Full length article

The role of episodic simulation in motivating commonplace harms

Adam Morris a, *, Brendan Bo O’Connor b, Fiery Cushman a

a Department of Psychology, Harvard University, Cambridge, MA, United States of America
b Department of Psychology, University at Albany, State University of New York, Albany, NY, United States of America

ARTICLE INFO

Keywords:
Episodic simulation
Imagination
Harm
Affect
Morality

ABSTRACT

Every day, people face choices which could produce negative outcomes for others, and understanding these decisions is a major aim of social psychology. Here, we show that episodic simulation – a key psychological process implicated in other types of social and moral decision-making – can play a surprising role. Across six experiments, we find that imagining performing actions which adversely affect others makes people report a higher likelihood of performing those actions in the future. This effect happens, in part, because when people construe the actions as morally justified (as they often do spontaneously), imagining doing it makes them feel good. These findings stand in contrast to traditional accounts of harm aversion in moral psychology, and instead contribute to a growing body of evidence that people often cast harming others in a positive light.

1. Introduction

People commonly make choices which produce negative outcomes for others, in ways big and small. They lie to a friend; take the last seat on the bus when they know someone else needs it more; take credit for another’s work; and so on. People’s choices to harm others produce enormous social costs, and understanding why people do it has been a major aim of social psychology (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Cushman, Young, & Hauser, 2006; Fehr & Fischbacher, 2004; Graham et al., 2013; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Henrich et al., 2010; Mikhail, 2000). (By “harming others”, we mean performing actions which canonically violate moral norms and produce negative outcomes for others—e.g. lying, stealing—even if those actions might be justified in context [Cushman et al., 2006]).

Here, we demonstrate a simple but novel route by which people arrive at a higher reported willingness to harm. The route lies at the juncture of two ideas. First, people evaluate actions by imagining them (i.e., episodic simulation), gauging how taking different actions might feel. Second, when imagining harmful actions, people sometimes imagine them in justified contexts, which makes them feel good about the thought of harming others. Thus, as we show, by asking a person to imagine committing small harms (e.g. taking the last bus seat from an elderly person), on average they will report a higher likelihood of actually performing those harms. This effect is somewhat striking: Naively, one might have thought that being asked to imagine harmful acts would reinforce one’s aversion to it. Although we show that this effect only occurs when people view the harm as morally justified, people often do so spontaneously (telling themselves, in the words of one participant, that “old people are usually rude anyways on public transportation”). Collectively, our results sharpen current models of episodic simulation, challenge some models of moral learning, and add to a burgeoning literature on the ways that harm and violence often get cast in a positive light (Cikara, Bruneau, Van Bavel, & Saxe, 2014; Cikara & Fiske, 2012; Fiske & Rai, 2014; Rai, Valdesolo, & Graham, 2017).

1.1. Episodic simulation and moral decision-making

Imagining oneself performing harm is an act of “episodic simulation”: the ability to produce detailed mental representations of specific events not directly experienced (Schacter, Addis, & Buckner, 2008). When a romantic fantasizes about his far-off wedding day, or when an office worker daydreams about telling off her boss, they are engaging in episodic simulation. A core human ability, it plays a role in many parts of cognition, such as creativity, emotion regulation, prospective memory, and spatial navigation (Bulley, Redshaw, & Suddendorf, 2020; Schacter, Benoit, & Szpunar, 2017).

Most relevant for our purposes, episodic simulation allows people to guess how their future actions might make them feel (Gilbert & Wilson, 2007; Schacter et al., 2008; Suddendorf & Corballis, 2007). People construct vivid simulations of action-outcome contingencies, and then alter their behavior according to the affective signal of those simulated
outcomes (Benoit, Szpunar, & Schacter, 2014; Bulley et al., 2020; Bulley, Henry, & Suddendorf, 2017; Ernst & D’Argembeau, 2017; Gershman, Markman, & Ross, 2014; Schacter et al., 2008; Suddendorf & Corballis, 2007; Szpunar, Addis, & Schacter, 2012). This influence of episodic simulation extends to moral judgment and decision-making (Amit & Greene, 2012; Caruso & Gino, 2011). For instance, Gaesser & colleagues have shown that when people simulate helping others, they report being more likely to help in the future (Gaesser & Fowler, 2020; Gaesser, Hirschfeld-Kroen, Wasserman, Horn, & Young, 2019; Gaesser, Keeler, & Young, 2018; Gaesser & Schacter, 2014; Sawczak, McAndrews, Gaesser, & Moscovitch, 2019), especially when the simulations elicit positive affect (Gaesser, DiBiase, & Kensinger, 2017).

