eyes; what I've seen doesn't feel like a fluke. I've seen enough and have hit a "tipping point."* Unbeknownst to participants, choosing the latter option ended the task; choosing the former loaded a new screen with "Occasion #2, some time later" during which this healthy person again displayed evidence of becoming unhealthy, and now at this point they again indicated their choice. This process continued until the participant chose the tipping point option, or until cycling through "Occasion #10" and the task ended automatically. Participants were never informed about how long the task could last. For sake of a parsimonious design, each encounter was described as going in the same negative way. "Improvement" participants followed identical procedures except they tracked the converse (an "unhealthy person" who had a wave of feeling better at each occasion). Our dependent variable was the number of observations that participants freely amassed before they deemed the target as officially changed. We hypothesized to replicate the same basic effect, such that these "standard hedging" participants should be quicker to diagnose change for the worse than equivalent change for the better. After all, the costs here are clear: waiting too long to diagnose an unhealthy person is often truly more alarming than diagnosing a healthy person as sick in the meantime just in case.

"Reverse hedging" participants followed the same procedures, except they were given opposing hedging information. After reading the same exact prompt as "standard" participants, they were presented with the following additional note, across valence:

> NOTE: In this world, it's extremely important to get people on medicine ONLY WHEN you think they "officially" might be getting sick [off medicine RIGHT WHEN you think they "officially" might be getting well]. Making this diagnosis too early [too late] can cause the bigger problem. As you track this person, keep in mind the first moment when you TRULY think they may be changing, if at all.

The rest of the task was identical to "standard" conditions, and we again assessed how much evidence that participants freely amassed before hitting a tipping point. Note the diverging predictions here.

"Reverse-decline" participants are informed that the stakes are more costly for misidentifying a healthy person as unhealthy *too soon*, which relates to a number of important real-world situations (e.g., needing to ensure that people start a powerful treatment or medication only if absolutely necessary). If "alarm" alone is the sole driver of the effect, this should motivate people to tip for decline less quickly than usual. Likewise, "reverse-improvement" participants are given similarly high stakes for misidentifying an unhealthy person as healthy *too late* (e.g., needing to ensure that people are not overprescribed a powerful treatment or medication). Again, if pure hedging is at work, people should tip for improvement more readily than usual. However, if decline simply seems more "real"—essentially the belief that healthy people have more universal capacities to get sick but not vice versa—then we might replicate the same asymmetry and still observe a relatively quick tipping point for decline (i.e., participants *truly think* healthy targets are becoming sick, even though forming this conclusion presents a risk).

After the task, all participants reported demographic information and responded to 2 manipulation checks, one regarding valence (*I tracked a healthy person who had waves of feeling worse*; *I tracked an unhealthy person who had waves of feeling better*) and the other regarding role-play (*Yes, I did get an extra note in the instructions; No, I did not get an extra note in the instructions*). Further ensuring that "reverse hedging" participants understood the task, they rated a third manipulation check about the specific text of the note (*Making the diagnosis to get OFF medicine TOO LATE causes the bigger problem*; *Making the diagnosis to get OFF medicine TOO EARLY causes the bigger problem*; *Making the diagnosis to get ON medicine TOO LATE causes the bigger problem*; *Making the diagnosis to get ON medicine TOO EARLY causes the bigger problem*).

Results and Discussion

Only 1.83% of participants (4 of 218) failed the manipulation check for valence; 11.47% of participants (25 of 218) failed the manipulation check for role; and 13.08% of "reverse" participants (14 of 107) failed the manipulation check for their specific task details. Eliminating them does not change any result, so they are retained in analyses.

Data were submitted to Univariate GLM analyses, with valence and role as fixed factors, and number of freely amassed observations as the dependent variable. First, we replicated the core asymmetry via a main effect of valence: collapsing across conditions, participants waited to observe significantly fewer occasions of a healthy target exhibit potential unhealthiness before perceiving official change for the worse ($M = 3.68$, $SD = 2.07$), relative to the number of occasions that an unhealthy target had to exhibit potential improvements in health to convince participants of official change for the better ($M = 5.19$, $SD = 3.00$), $F(1, 214) = 18.61$, $p < .001$, $\eta_p^2 = .08$, 95% $CI_{\text{difference}}$ [.82, 2.20]. This asymmetry is strikingly reflected further in terms of response distributions: whereas only 3.60% participants (4 of 110) in the "Decline" conditions waited to observe all 10 occasions before reaching a tipping point, a full 21.30% of participants (23 of 108) in the "Improvement" conditions did so. Overall, these findings emerged despite various objective parameters of "good" and "bad" being explicitly defined and matched within the stimuli and manipulation, providing a robust replication of the same basic effect.

We also found evidence against *pure* hedging, as shown by no main effect of role, $F(1, 214) = .009$, $p = .92$, $\eta_p^2 < .001$, and no interaction, $F(1, 214) = .09$, $p = .77$, $\eta_p^2 < .001$ (see Figure 2): As expected, "standard hedging" participants amassed less evidence before diagnosing decline versus improvement, $F(1, 214) = 10.85$, $p = .001$, $\eta_p^2 = .05$, 95% $CI_{\text{difference}}$ [.65, 2.58]. More surprisingly, this asymmetry was identical for "reverse hedging" participants: although these participants were incentivized to delay for decline and hasten for improvement, they still tipped significantly more quickly for decline than for improvement, $F(1, 214) = 7.91$, $p = .005$, $\eta_p^2 = .04$, 95% $CI_{\text{difference}}$ [.42, 2.39], and in fact tipped just as readily as their "standard" counterparts, Fs $< .08$, ps $> .78$, ηs$_p^2 < .001$.[5]

Finally, these effects were again emphasized by response distributions: just 3.70% of "standard hedging" participants (2 of 54) waited for all the evidence before they tipped for decline, but 19.30% did (11 of 57) before tipping for improvement; likewise, almost no "reverse hedging" participants (3.60%, 2 of 56) amassed all the evidence before they tipped for decline, but far more did (19.30%, 11 of 57) before tipping for improvement.

These results offer initial evidence that our basic effect is not driven by "alarm" alone, a candidate mechanism by which valence asymmetries are typically understood. When incentivized to really tell the truth in their tipping points (i.e., knowing that tipping too early would be mistaken and costly), participants tipped at the same rate as they normally do. This null difference at least hints at an entropy component. Quickly tipping here seems to at least partly reflect a *genuine* belief that the entity is getting worse.

To emphasize, "alarm" reasons surely contribute to the asymmetry at large, and entropy should not necessarily win out in all contexts for all people. Also, perhaps our manipulation did not fully override other intuitions about costs (e.g., "reverse hedging" participants ultimately may have worried more about signs of sickness than side effects of treatment). Study 6 was thus designed to actively support our entropy hypothesis. We directly pitted the many potential features of negativity against the proposed effects of plausibility. If the asymmetry indeed has an entropy component, it may be attenuated if the opposite trajectory is evoked—if ceilings seem universal and floors seem selective.

Study 6

Framing Improvement as Universally Accessible

In Study 6, we measured people's tipping points for a task that was framed as hard to scale upward on (which may replicate the

[5] Incidentally, as is often the case for data involving free response timing (Heathcote, Popiel, & Mewhort, 1991), the distribution of tipping points was positively skewed. The observed asymmetric effect remained highly significant after applying the recommended logarithmic transformation: no main effect of role, $F(1, 214) = .21$, $p = .65$, $\eta_p^2 = .001$, no interaction, $F(1, 214) = .29$, $p = .59$, $\eta_p^2 = .001$, and only the main effect of valence: $M_{\text{decline}} = .50$, $SD_{\text{decline}} = .24$; $M_{\text{improve}} = .64$, $SD_{\text{improve}} = .27$, $F(1, 214) = 14.34$, $p < .001$, $\eta_p^2 = .05$, 95% $CI_{\text{difference}}$ [.06, .20]. Pairwise comparisons also remained significant, Fs > 5.17, ps $< .024$, ηs$_p^2 > .02$. We report untransformed results in the main text for ease of conceptual interpretation.

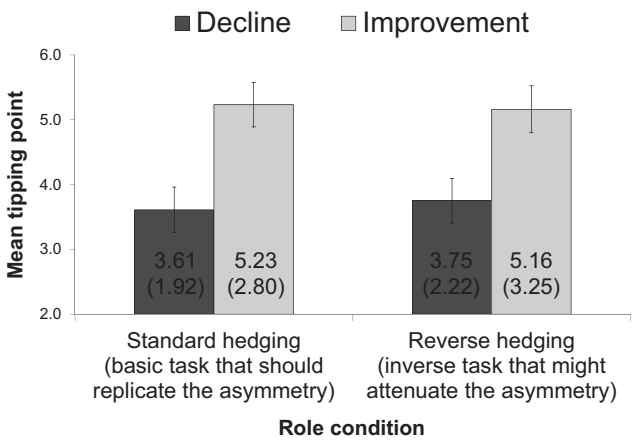


Figure 2. Study 5: Amount of evidence that participants freely amassed before hitting a tipping point in viewing the target as officially changed. Means and standard deviations (in parentheses) are presented, with error bars signifying ±1 standard error. "Reverse hedging" participants were instructed to cautiously tip for decline (and to quickly tip for improvement) to protect against costs, but they nonetheless tipped just as readily (and just as slowly) as they normally would have.

asymmetry, because improvement should seem selective and hard to make happen) or easy to scale upward on (which may reduce or even reverse the asymmetry, since the same logic of plausibility implies initial *bad* signs are flukish). But if people are purely risk averse, or otherwise react to negative information (by virtue of being negative) in extreme ways, people may still tip more readily in the face of bad signs even when knowing most cases will ultimately improve.

Participants. We recruited 339 participants from Amazon's Mechanical Turk (M_{age} = 33.70, SD_{age} = 11.14; 42.20% Female; 73.50% Caucasian) to complete a study about mental simulation and role-playing in exchange for $0.50.

Procedure. Participants were randomly assigned into a 2 (valence: *decline* or *improvement*) × 3 (framing: *selective ceilings*, *universal ceilings*, or *control*) between-subjects design (all *n*s ≥ 53). We used the same piecemeal evidence-gathering design as in Study 5. Participants read about a fictional community that plays the game "X-Ball," and were randomly assigned to 1 of 3 framings of X-Ball improvement. We first describe the "selective ceilings" versions of the manipulation, which served as a basic replication:

> For purposes of this study, you (the M-Turker reading this right now) do not need to know specifics about the game. What's important for you to know is that the game taps into certain tendencies and capacities that few people in the community naturally possess, whether they realize it now or not. X-Ball takes a lot to excel at; people have to do more than just show up and play to ultimately become one of the good players. Regardless of how things pan out over the course of learning, statistically it's true that few people here will turn out to be good X-Ball players.

Then, participants were tasked with role-playing as an impartial "overseer" identical to Study 5, in which they tracked a random person's performance at various occasions after the person had decided to try out X-Ball from scratch. "Decline" participants read that, some time later, they checked up on this person ("Occasion #1") and ". . . on this particular occasion, the person happened to be playing poorly." They saw 1 of 2 options. The first option was: *NO CHANGE: In my eyes, this may still be a fluke. I'll wait in my judgment; at this point, I feel unsure if this person is "officially" one of the bad X-Ball players in the making.* The second option was: *YES CHANGE: In my eyes, I've seen enough. I've hit a "tipping point" in my judgment; at this point, I feel sure that this person is "officially" one of the bad X-Ball players in the making.* As in Study 5, clicking the former option loaded a similar description of "Occasion #2," and so on, until participants tipped or the process cycled through "Occasion #10." "Improvement" participants followed identical, converse procedures (tracking "one of the good X-Ball players in the making" as a person "played well" at each occasion). The dependent variable was the number of observations that participants freely amassed before tipping. Here we hypothesized to replicate the basic effect: these "selective ceilings" participants should tip quicker in the face of bad signs versus good signs, presumably because more evidence is needed to convince them that the person in question is among the select few who will indeed change for the better.

Critically, other participants assigned to "universal ceilings" versions of the manipulation completed the study in the same exact way, but with one exception. Their opening description of X-Ball included the *opposite* chances of improvement. They read:

> What's important for you to know is that the game taps into certain tendencies and capacities that most people in the community naturally possess, whether they realize it now or not. X-Ball takes almost nothing to excel at; people just have to show up and play to ultimately become one of the good players. Regardless of how things pan out over the course of learning, statistically it's true that most people here will turn out to be good X-Ball players.

Note the competing predictions: If bad signs stand out by virtue of being negative, these "universal ceilings" participants should tip just readily as normal when faced with initial badness ("better safe than sorry"). However, if selectivity also informs tipping points, the asymmetry here may be attenuated or even flip; despite being negative, these bad signs may also seem relatively implausible and thus should demand more evidence, whereas initial *good* signs should now seem more diagnostic of "true" change.

Finally, we also added a control condition in this study, in which the task was framed as having an equal likelihood of scaling upward versus staying at the floor. These participants followed identical procedures, except their opening details were as follows:

> What's important for you to know is that the game taps into certain tendencies and capacities that 50% of people in the community naturally possess and 50% do not naturally possess, whether they realize it now or not. As people show up and play, it's unclear whether they will ultimately become one of the good players. Regardless of how things pan out over the course of learning, statistically it's true that half of the people here will turn out to be good X-Ball players.

This allows us to assess whether the basic effect reemerges when conditions of change are more ambiguous and improvement and decline are better matched, as in our previous studies. That is, control participants may resemble "selective ceilings" participants, even though the objective information they possess should compel them to tip at equal rates.

After the task, all participants reported demographic information and responded to 2 manipulation checks, one regarding valence

(*I tracked a person who kept playing well at X-Ball*; *I tracked a person who kept playing poorly at X-Ball*) and the other regarding framing (*I was told that, statistically, most people turn out to be good X-Ball players; I was told that, statistically, few people turn out to be bad X-Ball players; I was told that, statistically, exactly 50% of people turn out to be good X-Ball players*).

Results and Discussion

Only 5.90% of participants (20 of 339) failed the manipulation check for valence, and only 5.00% of participants (17 of 339) failed the manipulation check for framing. Eliminating them does not change any result, so they are retained in analyses.

Data were submitted to Univariate GLM analyses, with valence and framing as fixed factors, and number of freely amassed observations as the dependent variable. In line with our hypothesis, there was no main effect of valence, $F(1, 333) = .07$, $p = .80$, $\eta_p^2 < .001$, a main effect of framing, $F(1, 333) = 8.15$, $p < .001$, $\eta_p^2 = .05$, and most critically, a significant interaction, $F(1, 333) = 19.59$, $p < .001$, $\eta_p^2 = .11$ (see Figure 3).

Parsing apart this interaction, we first replicated the same core asymmetry within "selective ceilings" conditions: these participants indeed waited to observe significantly more instances of good performance before concluding that the target seemed officially on track for substantive improvement, relative to the number of bad performances before concluding otherwise, $F(1, 333) = 7.44$, $p = .007$, $\eta_p^2 = .02$, 95% $CI_{\text{difference}}$ [.33, 2.01]. Ceilings were explicitly described as selective, and in turn participants demanded more evidence before diagnosing lasting change for the better—as in all our previous studies. More important for purposes of the current study, this robust basic effect *reversed* within "universal ceilings" conditions: when participants believed that most targets can and will improve rather than decline, they actually waited to observe relatively *few* instances of good performance and relatively *many* instances of bad performance before tipping in their perceptions of official change, $F(1, 333) = 27.35$, $p < .001$, $\eta_p^2 = .08$, 95% $CI_{\text{difference}}$ [1.44, 3.17]. This finding suggests that (at least in this context) people indeed draw upon broader beliefs about the plausibility of improvement and decline to evaluate tipping points of change, despite the fact that the evidence is still negative. Conditions that flip the traditional direction of "entropy" flip the asymmetric tipping point in kind.

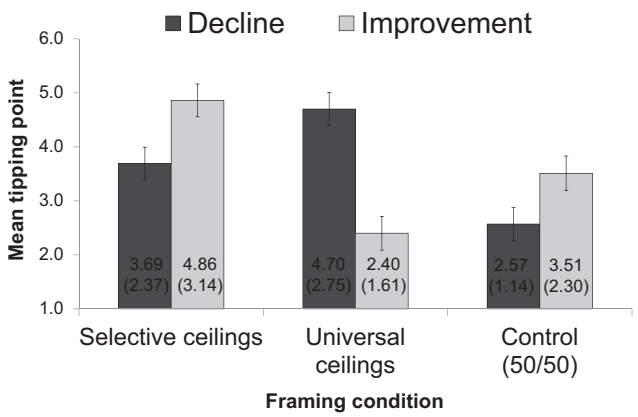


Figure 3. Study 6: Amount of evidence that participants freely amassed before hitting a tipping point in viewing the target as officially changed. Means and standard deviations (in parentheses) are presented, with error bars signifying ±1 standard error. Of note, the asymmetry observed throughout studies flipped among "universal ceilings" participants, who relatively slowly tipped for decline and relatively quickly tipped for improvement.