1.2. Episodic simulation and committing harm

Little is known, however, about the role of episodic simulation in motivating an especially important class of moral choices: choices to harm others. As noted above, we use “harm” to denote actions which canonically violate moral norms and produce negative outcomes for others – e.g. lying, cutting in line, taking another person’s seat, and so on – even if those actions are prescriptively justified in their context (e.g. by producing additional benefits). For instance, on this definition, when Robin Hood steals from the rich to give to the poor, we consider him to have harmed the rich (i.e. taken an action which adversely affected them) even if the act was ethically justified and produced benefits for the poor. This definition aligns with research showing that people are often averse to canonically-harmful actions themselves, divorced from their consequences or context (Crockett, 2013; Cushman, 2013; Cushman, Gray, Gaffey, & Mendes, 2012; Greene et al., 2001; Lockwood, Klein-Fligge, Abdurahman, & Crockett, 2020; R. Miller & Cushman, 2013; R. M. Miller, Hannikainen, & Cushman, 2014; Rozin, Millman, & Nemeroff, 1986), as well as research on schadenfreude and motivated violence showing how people cause negative outcomes for other in contexts where they believe they are justified to do so (Feather, 2006a, 2006b; Feather & Sherman, 2002).

We know that people imagine harmful acts. For instance, in surveys of college students, well more than half of students describe having fantasized about killing somebody (Buss, 2006; Kenrick & Sheets, 1993). How does the mental act of imagining harm affect people’s attitudes and beliefs about the actual event? Does it dissuade people from harm, or prepare them for it? Several theories make divergent predictions on this question.

1.2.1. Harm aversion hypothesis

An intuitive prediction is that imagining harming others would decrease one’s willingness to commit harm because the thought of harming others feels aversive. Indeed, a central tenant of moral psychology – from theoretical perspectives as diverse as domain theory (Nucci & Turiel, 1978; Turiel, 1983), dyadic morality (Schein & Gray, 2018), moral foundations theory (Graham et al., 2013), and more (Blair, 1995) – is that people are averse to harming others (Crockett et al., 2014). People judge harm to be morally wrong (Cushman et al., 2013), even when causing harm to one person could prevent harm to several others (Cushman et al., 2006; Greene et al., 2001; Mikhail, 2006; Petrinovich & O’Neill, 1996). They show physiological signs of aversion to canonically harmful acts such as shooting, even if they know the gun is fake and harmless (Cushman et al., 2012). They avoid throwing darts at a picture of someone’s face (Rozin et al., 1986), and will pay money to avoid shocking others (Crockett et al., 2014). There is even anecdotal evidence that soldiers in front-line conflict will deliberately avoid shooting at hostile enemy forces (Grossman, 2009). Given the apparently predominant negative emotional response to harming, it is plausible that episodically simulating harm would reinforce an aversion to it. We label this the “harm aversion” hypothesis (Table 1).

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harm aversion</td>
<td>Imagining committing harm will make people feel bad and report a lower likelihood of harming.</td>
</tr>
<tr>
<td>Schadenfreude</td>
<td>Imagining committing harm, when the harm is construed as morally justified, will make people feel good and report a higher likelihood of harming.</td>
</tr>
<tr>
<td>Imagination inflation</td>
<td>Imagining harm will make people report a higher likelihood of harming, independent of justification or affect.</td>
</tr>
</tbody>
</table>