Finally, recall that control participants were told that the likelihood of a person's X-Ball improvement was at equal chance (rather than explicitly "select" or "universal"). Instead of then demanding equal evidence, these participants waited for significantly more positive signs before diagnosing lasting change for the better, compared with their quick tipping point in the face of initial negative signs, $F(1, 333) = 4.42$, $p = .036$, $\eta_p^2 = .01$, 95% $CI_{\text{difference}}$ [.06, 1.82].[6] The basic effect reemerged. Hence, as in throughout all of our studies, people's "default" process of diagnosing tipping points may be to tip more readily for decline unless given clear reason otherwise. At the end of the current paper, we return to how this kind of *overgeneralization* may pose particular problems.

General Discussion

Is one grain of sand a heap of sand?

The answer to this simple question is clearly no.

If a second grain is added to the first, is there a heap?

Again the answer is no.

If a third grain is added, is there a heap?

For a third time the answer is no.

— The Sorites Paradox (Paradox of the Heap), in Fisher (2000)

Observations like the one above can be traced as far back as the Ancient Greeks (circa 400 BCE) and were recently named as representing one of the most bewildering and profound problems in contemporary philosophy (see Sainsbury, 2009). A collection of individual grains will eventually become a heap, but the objective point at which they do is effectively ambiguous. Here we have explored a related theme, more broadly in terms of the subjective perception of when change first seems to emerge: the point at which people feel they have seen sufficient evidence that an entity has begun deviating from some baseline in a meaningful, lasting, "official" way. Many important experiences in everyday life unfold in such ambiguous fashion. Smaller fluctuations in health, wealth, happiness, character, abilities, and interpersonal states at some point seem to "add up"—our judgments of them tip—which can guide many consequential decisions, feelings, and behaviors in response. What quantity of smaller fluctuations influences this tipping point? How many grains of sand must compound to elicit our diagnosis of official change?

Ten studies reveal that the answer depends on the kind of heap being formed. People are quicker to diagnose change for the worse (decline) than change for the better (improvement), given similar reason to tip. This asymmetry proved strikingly pervasive.

[6] Per Note 5, all effects remained highly significant after applying the recommended logarithmic transformation: no main effect of valence, $F(1, 333) = 1.21$, $p = .27$, $\eta_p^2 = .004$, a main effect of framing, $F(1, 333) = 6.38$, $p = .002$, $\eta_p^2 = .04$, and an interaction, $F(1, 333) = 19.83$, $p < .001$, $\eta_p^2 = .11$. All pairwise comparisons also remained significant, $Fs > 2.91$, $ps < .089$, $\eta s_p^2 > .09$.

First, we established this basic effect across many methods, measures, and contexts, from judging changes in mood, academics, athletic performance, personality, and a diversity of life domains (Studies 1a, 1b, and 1c); to interpreting the same societal trends merely framed as decline or improvement (Study 2); to real behavior—people were more likely to bet on bad luck versus good luck, despite identical chances. Next, we sought to explain this asymmetry. It replicated across various exogenous features of our stimuli (Studies 4a, 4b, and 4c), suggesting a generalized negativity bias. Teasing this apart, we tested whether this reflects people's high sensitivity to costs ("alarm" effects may lead people to perceive early signs of decline in more extreme ways), or that early signs of decline seem more immediately credible in sorting fact from fluke (an "entropy" effect). We found some direct evidence for the novel role of entropy beyond pure alarm (pilot study, Studies 5–6).

More research is needed to document this entropy construct, to test when and how it affects change perception, and to further disentangle entropy from other contributors. Throughout we have considered the extent to which each of our studies may or may not have reflected an entropy component (review on pp. 28–29; Studies 4–6), and to be clear, multiple drivers likely contribute to some degree depending on the study or task at hand. In all, we find robust evidence for a general tipping point asymmetry, which does appear to at least partly reflect an "entropy" effect beyond reasons like risk aversion alone.

Study 6 makes the strongest case for this latter claim. Prior theories of valence asymmetries do not account for these results (i.e., regardless of the many potential features, effects, and "alarms" of negative information, people no longer reacted to early bad signs when ultimate decline seemed uncommon). This reveals the need for a better understanding of how assumptions of plausibility may also shape change perception, in novel ways. Therefore, throughout the General Discussion, we consider the value of the basic effect while highlighting unique implications raised by an entropy component.

Theoretical Implications

First, the construct of a tipping point significantly complements and extends existing literatures on change. As outlined, past studies shed little light on the underlying *evidence-gathering* dynamics guiding our initial diagnosis that something has changed. This process often unfolds freely, gradually, and across extended fluctuations; a person might observe smaller ambiguous changes in a relationship over time and must consider *when* these changes stop reflecting a fluke and start reflecting a stable indicator of how things are. We sought to capture this everyday stepwise process, illuminating how and why people infer "official" change and thus the point at which people may actually act.

Further, the clear asymmetry in this tipping point process provides a compelling case against weighted averaging, which predicts that valence should not have made such a critical difference in perceiving things like streaks (Anderson, 1981). To our knowledge these studies count among the most diverse and general cases of negativity bias to date, comprising many subjects and contexts far beyond hedging or morality as in prior work on the topic (Baumeister et al., 2001; Reeder & Brewer, 1979; Rozin & Royzman, 2001; Skowronski & Carlston, 1987). That such a tendency extends to tipping points may be seen as an important "prequel": we reveal negativity bias in the *initial emergence* of judgment (the first point when a good target seems changed for the worse), while past work might be seen as then highlighting the uphill battle of such conclusions *once formed* (how a definitively categorized bad target then gets treated). Together these findings help paint a fuller portrait of negativity bias across the time-course of impression formation.

At the same time, one cannot help but bring to mind the equally robust literature on *positivity biases* like unrealistic optimism and axiomatic notions of self-enhancement and motivated reasoning (Kunda, 1990; Markus & Ruvolo, 1989; Schacter & Addis, 2007; Sedikides & Hepper, 2009; Sharot, 2012; Taylor & Brown, 1988; Weinstein, 1980). This literature seems to make the exact opposite predictions, yet we found little evidence that people readily claim signs of improvement as "real" (even when evaluating themselves). This presents an opportunity for fruitful theoretical intersection. How can people be so sensitive to decline while also blind to it?

We consider two general answers. First, our entropy-specific findings could be key. Recall that the only study in which people did readily claim positive signs as "real" was Study 6, when we framed ultimate improvement as likely, common, and plausible. This suggests people who hold this view of improvement by other means should also readily embrace positive signs and be skeptical of signs of decline, confirming what one would predict from the optimism literature. Our studies simply may not capture the conditions that reveal these beliefs (e.g., being passionately engaged in goal pursuit; having strong personal incentives to see something come to fruition; having strong agency to make or stop change; focusing on other timescales). More research should map out the individual and situational demands that alter "entropy" assumptions (and thereby tipping points).

Second and more broadly, a tipping point framework shines light on ways to potentially integrate the literatures on negativity bias and positivity bias rather than attributing them to some individual difference. That is, perhaps most people do feel that good things can readily lose their positive qualities and that decline can readily emerge, *while also* expressing the many established variants of optimism. One critical feature of our studies is that each piece of evidence for possible change actually occurred (e.g., back to Study 1a: a good athlete *did* play a bad game and a bad athlete *did* play a good game). Thus, perhaps negativity bias dominates once the evidence starts to pile up; bad signs pile up more quickly. But in studies on positivity bias, participants almost always start at a zero-point with no evidence yet emerging (e.g., in Weinstein's (1980) classic research on unrealistic optimism: perfectly healthy students judge their likelihood of a future heart attack); see Snyder, Sympson, Michael, and Cheavens (2001) for a review. Thus, perhaps positivity biases dominate before any objective counterevidence accrues. People could be well aware of the universality of decline (as in our entropy account) but be optimistic by "default," until concrete signs of this likely fate begin to string together. This would be consistent with traditional theorizing about motivated reasoning, which posits that people are self-serving only to the extent there is high ambiguity that they can distort in their favor (see Kunda, 1990). The same person may therefore dynamically express both tendencies (e.g., before the season starts, "My team will win it all!"; after a quick string of losses, "We're officially done!"). In any case, this is one of many possibilities for integration; our general tipping point framework invites rich ways for understanding the longer-term dynamics of change perception. More complete models

may be required for integrating when and why a person will overweight positive versus negative information.

Practical Implications

People care about change. The folk-tale of villain-turned-saint is an attractive one, featured in well-known stories like Dickens' (1834) *A Christmas Carol* and movies like *Groundhog Day* (Albert & Ramis, 1993). Acclaimed TV shows like *The Sopranos, Breaking Bad*, and *Mad Men* all hinge on the question of how many acts a person must commit before they gain or lose their humanity (Harris, 2014). Headlines speculate when players have bounced back and celebrities have washed up. Politicians debate crashes and recoveries. Aging relatives lament generational shifts. And people more broadly monitor their own everyday fluctuations. Such observations often rely on subjective perception alone. Hence, our documented asymmetry also raises a variety of practical implications.

To the extent a tipping point indeed reflects when people become more liable to actually act, our findings suggest inequitable behavior in response to improvement and decline. People might be overly quick to punish others but overly slow to reward others who are trying to change, and likewise too slow to appreciate their own progress but too quick to feel hopeless from equivalent regression—and even more worrisome, an entropy component suggests at least some people in some contexts will construe these changes as objective (potentially unfixable) truths, even following changes of purely random chance akin to Study 3. In general, people should bear in mind how suddenly an "official" cold streak can seem to form and adjust accordingly. In treating others, for example, teachers could purposefully add an easy assignment if students receive 2 or 3 low grades in a row (helping counter an immediate sense of lasting descent). In managing oneself, people are wise to consider the fickleness of a positive reputation and reinforce their standing with ongoing affirmation; in our studies, being a happy person or having a good reputation at the start counted for less than one would hope, in the face of just a few opposing signals.

These findings also invite a novel perspective for rethinking policy debates over broader societal changes, such as climatic and economic cycles. Traditionally, "effective political persuasion" is thought to involve convincing another side that a problem indeed exists and should be addressed, even if one's precise method of convincing (e.g., central or peripheral routes) might vary (Blumler & Kavanagh, 1999; Campbell, 2002; Hoffman, 2011; Risse, 2000). Our findings suggest a more crucial piece: as hard as it is to convince others that one's chosen societal issue is a problem (a negative change demanding little evidence), it may be even harder to convince others that policy *solutions* are working (a positive change demanding a lot of evidence). Study 2 directly supports this claim: the same data may be less impactful simply when depicting improvement. Policymakers might put more effort into this back-end of the persuasion process and emphasize the gradual nature of improvement from a proposed intervention beyond the initial existence of decline. Again, this becomes all the more troubling in light of an entropy effect, in that people may *truly believe* the declines and *truly doubt* the improvements that are reported.

Future Directions

The current studies pave 3 valuable avenues for empirical work, in addition to the various theoretical directions that we raised earlier. First, the basic construct of a tipping point can be extended in many interesting directions. What factors beyond "how much" evidence accelerate or suspend our diagnosis that things have officially changed? Some likely candidates are extremity (e.g., one large event may wield the same power as a string of small events) and the willingness to believe in change (e.g., *any* evidence of climate change may fail to sway those who reject the premise on ideological grounds), but many interesting possibilities remain (e.g., how current emotions alter perceptions of the evidence: Niedenthal, Brauer, Halberstadt, & Innes-Ker, 2001). Moreover, given that our entropy account cites plausibility as key, other factors that boost plausibility should accelerate tipping points, like hearing about a change from a credible source (Cialdini & Goldstein, 2004) or reducing the number of alternative sources to which one can attribute the change (Schwarz, 2012). Designs with more varied evidence may also prove valuable (e.g., "noisier" fluctuations than strings of all equally positive or all equally negative evidence). Last, research could extend tipping points to broader categorization processes, exploring tipping points not only across the same spectrum (e.g., when a good player seems to become a bad player) but across discrete identities (e.g., when someone seems like a different "person" altogether: Massey & Wu, 2005; Strohminger & Nichols, 2014).

Second, the parameters of the basic asymmetry should be investigated further, especially in terms of an entropy component. Study 6 suggests that people who *naturally* view improvement in a domain as "universally accessible" should also show the opposite tipping points; cultural differences in construing the connectedness of events (Masuda & Nisbett, 2006) and individual differences in optimism or pessimism (Scheier, Carver, & Bridges, 1994) may prove relevant. Other research should further rule out incidental effects that render the evidence objectively different to begin with, beyond the difference of interest; we sought to address this issue as best as we could conceive, but it remains a notorious problem for interpreting valence asymmetries (see Rozin & Royzman, 2001).

On the other hand, it seems useful to keep *expanding* the asymmetry by exploring real-world contexts outside the lab. For example, the effect suggests novel ways to understand big data. Just a single GDP down-year may shift spending and financial satisfaction, but many GDP up-years may be needed for converse changes; attitudes might track closer with down-years (e.g., the worse a GDP down-year, the bigger the negative change in public opinion) versus up-years (e.g., better GDP up-years may not correlate in kind), if spikes in decline seem lasting but similar spikes in improvement seem flukish; and so on.

Last, a robust asymmetry in diagnosing decline versus improvement raises the question of whether this bias can be considered "rational" or "irrational." Rationality has a long history in judgment research, with many definitions for what constitutes a rational judgment (see Hastie & Dawes, 2001). A common metric is whether a bias leads to more accurate assessments across most situations (Gigerenzer et al., 1999; Hastie, 2001; Payne, Bettman, & Johnson, 1993). Does more readily tipping for decline than for improvement make people wiser or better off in the long run? On the one hand, both "alarm" and entropy are largely rooted in this vein: in many situations, it may be right to worry about the costs of possible decline more than celebrate possible improvement, and the entity may indeed be more on track to decline than improve after a few early signs. On the other hand, the sheer pervasiveness of the asymmetry raises concern. First, because it emerged across a striking variety of situations, it likely reflects an

overgeneralized heuristic—valid in many cases, but unwittingly applied to others even when it poorly approximates reality (Baron, 1990). Readily detecting declines in health seems rational, but cutting off nascent relationships because of a small slight might result in unnecessarily losing friends and social support (which can require time to cultivate); and although the actual difficulty of sustaining good athletic performance may justify a quick tip for decline, good luck is no harder to sustain than bad luck. An entropy explanation may be especially problematic, given that people may attribute such fluctuations to objective reality. Study 2 (graphs) and Study 3 (bets) best highlight this kind of overgeneralization. Second, even if the asymmetry is a mostly accurate judgment, it may come with undesirable consequences. Appreciating the reality of decline at all times and for all cases may foster a bad mood or sense of cynicism, leading people to feel unmotivated to help despite a ready diagnosis; early detection might too-often lead people to "write off" the entity rather than intervene.

Concluding Thoughts

Change is an inevitable fact of life, but the point at which things actually do is often more a work of fiction. Smaller fluctuations in our everyday experiences create ambiguity about when they reflect a substantive shift in quality versus simply a passing trend, relegating this *tipping point* to the whims of subjective perception.

The current paper was an attempt to shed light on this process. We specifically focused on the question of "how much" evidence is needed before people first come to diagnose official change. Holding constant a wide variety of features and generalizing across a host of contexts, results converge to reveal a robust principle: people tip for the worse much more readily than they tip for the better. Doing so might frequently be to our advantage, but not always and not without consequence. Regardless, the ubiquity of this tendency not only affords fruitful offshoots for research, but also a fuller understanding of how people might (inequitably) navigate their ever-changing lot.

References

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. http://dx.doi.org/10.1037/0096-1523.12.4.445

Albert, S. (1977). Temporal comparison theory. *Psychological Review, 84,* 485–503. http://dx.doi.org/10.1037/0033-295X.84.6.485

Albert, T., & Ramis, H. (1993). *Groundhog Day* [motion picture]. Culver City, CA: Columbia Pictures.