1.2.2. Schadenfreude hypothesis

This picture is complicated, however, by other work showing that people sometimes feel good about other’s misfortunes – a phenomenon known as “schadenfreude” (Cikara, 2015; Cikara, Bruneau, & Saxe, 2011; Cikara & Fiske, 2012; Feather, 2006b; Feather & Sherman, 2002; Leach, Spears, Branscombe, & Doosje, 2005; Smith et al., 1996; Smith, Powell, Combs, & Schurtz, 2009; Takahashi et al., 2009). People commonly feel schadenfreude when they stand to gain from the other person’s misfortune (Cikara et al., 2014; Smith, Eyre, Powell, & Kim, 2006), or when they believe the harm to be ethically justified (Feather, 2006a, 2006b; Feather & Sherman, 2002). For instance, people report feeling good when those in competition with them experience misfortune (Cikara et al., 2014), or when immoral actors get their comeuppance (Feather, 2006b). And this phenomenon is not only seen in the laboratory; think of Americans celebrating the death of Osama bin Laden (Bowman, Lewis, & Tambornini, 2014).

Moreover, when people commit harm in real life, they often believe that it is morally justified, or has a virtuous purpose (Fiske & Rai, 2014; Rai & Fiske, 2012); think of Nazis who believed their genocides were purifying their race, or a violent father who believe corporal punishment is for his child’s good. In ethnological study of violence in cultures across the world, Fiske and Rai (2014) find that most violence is done, in the mind of the perpetrator, for morally justified reasons (Rai et al., 2017). Together, these bodies of work render a different prediction. They predict that the effect of simulating harm will depend on whether the harm is believed to be justified. If people construe the harm as morally justified (which they often do; Fiske & Rai, 2014), then simulating performing the harm will make them feel good, not bad, and this experience will make them more willing to perform the harm in the future. We label this the “schadenfreude” hypothesis.

1.2.3. Imagination inflation hypothesis

Finally, one other popular view predicts that simulating performing harm will consistently increase people’s willingness to harm, regardless of justification or affect. A large body of evidence suggests a general “imagination inflation” effect (Garry & Polaschek, 2000), whereby imagining any event makes the representation of the event more accessible, and hence makes the event seem more likely. In a classic example, people who imagine a presidential candidate winning the election subsequently believe that that candidate is more likely to win (Carroll, 1978). Importantly, the imagination inflation effect is thought to be agnostic to the valence of the event; people who imagine getting their teeth pulled out, or their skin removed, subsequently believe that those events are more likely to occur (Mazzaoni, Loftus, & Kirsch, 2001; Mazzaoni & Memon, 2003; Szpunar & Schacter, 2013). Applying this logic to our domain, simulating performing harm should make people report that they are more likely to perform harm in the future, independent of the valence of the harm. We label this the “imagination inflation” hypothesis.

Simulating performing harm, then, is an attractive test case because different literatures make starkly different predictions. In six experiments, we tested these predictions by asking participants to simulate themselves performing mild, commonplace harms (like lying to a house guest or taking a bus seat from an elderly person). To preview our results, we find that episodic simulation can influence people’s beliefs that
they would commit harm, but not in the way predicted by accounts of harm aversion or imagination inflation. Rather, we find evidence most consistent with the schadenfreude account. When people imagine themselves committing harm, they subsequently say that they would be more likely to commit the harm in the future, and this effect occurs because many people spontaneously construe the harm as morally justified (Experiments 1a-b) and because simulating justified harm makes people feel good (Experiments 2–5). These results shed light on the role of episodic simulation in moral decision-making and suggest that, for mild, commonplace harms, imagination pushes people towards believing they would commit harm.

2. Experiment 1

2.1. Overview

In Experiment 1, we tested whether vividly imagining harming someone increases or decreases people’s reported likelihood of actually harming them. We base our design on Gaesser et al.’s paradigm, which examined the effect of episodic simulation on self-reported willingness to help others (Gaesser et al., 2018; Gaesser, DiBiase, & Kensing, 2017; Gaesser & Schacter, 2014). In the studies of Gaesser et al., vividly simulating commonplace helpful actions increased people’s reported willingness to help people in need. We adopted Gaesser et al.’s design, except instead of imagining performing commonplace helpful actions people imagined performing commonplace harms. This allowed us to test, in an established paradigm, whether simulating harmful behaviors increases or decreases people’s reported likelihood of harming.