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55,* 387–395. http://dx.doi.org/10.1037/0022-3514.55.3.387

Ambady, N., Bernieri, F., & Richeson, J. (2000). Towards a histology of social behavior: Judgmental accuracy from thin slices of behavior. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 201–271). New York, NY: Academic Press. http://dx.doi.org/10.1016/S0065-2601(00)80006-4

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

Baron, J. (1990). Harmful heuristics and the improvement of thinking. In D. Kuhn (Ed.), *Developmental perspectives on teaching and learning thinking skills* (pp. 28–47). Basel, Switzerland: Karger. http://dx.doi.org/10.1159/000418979

Baron, M. E. (1969). *The origins of infinitesimal calculus*. London, UK: Pergamon Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323–370. http://dx.doi.org/10.1037/1089-2680.5.4.323

Beck, D. M., Rees, G., Frith, C. D., & Lavie, N. (2001). Neural correlates of change detection and change blindness. *Nature Neuroscience, 4,* 645–650. http://dx.doi.org/10.1038/88477

Blumler, J. G., & Kavanagh, D. (1999). The third age of political communication: Influences and features. *Political Communication, 16,* 209–230. http://dx.doi.org/10.1080/105846099198596

Burgers, J. M. (1963). On the emergence of pattern of order. *Bulletin of American Mathematics, 69,* 1–26. http://dx.doi.org/10.1090/S0002-9904-1963-10832-0

Campbell, J. L. (2002). Ideas, politics, and public policy. *Annual Review of Sociology, 28,* 21–38. http://dx.doi.org/10.1146/annurev.soc.28.110601.141111

Campbell, T., O'Brien, E., Van Boven, L., Schwarz, N., & Ubel, P. A. (2014). Too much experience: A desensitization bias in emotional perspective taking. *Journal of Personality and Social Psychology, 106,* 272–285.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social and cognitive psychology* (pp. 73–96). New York, NY: Guilford Press.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55,* 591–621. http://dx.doi.org/10.1146/annurev.psych.55.090902.142015

Cone, J., & Ferguson, M. J. (2015). He did *what*? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108,* 37–57. http://dx.doi.org/10.1037/pspa0000014

Conway, M., & Ross, M. (1984). Getting what you want by revising what you had. *Journal of Personality and Social Psychology, 47,* 738–748. http://dx.doi.org/10.1037/0022-3514.47.4.738

Davidai, S., & Gilovich, T. (2015). What goes up apparently needn't come down: Asymmetric predictions of ascent and descent in rankings. *Journal of Behavioral Decision Making, 28,* 491–503. http://dx.doi.org/10.1002/bdm.1865

Demany, L., Trost, W., Serman, M., & Semal, C. (2008). Auditory change detection: Simple sounds are not memorized better than complex sounds. *Psychological Science, 19,* 85–91. http://dx.doi.org/10.1111/j.1467-9280.2008.02050.x

Devine, P. G., Hirt, E. R., & Gehrke, E. M. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality and Social Psychology, 58,* 952–963. http://dx.doi.org/10.1037/0022-3514.58.6.952

Dickens, C. (1834). *A Christmas carol*. London, England: Chapman & Hall.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia, PA: Taylor and Francis/Psychology Press.

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. http://dx.doi.org/10.1037/0022-3514.84.5.917

Falk, R., & Konold, C. E. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104,* 301–318. http://dx.doi.org/10.1037/0033-295X.104.2.301

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Fischhoff, B., & Beyth, R. (1975). "I knew it would happen": Remembered probabilities of once-future things. *Organizational Behavior & Human Performance, 13,* 1–16. http://dx.doi.org/10.1016/0030-5073(75)90002-1

Fisher, P. (2000). Sorites paradox and vague geographies. *Fuzzy Sets and Systems, 113,* 7–18. http://dx.doi.org/10.1016/S0165-0114(99)00009-3

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. http://dx.doi.org/10.1037/0022-3514.38.6.889

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology, 15,* 183–217. http://dx.doi.org/10.1080/10463280440000026

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. http://dx.doi.org/10.1037/0033-2909.132.5.692

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Gilbert, D. T., & Wilson, T. D. (2000). Miswanting: Some problems in the forecasting of future affective states. In J. Forgas (Ed.), *Feeling and thinking: The role of affect in social cognition* (pp. 178–197). New York, NY: Cambridge University Press.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology, 38,* 129–166. http://dx.doi.org/10.1006/cogp.1998.0710

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6,* 517–523. http://dx.doi.org/10.1016/S1364-6613(02)02011-9

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90,* 1–20. http://dx.doi.org/10.1037/0022-3514.90.1.1

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435. http://dx.doi.org/10.1016/0010-0285(92)90013-R

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science, 16,* 152–160. http://dx.doi.org/10.1111/j.0956-7976.2005.00796.x

Grimes, J. (1996). On the failure to detect changes in scenes across saccades. In K. Akins (Ed.), *Vancouver studies in cognitive science, Vol. 2: Perception* (pp. 89–110). New York, NY: Oxford University Press.

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

Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology, 52,* 653–683. http://dx.doi.org/10.1146/annurev.psych.52.1.653

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA: Sage.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin, 109,* 340–347. http://dx.doi.org/10.1037/0033-2909.109.2.340

Heraclitus. (translated by D. W. Graham, 2015). *From The Stanford encyclopedia of philosophy*. Retrieved from http://plato.stanford.edu/archives/fall2015/entries/heraclitus/

Hoffman, A. J. (2011). Talking past each other? Cultural framing of skeptical and convinced logics in the climate change debate. *Organization & Environment, 24,* 3–33. http://dx.doi.org/10.1177/1086026611404336

Ji, L. J., Nisbett, R. E., & Su, Y. (2001). Culture, change, and prediction. *Psychological Science, 12,* 450–456. http://dx.doi.org/10.1111/1467-9280.00384

Kahneman, D., & Snell, J. (1992). Predicting a changing taste: Do people know what they will like? *Journal of Behavioral Decision Making, 5,* 187–200. http://dx.doi.org/10.1002/bdm.3960050304

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes, 99,* 49–65. http://dx.doi.org/10.1016/j.obhdp.2005.07.002

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,* 149–166. http://dx.doi.org/10.1521/soco.2016.34.2.149

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480

Levine, M., & Shefner, J. (1981). *Fundamentals of sensation and perception*. Reading, MA: Addison Wesley.

Lieb, E. H., & Yngvason, J. (1999). The physics and mathematics of the second law of thermodynamics. *Physics Reports, 310,* 1–96. http://dx.doi.org/10.1016/S0370-1573(98)00082-9

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127,* 267–286. http://dx.doi.org/10.1037/0033-2909.127.2.267

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,* 823–849. http://dx.doi.org/10.1037/pspa0000021

Markus, H., & Ruvolo, A. (1989). Possible selves: Personalized representations of goals. In L. A. Pervin (Ed.), *Goal concepts in personality and social psychology* (pp. 211–241). Hillsdale, NJ: Erlbaum.

Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under- and overreaction. *Management Science, 51,* 932–947. http://dx.doi.org/10.1287/mnsc.1050.0386

Masuda, T., & Nisbett, R. E. (2006). Culture and change blindness. *Cognitive Science, 30,* 381–399. http://dx.doi.org/10.1207/s15516709cog0000_63

Nesse, R. (2005). Natural selection and the regulation of defenses: A signal detection analysis of the smoke detector principle. *Evolution and Human Behavior, 26,* 88–105. http://dx.doi.org/10.1016/j.evolhumbehav.2004.08.002

Niedenthal, P. M., Brauer, M., Halberstadt, J. B., & Innes-Ker, A. H. (2001). When did her smile drop? Facial mimicry and the influences of emotional state on the detection of change in emotional expression. *Cognition and Emotion, 15,* 853–864. http://dx.doi.org/10.1080/02699930143000194

O'Brien, E. (2013). Easy to retrieve but hard to believe: Metacognitive discounting of the unpleasantly possible. *Psychological Science, 24,* 844–851. http://dx.doi.org/10.1177/0956797612461359

O'Brien, E. (2015a). Feeling connected to younger versus older selves: The asymmetric impact of life stage orientation. *Cognition and Emotion, 29,* 678–686.