Moreover, to examine the role of moral justification in this effect (which is crucial for distinguishing between the three hypotheses in Table 1), we also asked participants to report the extent to which they construed the behavior as morally justified. On the harm aversion and imagination inflation accounts, the effect of simulating performing harm should not depend on justification; simulation should either reduce (harm aversion) or increase (imagination inflation) people’s reported willingness to harm. In contrast, the schadenfreude account predicts that the effect will depend on whether people construe the harm as morally justified: Simulating more justified harm will make people want to do it more. To preview our results, we find evidence most consistent with the schadenfreude account. People often spontaneously found ways to rationalize or justify the harms, and simulating performing harm increased people’s reported willingness to harm only when the harms were construed as relatively more justified.

Following Gaesser’s paradigm, all our experiments focus on mild, common behaviors – e.g. lying to a friend, taking someone else’s seat on public transportation – rather than more extreme and rare harms. This approach has the advantage of staying within the realm of behaviors and choices people often encounter in their normal lives, and aligns with some prior experimental work on schadenfreude (which focused on relatively mild misfortunes like failing an exam or getting soaked by a passing taxi; Gikara & Fiske, 2012; Feather & Sherman, 2002). Nonetheless, this experiment – and the ones that follow – do not directly speak to the effects of episodic simulation on more extreme harms, and we discuss this limitation further in the General Discussion.

In this section, we describe Experiment 1a, our initial test of the effect of imagining performing harm; Experiment 1b, a pre-registered replication of Experiment 1a; and two supplementary analyses. All data and analysis code for this manuscript can be found at https://researchbox.org/260.

2.2. Experiment 1a

Experiment 1a is our initial test of the effect of imagining performing harm.

2.2.1. Methods

2.2.1.1. Participants. All participants gave informed consent, and all experiments were approved by Harvard’s Committee on the Use of Human Subjects. All our participants were recruited on Amazon Mechanical Turk and required to (a) live in the US (cross-checked against their IP address), (b) have an approval rating of at least 98%, (c) have completed at least 100 studies before this one, and (d) not have participated in any of our prior studies on this topic. Furthermore, in all our studies, participants were excluded for ending the study prematurely, or for giving (on average) fewer than 50 characters per simulation description. Throughout all our experiments, we ended data collection before analyzing the results.

In Experiment 1, 119 participants were recruited on Mechanical Turk, and 23 (19% of the sample) were excluded, leaving 96 for analysis. A post-hoc power analysis (using the R package mixdpower; Kumie, Vo, & Draschckow, 2021) indicated that this sample size gave us over 80% power to detect the effect of imagining harm on reported likelihood of harming (see SM for details). Due to an error, we failed to collect demographic information for Experiments 1–4. Experiment 5 reports demographic information. This error is unlikely to affect the generalizability of our conclusions; Mechanical Turk samples are known to typically be more representative of the general US population than in-person convenience samples (Berinsky, Huber, & Lenz, 2012), and our sample in Experiment 5 proved to be demographically balanced. Moreover, in Experiment 5 we find that there are no detectable differences in our effects across age, gender, or race. Nonetheless, the lack of demographic information in Experiments 1–4 should be noted as a limitation of our results.

2.2.1.2. Design and stimuli. Participants read vignettes like the following: “You’re about to sit down on a crowded train, when you see an elderly person searching for a seat. You sit down in the seat.” The final sentence was underlined in order to clearly demarcate the important behavior.

There were 10 vignettes in total. On half of those (“imagine” trials), people were asked to vividly imagine themselves performing the underlined behavior in a plausible way, with as much detail as possible. On the other half of trials (“control” trials), people were asked to generate words that rhymed with words in the vignette. They were asked to generate as many rhymes as possible, and were forced to spend an equal amount of time as on the imagine trials. This control task, similar to those used in past work (Gaesser & Schacter, 2014), encouraged participants to engage with the vignettes without recruiting episodic simulation. See SM for the exact vignettes and instructions.

2.2.1.3. Procedure. Participants received instructions on the task, including how to simulate episodes with high detail and vividness. They were then given the 10 vignettes in a random order. Half of the vignettes were randomly assigned to be imagine trials, and the other half to be control trials. On all trials, people were asked to read the vignette and continue to the next screen when finished. On the next screen, for 1 min, participants were asked to imagine themselves performing the behavior (imagine trial) or generate words that rhymed with words in the vignette (control trial). After a minute, the screen advanced, and participants were given 2 min to write (i.e. enter into a textbox) what they had imagined or what rhymes they had generated. When they clicked ‘submit’, they were taken to a screen where they answered the following questions. First, they were asked: “If you were in this situation, how likely would you be to perform the underlined behavior?” They responded on a continuous Likert scale from 1 (“very unlikely”) to 7 (“very likely”). Second, they were asked: “How morally justified was the underlined behavior?”, on a continuous Likert scale from 1 (“very unjustified”) to 7 (“very justified”).
2.2.1.4. Statistical analysis. To analyze the effect of trial type (imagine vs. control) on our dependent variable (reported likelihood of harming), we estimated a linear mixed-effects model (LMM). Mixed-effects models are being increasingly used in contemporary analyses of episodic simulation (Gaesser, Shimura, & Cikara, 2020; Vollberg, Gaesser, & Cikara, 2021), and are theoretically preferred for many reasons to traditional analyses like ANOVA (Barr, Levy, Scheepers, & Tily, 2013; Boisgontier & Cheval, 2016; Gueorguieva & Krystal, 2004; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017; Yu et al., 2022). We regressed reported likelihood of harming on trial type, with random effects grouped by subject and vignette. We used maximal random slopes and intercepts (i.e. random intercepts and trial-type slopes for both subject and vignette; in R syntax, reported likelihood ~ trial type + (trial type | subject) + (trial type | vignette)), unless they prevented model convergence (in which case we dropped slopes one at a time, starting with the highest-order slope, until the model converged). This procedure is a standard recommendation for estimating mixed effects models (Barr et al., 2013; Bates, Kliegl, Vasishth, & Baayen, 2018; Matuschek et al., 2017), balancing the need to prevent false positives with the need to ensure model convergence. (Another way to put it is that this approach ensures convergence by only including random effects which capture nonzero variance in the data; see SM for details.) Due to this procedure, the final LMM estimated for each analysis includes a different set of random effects; we report the final LMMs for each analysis in the SM.

Models were estimated with the afex (Singmann et al., 2021) and lme4 packages in R (Bates, Mächler, Bolker, & Walker, 2015). We assessed the significance of all parameters via t-tests using the standard Satterthwaite approximation to the denominator degrees of freedom in the package lmerTest (Luke, 2017). Numbers in brackets are 95% confidence intervals, estimated using the “confint” function from the lme4 package with the default likelihood profile method. All results are rounded to two significant figures.

To analyze the role of moral justification, we dichotomized justification ratings using a median split. We did this so that our tests would not depend on how people were interpreting the scale (i.e. how they were mapping “Very unjustified” or “Very justified” onto their internal moral representations). Dichotomizing this variable with a median split allowed us to simply test whether the half of trials on which people construed the harms as relatively more justified differed from the half of trials on which people construed them as relatively less justified. Using this dichotomized, relative justification variable, we estimated a LMM regressing reported likelihood of harming on relative moral justification (high vs low), trial type (imagine vs control), and their interaction. Random effects were included in the same manner as above, starting with maximal random effects (i.e. reported likelihood ~ justification * trial type + (justification * trial type | subject) + (justification * trial type | vignette)) and then systematically removing random slopes until the model converged (Barr et al., 2013; Bates et al., 2018; Matuschek et al., 2017). (In Section 2.4 below, we also perform a finer-grained analysis of the role of moral justification using continuous justification ratings.)

2.2.2. Results

2.2.2.1. Effect of imagining harm. We regressed the reported likelihood of harming on trial type (imagine vs. control). As predicted by the schadenfreude and imagination inflation accounts, imagining performing the harmful action significantly increased people’s reported likelihood of performing the action (Fig. 1A; b = 0.3 [0.088, 0.51], p = 0.15 [0.045, 0.26], t(7.8) = 2.9, p = 0.019). Imagining performing the harm did not affect the degree to which people believed the harm was justified (b = 0.044 [−0.13, 0.21], p = 0.626 [−0.073, 0.12], t(850) = 0.51, p = 0.61).