O'Brien, E. (2015b). Mapping out past and future minds: The perceived trajectory of rationality versus emotionality over time. *Journal of Experimental Psychology: General, 144,* 624–638. http://dx.doi.org/10.1037/xge0000064

O'Brien, E., & Kardas, M. (2016). The implicit meaning of (my) change. *Journal of Personality and Social Psychology, 111,* 882–894. http://dx.doi.org/10.1037/pspi0000073

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44,* 369–378. http://dx.doi.org/10.3758/BF03210419

Payne, J. W., Bettman, J. R., & Johnson, E. L. (1993). *The adaptive decision maker*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139173933

Quoidbach, J., Gilbert, D. T., & Wilson, T. D. (2013). The end of history illusion. *Science, 339,* 96–98. http://dx.doi.org/10.1126/science.1229294

Reeder, G. D. (1993). Trait-behavior relations and dispositional inference. *Personality and Social Psychology Bulletin, 19,* 586–593. http://dx.doi.org/10.1177/0146167293195010

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86,* 61–79. http://dx.doi.org/10.1037/0033-295X.86.1.61

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition, 4,* 1–17. http://dx.doi.org/10.1521/soco.1986.4.1.1

Reeder, G. D., & Fulks, J. L. (1980). When actions speak louder than words: Implicational schemata and the attribution of ability. *Journal of Experimental Social Psychology, 16,* 33–46. http://dx.doi.org/10.1016/0022-1031(80)90034-7

Reeder, G. D., Pryor, J. B., & Wojciszke, B. (1992). Trait-behavior relations in social information processing. In G. R. Semin & K. Fiedler (Eds.), *Language, interaction and social cognition* (pp. 37–57). Thousand Oaks, CA: Sage.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53,* 245–277. http://dx.doi.org/10.1146/annurev.psych.53.100901.135125

Risse, T. (2000). "Let's argue!": Communicative action in world politics. *International Organization, 54,* 1–39. http://dx.doi.org/10.1162/002081800551109

Robinson, M. D., & Clore, G. L. (2002). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology, 83,* 198–215. http://dx.doi.org/10.1037/0022-3514.83.1.198

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96,* 341–357. http://dx.doi.org/10.1037/0033-295X.96.2.341

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5,* 296–320. http://dx.doi.org/10.1207/S15327957PSPR0504_2

Sainsbury, R. M. (2009). *Paradoxes*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511812576

Schacter, D. L., & Addis, D. R. (2007). The optimistic brain. *Nature Neuroscience, 10,* 1345–1347. http://dx.doi.org/10.1038/nn1107-1345

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67,* 1063–1078. http://dx.doi.org/10.1037/0022-3514.67.6.1063

Schwarz, N. (2012). Feelings-as-information theory. In P. A. M. Van Lange, A. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 289–308). Thousand Oaks, CA: Sage.

Sedikides, C., & Hepper, E. G. D. (2009). Self-improvement. *Social and Personality Psychology Compass, 3,* 899–917. http://dx.doi.org/10.1111/j.1751-9004.2009.00231.x

Sharot, T. (2012). *The optimism bias: A tour of the irrationally positive brain*. New York, NY: Random House.

Simons, D. J., & Ambinder, M. S. (2005). Change blindness: Theory and consequences. *Current Directions in Psychological Science, 14,* 44–48. http://dx.doi.org/10.1111/j.0963-7214.2005.00332.x

Simons, D. J., Franconeri, S. L., & Reimer, R. L. (2000). Change blindness in the absence of a visual disruption. *Perception, 29,* 1143–1154. http://dx.doi.org/10.1068/p3104

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. http://dx.doi.org/10.1037/0022-3514.52.4.689

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105,* 131–142. http://dx.doi.org/10.1037/0033-2909.105.1.131

Snyder, C. R., Sympson, S. C., Michael, S. T., & Cheavens, J. (2001). Optimism and hope constructs: Variants on a positive expectancy theme. In E. C. Chang (Ed.), *Optimism & pessimism: Implications for theory, research, and practice* (pp. 101–125). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10385-005

Staley, D. J. (2002). A history of the future. *History and Theory, 41,* 72–89. http://dx.doi.org/10.1111/1468-2303.00221

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131,* 159–171. http://dx.doi.org/10.1016/j.cognition.2013.12.005

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193–210. http://dx.doi.org/10.1037/0033-2909.103.2.193

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522. http://dx.doi.org/10.1038/381520a0

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323. http://dx.doi.org/10.1007/BF00122574

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, NY: Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199730018.013.0004

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39,* 806–820. http://dx.doi.org/10.1037/0022-3514.39.5.806

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4,* 1120–1135. http://dx.doi.org/10.1167/4.12.11

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. http://dx.doi.org/10.1037/0022-3514.80.4.572

Withey, S. B. (1954). Reliability of recall of income. *Public Opinion Quarterly, 18,* 197–204. http://dx.doi.org/10.1086/266505

Ybarra, O. (2002). Naive causal understanding of valenced behaviors and its implications for social information processing. *Psychological Bulletin, 128,* 421–441. http://dx.doi.org/10.1037/0033-2909.128.3.421

Appendix A

Stimuli Used in Study 1

I. Study 1a (Change Scaled in Terms of *Frequency*; Each Item Rated From 1 to 10)

Athletic performance. Think about a professional athlete who plays a long season of games. They began the season with quite a good [bad] stretch of games in terms of their performance. Out of the next 10 games: how many must be bad [good] for you to feel a cold streak [hot streak] is "officially" here? In other words, how many bad [good] games would you need to see to be convinced their performance has indeed changed for the worse [better] in a lasting way?

Academic performance. Think about a college student who has to take many tests and quizzes over the semester. So far, they have received a number of high [low] grades on these assignments. Out of the next 10 grades: how many must be low [high] for you to feel this student has "officially" become bad [good]? In other words, how many low [high] grades would you need to see to be convinced their performance has indeed changed for the worse [better] in a lasting way?

Physical health. Imagine a person's recent healthy [unhealthy] weekend routines have led them to lose [gain] a few pounds. Out of the next 10 weekends: how many must be unhealthy [healthy] for you to feel this person is in "officially" bad [good] shape? In other words, how many healthy weekends would you need to see to be convinced their shape has indeed changed for the worse [better] in a lasting way?

Mood. Imagine a person has felt quite happy [sad] recently, and their mood has been consistently positive [negative]. Out of the next 10 days: how many must be sad [happy] for you to feel this person's mood is "officially" negative [positive]? In other words, how many sad [happy] days would you need to see to be convinced their mood has indeed changed for the worse [better] in a lasting way?

Luck. Imagine a person has taken a trip to the local casino. The outcomes of the games have been going somewhat well [poorly]. Out of the next 10 games: how many must they lose [win] for you to feel their luck is "officially" bad [good]? In other words, how many losses [wins] would you need to see to be convinced their luck has indeed changed for the worse [better] in a lasting way?

Habits. Imagine a person has a good [bad] routine of going to bed very early [late], such that they get high [low]-quality sleep. Out of the next 10 nights: how many must end late [early] for you to feel a good habit has "officially" developed? In other words, how many early nights would you need to see to be convinced their habit has indeed changed for the worse [better] in a lasting way?

Friendship. Imagine you meet a new co-worker of the same gender and you two get off to a good [bad] start. Out of the next 10 interactions: how many must be unpleasant [pleasant] for you to feel you're "officially" bad [good] as friends? In other words, how many unpleasant [pleasant] interactions would you need to see to be convinced the relationship has indeed changed for the worse [better] in a lasting way?

Personality. Imagine a person recently has been exhibiting a lot of prosocial tendencies, like being calm, confident, and tolerant [antisocial tendencies, like being neurotic, insecure, and close-minded]. Out of the next 10 weeks: during how many must they exhibit antisocial [prosocial] tendencies for you to feel their personality is "officially" bad [good]? In other words, how many antisocial [prosocial] weeks would you need to see to be convinced their personality has indeed changed for the worse [better] in a lasting way?

II. Study 1b (Change Scaled in Terms of *Duration*; Each Item Rated From 1 to 10)

Athletic performance. Think about a professional athlete who plays a long season of games. They began the season with quite a good [bad] stretch of games in terms of their performance. However, they've recently hit a cold streak [hot streak] of playing poorly [well]. For how long must this trend continue for you to feel they have "officially" become bad [good]? In other words, how long would things need to stay bad [good] for you to be convinced their performance has indeed changed for the worse [better] in a lasting way?

Academic performance. Think about a college student who has to take many tests and quizzes over the semester. So far, they have received a number of high [low] grades. However, they've received low [high] grades on the past few assignments. For how long must this trend continue for you to feel they have "officially" become bad [good]? In other words, how long would things need to stay bad [good] for you to be convinced their performance has indeed changed for the worse [better] in a lasting way?

Physical health. Imagine a person's healthy [unhealthy] weekend routines have led them to lose [gain] a few pounds. However, they've been rather unhealthy [healthy] during the past few weekends and are starting to look worse [better]. For how long must this trend continue for you to feel this person is in "officially" bad [good] shape? In other words, how long would things need to stay bad [good] for you to be convinced their shape has indeed changed for the worse [better] in a lasting way?

(*Appendices continue*)

Mood. Imagine a person has felt quite happy [sad] recently, and their mood has been consistently positive [negative]. However, they've started to show signs of feeling negative [positive]. For how long must this trend continue for you to feel their mood is "officially" negative [positive]? In other words, how long would things need to stay bad [good] for you to be convinced their mood has indeed changed for the worse [better] in a lasting way?