2.2.2.2. Role of moral justification. Qualitatively, based on participants’ descriptions of their episodic simulations, people appeared to spontaneously construe the harms as morally justified on many trials. For example, one participant in the harm condition who imagined taking a seat from an elderly person noted in their description that “old people are usually rude anyways on public transportation and feel entitled.” Another participant, after imagining lying to a vegetarian house guest about whether their meal was vegetarian, wrote: “I really wish people would tell me in advance if they are coming to my house specifically to eat and they have special diet requirements. I have lost patience with this kind of thing, am not running a restaurant at home and am not going to be judged by a guest. She should have told me in advance.” (To give readers a fuller sense of these simulation descriptions, we include 25 randomly selected descriptions in the SM, and the full set of responses is available online in our ResearchBox.) Quantitatively, people did not often rate the harms as especially justified; the justification ratings had a mean of 3.1 (standard deviation 1.7) and median of 3.0. The absolute magnitude of these ratings, however, are difficult to interpret (for reasons discussed below). In Experiment 1c below, we show that an independent sample of observers rated people’s simulations as much more justified on average.

To examine the role that this moral justification played in the imagination effect, we first analyzed the effect of relative justification levels, dichotomized via a median split (i.e. comparing the half of trials which were more justified to the half of trials which were less justified). We regressed the reported likelihood of harming on trial type (imagine vs. control), relative justification (high vs. low), and their interaction. Justification moderated the effect of imagination: Imagining performing harm only increased its reported likelihood on trials when the harm was construed as relatively more justified (Fig. 1B; interaction, b = 0.4 [0.047, 0.75], β = 0.2 [0.024, 0.38], t(350) = 2.2, p = 0.025; simple effect when harm was relatively more justified, b = 0.51 [0.24, 0.77], β = 0.29 [0.14, 0.45], t(410) = 3.8, p < 0.001; simple effect when harm was relatively less justified, b = 0.12 [−0.1, 0.34], β = 0.087 [−0.077, 0.25], t(100) = 1, p = 0.3).

2.2.3. Discussion

Imagining performing harm increased people’s reported likelihood of performing it – but only when the harm was (spontaneously) construed as relatively more justified. These findings are difficult to reconcile with the harm aversion hypothesis (which predicts that imagining doing harm would decrease the reported likelihood of doing it) and the imagination inflation hypothesis (which predicts that imagining doing harm would increase its perceived likelihood, regardless of moral justification). In contrast, these findings are most consistent with the schadenfreude hypothesis, according to which simulating justified harm will make people want to do it more.

To ensure the replicability of these effects, we ran a pre-registered replication of Experiment 1a.

2.3. Experiment 1b

Experiment 1b is a pre-registered replication of Exp. 1a. The pre-registration document can be found here: https://aspredicted.org/nq277.pdf.

2.3.1. Methods

2.3.1.1. Participants. 286 participants were recruited on Amazon Mechanical Turk, using the same qualifications as above (must live in U.S., have a > 98% approval rating, and have completed >100 studies). As an additional Turk screening mechanism, we used the “Approved Participants’” mechanism from the Turk-interfacing platform CloudResearch (which only includes participants who have been thoroughly vetted; Hauser et al., 2021; Litman & Robinson, 2020). Of the participants recruited, 50 (17% of the sample) were further excluded using the pre-registered exclusion criteria (identical to Experiment 1a), leaving 236
Fig. 1. Results of Expts 1a–c. (A) Reported likelihood of performing harm on imagine vs. control trials in Expt 1a. People reported a higher likelihood of performing the harm after imagining performing it. (B) Reported likelihood of harming as a function of trial type (imagine vs. control) and relative justification levels. The effect of imagination was moderated by relative justification: People only showed an effect of imagination on the half of trials where the harm was construed as relatively more justified. (C–D) Results of a direct replication in Expt. 1b. (E) Aggregating data from Expts 1a–b, each bar indicates the standardized beta coefficient from regressing reported likelihood on trial type for each absolute level of justification (i.e. for justification rating between 1 and 2, between 2 and 3, and so on). Significance tests indicate whether each coefficient is different from zero; the red line indicates the median justification rating. (F) Histogram of observer-rated justification levels for imagine trials in Expt 1a. Error bars indicate standard error of the mean; * = p < 0.05; † = 0.05 < p < 0.10; n.s. = p > 0.10. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
for analysis. An a priori power analysis using our results from Experiment 1a found that this sample size gave us well over 80% power to detect the effects observed in Experiment 1a (see SM for details).