Luck. Imagine a person has taken a trip to the local casino. The outcomes of the games have been going somewhat well [poorly]. However, they've recently hit a cold streak of losing [hot streak of winning]. For how long must this trend continue for you to feel their luck is "officially" bad [good]? In other words, how long would things need to stay bad [good] for you to be convinced their luck has indeed changed for the worse [better] in a lasting way?

Habits. Imagine a person has a good [bad] routine of going to bed very early [late], such that they get high [low]-quality sleep. However, they've recently started to follow a bad [good] routine and have gone to bed later [earlier]. For how long must this trend continue for you to feel a bad [good] habit has "officially" developed? In other words, how long would things need to stay bad [good] for you to be convinced their habit has indeed changed for the worse [better] in a lasting way?

Friendship. Imagine you meet a new co-worker of the same gender and you two get off to a good [bad] start. Recently, however, your interactions have been more unpleasant [pleasant]. For how long must this trend continue for you to feel you're "officially" bad [good] as friends? In other words, how long would things need to stay bad [good] for you to be convinced the relationship has indeed changed for the worse [better] in a lasting way?

Personality. Imagine a person recently has been exhibiting a lot of prosocial tendencies, like being calm, confident, and tolerant [antisocial tendencies, like being neurotic, insecure, and close-minded]. Recently, however, they've started to exhibit more antisocial [prosocial] tendencies. For how long must this trend continue for you to feel their personality is "officially" bad [good]? In other words, how long would things need to stay bad [good] for you to be convinced their personality has indeed changed for the worse [better] in a lasting way?

III. Study 1c (Change Scaled in Terms of *Magnitude*; Each Item Rated From 1 to 10)

Athletic performance. Think about a professional athlete who plays the same sport every season. Last season was mostly filled with good [bad] games in terms of their performance. How bad [good] must next season be for you to feel this time around is "officially" worse [better]? In other words, how much decline [improvement] would convince you their performance has indeed changed for the worse [better] in a lasting way?

Academic performance. Think about a college student who has to take many tests and quizzes over the next two semesters. They received a number of high [low] grades during Semester 1. How bad [good] must Semester 2 be for you to feel this time around is "officially" worse [better]? In other words, how much decline [improvement] would convince you their performance has indeed changed for the worse [better] in a lasting way?

Physical health. Imagine a person abstained from [engaged in] unhealthy eating during the last few social events, which led them to lose [gain] a few pounds. How bad [good] must the next few social events be for you to feel this time around is "officially" worse [better]? In other words, how much decline [improvement] would convince you their eating behavior at social events has indeed changed for the worse [better] in a lasting way?

Mood. Imagine a person felt quite happy [sad] during the past year, and their mood was consistently positive [negative]. How bad [good] must the next year be for you to feel this time around is "officially" worse [better]? In other words, how much decline [improvement] would convince you their mood has indeed changed for the worse [better] in a lasting way?

Luck. Imagine a person takes trips to the local casino on weekends. Last weekend, the outcomes of the games went somewhat well [poorly]. How bad [good] must next weekend be for you to feel this time around is "officially" worse [better]? In other words, how much decline [improvement] would convince you their luck has indeed changed for the worse [better] in a lasting way?

Habits. Imagine a person has a good [bad] routine of going to bed very early [late], such that they get high [low]-quality sleep. This has been going on for the past year or so. How bad [good] must next year be for you to feel this time around is "officially" worse [better]? In other words, how much decline [improvement] would convince you their habit has indeed changed for the worse [better] in a lasting way?

Friendship. Imagine you meet a new co-worker of the same gender and you two get off to a good [bad] start during Quarter 1 together. How bad [good] must Quarter 2 be for you to feel this time around is "officially" worse [better]? In other words, how much decline [improvement] would convince you the relationship has indeed changed for the worse [better] in a lasting way?

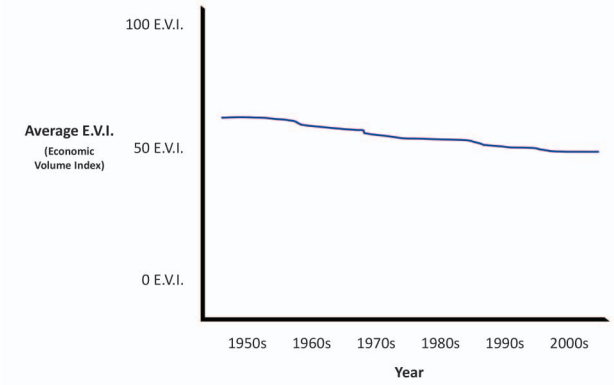
Personality. Imagine a person exhibited a lot of prosocial tendencies, like being calm, confident, and tolerant [antisocial tendencies, like being neurotic, insecure, and close-minded], during the past year or so. How antisocial [prosocial] must they be during the next year or so for you to feel this time around is "officially" worse [better]? In other words, how much decline [improvement] would convince you their personality has indeed changed for the worse [better] in a lasting way?

(*Appendices continue*)

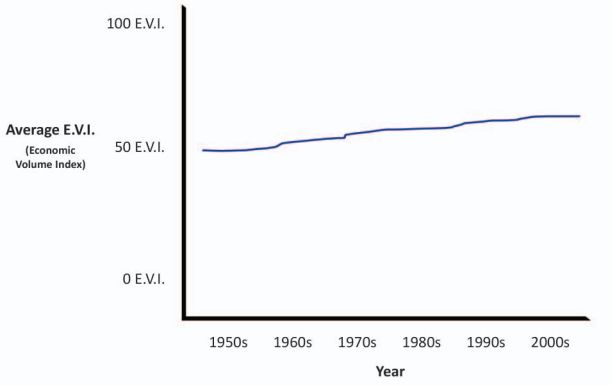
Appendix B

Stimuli used in Study 2

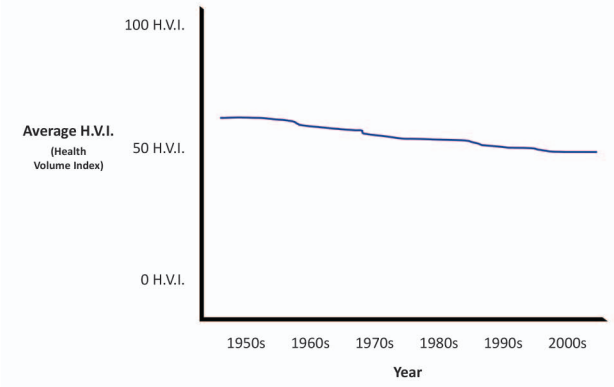
Economy – downward sloping (with "downward" framed as reflecting either decline or improvement)



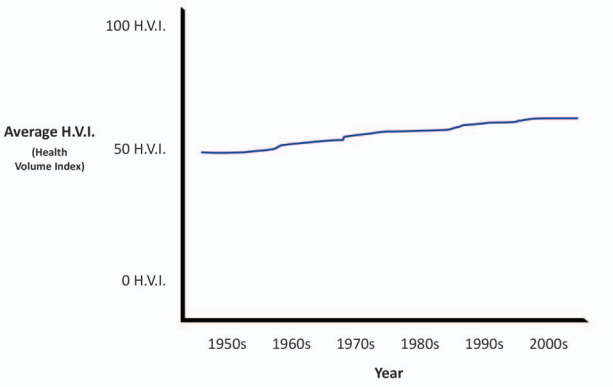
Economy – upward sloping (with "upward" framed as reflecting either decline or improvement)



Health – downward sloping (with "downward" framed as reflecting either decline or improvement)



Health – upward sloping (with "upward" framed as reflecting either decline or improvement)



(*Appendices continue*)

Appendix C

Stimuli Used in Study 4

I. Study 4a (Changes in Self Versus Others; Each Item Rated From 1 to 10)

Work performance: Self. Imagine you have started showing signs of becoming a worse [better] worker at the office, and have completed the last few assignments quite poorly [well]. However, because work efforts can ebb and flow for a variety of reasons, it's unclear whether your new lazy [hard-working] attitude is "real" and will last. For how long must you keep exhibiting this attitude for you to feel you have become "officially" bad [good]? In other words, for how long would this need to continue for you to be convinced your office performance has indeed changed for the worse [better], that your initial bad [good] signs weren't just a fluke?

Work performance: Other. Imagine another person has started showing signs of becoming a worse [better] worker at the office, and has completed the last few assignments quite poorly [well]. However, because work efforts can ebb and flow for a variety of reasons, it's unclear whether their new lazy [hard-working] attitude is "real" and will last. For how long must they keep exhibiting this attitude for you to feel they have become "officially" bad [good]? In other words, for how long would this need to continue for you to be convinced their office performance has indeed changed for the worse [better], that their initial bad [good] signs weren't just a fluke?

Health: Self. Imagine that your levels of dieting and physical activity have quickly led you to gain [lose] a few pounds and get in worse [better] shape. However, because body tone can fluctuate rather often and easily, it's unclear whether your new look is "real" and will last. For how long must you keep in worse shape for you to feel your look has become "officially" bad [good]? In other words, for how long would this need to continue for you to be convinced your shape has indeed changed for the worse [better], that your initial gain [loss] of weight wasn't just a fluke?

Health: Other. Imagine that another person's levels of dieting and physical activity have quickly led them to gain [lose] a few pounds and get in worse [better] shape. However, because body tone can fluctuate rather often and easily, it's unclear whether their new look is "real" and will last. For how long must they keep in worse shape for you to feel their look has become "officially" bad [good]? In other words, for how long would this need to continue for you to be convinced their shape has indeed changed for the worse [better], that their initial gain [loss] of weight wasn't just a fluke?

Luck: Self. Imagine that you have incorrectly [correctly] guessed the outcomes of a few recent games involving coin flips. However, because these games are random, it's unclear whether your streak is "real" and will last. For how long must you keep losing for you to feel your luck has become "officially" bad [good]? In other words, for how long would this need to continue for you to be convinced your luck has indeed changed for the worse [better], that your initial losses [wins] weren't just a fluke?

Luck: Other. Imagine that another person has incorrectly [correctly] guessed the outcomes of a few recent games involving coin flips. However, because these games are random, it's unclear whether their streak is "real" and will last. For how long must they keep losing for you to feel their luck has become "officially" bad [good]? In other words, for how long would this need to continue for you to be convinced their luck has indeed changed for the worse [better], that their initial losses [wins] weren't just a fluke?

II. Study 4b (Additive Versus Subtractive Changes; Each Item Rated From 1 to 10)

Community news: Additive. Imagine you hear a "normal" range of stories and events that go on in your neighborhood: sometimes you hear especially good, sometimes you hear especially bad, and a lot of times you hear relatively neutral news. However, you notice a change: you've started to hear more and more bad [good] news. For how long must this continue for you to feel things are "officially" bad [good]? In other words, how long would this "new presence" of bad [good] news need to persist for you to be convinced things have indeed changed for the worse [better] in a lasting way?

Community news: Subtractive. Imagine you hear a "normal" range of stories and events that go on in your neighborhood: sometimes you hear especially good, sometimes you hear especially bad, and a lot of times you hear relatively neutral news. However, you notice a change: you've started to hear less and less good [bad] news. For how long must this continue for you to feel things are "officially" bad [good]? In other words, how long would this "new absence" of good [bad] news need to persist for you to be convinced things have indeed changed for the worse [better] in a lasting way?

Work feedback: Additive. Imagine you get a "normal" range of feedback about your performance at work: sometimes you get great positive feedback, sometimes you get poor negative feedback, and a lot of times you get standard routine feedback. However, you notice a change: you've started to get more and more poor [great] feedback. For how long must this continue for you to feel things are "officially" bad [good]? In other words, how long would this "new presence" of poor [great] feedback need to persist for you to be convinced things have indeed changed for the worse [better] in a lasting way?

(*Appendices continue*)

Work feedback: Subtractive. Imagine you get a "normal" range of feedback about your performance at work: sometimes you get great positive feedback, sometimes you get poor negative feedback, and a lot of times you get standard routine feedback. However, you notice a change: you've started to get less and less great [poor] feedback. For how long must this continue for you to feel things are "officially" bad [good]? In other words, how long would this "new absence" of great [poor] feedback need to persist for you to be convinced things have indeed changed for the worse [better] in a lasting way?

Emotion: Additive. Imagine you experience a "normal" range of emotions in your daily life: sometimes you feel especially happy, sometimes you feel especially sad, and a lot of times you feel just okay. However, you notice a change: you've started to experience more and more sad [happy] moments. For how long must this continue for you to feel things are "officially" bad [good]? In other words, how long would this "new presence" of sad [happy] moments need to persist for you to be convinced things have indeed changed for the worse [better] in a lasting way?

Emotion: Subtractive. Imagine you experience a "normal" range of emotions in your daily life: sometimes you feel especially happy, sometimes you feel especially sad, and a lot of times you feel just okay. However, you notice a change: you've started to experience less and less happy [sad] moments. For how long must this continue for you to feel things are "officially" bad [good]? In other words, how long would this "new absence" of happy [sad] moments need to persist for you to be convinced things have indeed changed for the worse [better] in a lasting way?

III. Study 4c (Intended Versus Unintended Changes; Each Item Rated From 1 to 10)

Athletic performance: Intended. Imagine that you're monitoring a sports team over time. As you watch game after game and see them start to lose [win], you're trying to figure out whether a lasting cold streak [hot streak] might be forming versus whether this might just be a passing fluke. One thing to note is that the team and players are actively hoping to decline [improve] – they're truly trying to be bad [good] for various reasons. And here they are, happening to lose [win]. As you keep monitoring: For how long must this trend of losses [wins] continue for you to feel that a cold streak [hot streak] is "officially" here? In other words, how long would the team need to happen to keep losing [winning] in this way to convince you that something "real" is going on rather than just a passing fluke?

Athletic performance: Unintended. Imagine that you're monitoring a sports team over time. As you watch game after game and see them start to lose [win], you're trying to figure out whether a lasting cold streak [hot streak] might be forming versus whether this might just be a passing fluke. One thing to note is that the team and players are actively hoping to improve [decline] – they're truly trying to be good [bad] for various reasons. And here they are, happening to lose [win]. As you keep monitoring: For how long must this trend of losses [wins] continue for you to feel that a cold streak [hot streak] is "officially" here? In other words, how long would the team need to happen to keep losing [winning] in this way to convince you that something "real" is going on rather than just a passing fluke?

Romantic life: Intended. Imagine that you're monitoring your current romantic relationship over time. You notice that date after date is starting to go poorly [well]. As this is unfolding, you're trying to figure out whether lasting change for the worse [better] might be forming versus whether this might just be a passing fluke. One thing to note is that you are actively hoping to decline [improve] – you want to make this relationship bad [good] for various reasons. And here the dates are, happening to go poorly [well]. As you keep monitoring: For how long must this trend of bad [good] dates continue for you to feel that decline [improvement] is "officially" here? In other words, how long would things need to happen to keep going poorly [well] in this way to convince you that something "real" is going on rather just a passing fluke?

Romantic life: Unintended. Imagine that you're monitoring your current romantic relationship over time. You notice that date after date is starting to go poorly [well]. As this is unfolding, you're trying to figure out whether lasting change for the worse [better] might be forming versus whether this might just be a passing fluke. One thing to note is that you are actively hoping to improve [decline] – you want to make this relationship good [bad] for various reasons. And here the dates are, happening to go poorly [well]. As you keep monitoring: For how long must this trend of bad [good] dates continue for you to feel that decline [improvement] is "officially" here? In other words, how long would things need to happen to keep going poorly [well] in this way to convince you that something "real" is going on rather just a passing fluke?

Academic performance: Intended. Imagine that you're monitoring a friend's academic performance over time. You notice that assignment after assignment is starting to go poorly [well]. As this is unfolding you're trying to figure out whether lasting change for the worse [better] might be forming versus whether this might just be a passing fluke. One thing to note is that your friend is actively hoping to decline [improve] – they want to become a bad [good] student for various reasons. And here the assignments are, happening to go poorly [well]. As you keep monitoring: For how long must this trend of bad [good] assignments continue for you to feel that decline [improvement] is "officially" here? In other words, how long would things need to happen to keep going poorly [well] in this way to convince you that something "real" is going on rather than just a passing fluke?

Academic performance: Unintended. Imagine that you're monitoring a friend's academic performance over time. You notice that assignment after assignment is starting to go poorly [well]. As this is unfolding you're trying to figure out whether lasting change for the worse [better] might be forming versus whether this might just be a passing fluke. One thing to note is that your friend is actively hoping to improve [decline] – they want to become a good [bad] student for various reasons. And here the assignments are, happening to go poorly [well]. As you keep monitoring: For how long must this trend of bad [good] assignments continue for you to feel that decline [improvement] is "officially" here? In other words, how long would things need to happen to keep going poorly [well] in this way to convince you that something "real" is going on rather than just a passing fluke?

Received December 29, 2015
Revision received October 26, 2016
Accepted November 11, 2016 ■