Note that our pre-registration document indicated a sample size of 250. We collected this data by requesting a sample of 250 participants on CloudResearch, but some participants who started but did not complete the study were not counted towards this 250. Hence, the 286 participants include people who did not complete the study and were subsequently excluded.

All methods are identical to Expt. 1a.

2.3.2. Results

2.3.2.1. Effect of imagination. As in Expt. 1a, people reported a higher likelihood of performing the harm after imagining it, compared to control trials (Fig. 1D; b = 0.22 [0.06, 0.38], β = 0.11 [0.031, 0.19], t(7.9) = 2.9, p = 0.022). Imagining harm also marginally increased the degree to which people thought the harm was justified (0.15 [0.0063, 0.29], β = 0.081 [0.00013, 0.16], t(6.4) = 2.1, p = 0.072). As this finding was only marginal and not present in Experiment 1a, we do not interpret it here.

2.3.2.2. Role of moral justification. The effect of imagining harm was again moderated by moral justification: Imagination only increased reported likelihood when the imagined harm was construed as relatively morally justified (Fig. 1E; interaction, b = 0.27 [0.053, 0.48], β = 0.14 [0.027, 0.25], t(1900) = 2.5, p = 0.014; simple effect when the harm was more justified, b = 0.3 [0.12, 0.48], β = 0.17 [0.07, 0.28], t(1100) = 3.3, p = 0.001; simple effect when the harm was less justified, b = 0.082 [−0.09, 0.27], β = 0.046 [−0.06, 0.15], t(6) = 1.95, p = 0.38).

We note one deviation from our pre-registered analysis plan. When we wrote the pre-registration document, we were only estimating subject-level random effects and not using the model-trimming procedure described in the SM. If the models are estimated using the pre-registered-random-effects structure, all results are qualitatively identical (see SM).

2.3.3. Discussion

Expt. 1b replicated the key findings of Expt. 1a: People report a higher likelihood of performing a harm after simulating it, and this effect only occurs on trials when the harm is construed as relatively more morally justified. These findings are again most consistent with the schadenfreude perspective, and are difficult to explain on either the harm aversion or imagination inflation accounts.

2.4. Examining the finer-grained effect of justification

In Expts 1a-b, we analyzed the role of moral justification by comparing the effect of imagination when this justification was above or below its median level. This approach enables a simple test of whether justification moderates the effect of imagination, but loses important information about the effect of justification across its full range. Since Experiments 1a-b were not individually well-powered to analyze the effect of justification across its full range, we combined the data from the two experiments and re-analyzed the results using continuous justification ratings.

The primary finding from this analysis is that the effect of justification appears nonlinear. Fig. 1E shows the effect of imagination (i.e. the standardized beta coefficient from regressing reported likelihood of harming on trial type) at each level of justification. Imagining harm increased the reported likelihood of harming specifically for moderately high levels of moral justification (e.g. between 4 and 6 on the 1–7 scale), but not maximally high levels (between 6 and 7). This nonlinearity was borne out statistically. We regressed reported likelihood of harming on the interaction between trial type (imagine vs. control), continuous justification ratings, and the (centered) justification ratings squared. We found a significant three-way interaction: Imagining harm had a strongest effect when justification was higher and squared (centered) justification was lower (b = −0.026 [−0.041, −0.01], β = −0.076 [−0.12, −0.03], t(2900) = −3.2, p = 0.0012) – in other words, when justification was higher but not too far from the scale midpoint. The simple effects showed a similar pattern: Imagining harm had no significant effect for justification levels between 1 and 2 (b = 0.12 [−0.014, 0.26], β = 0.12 [−0.014, 0.26], t(7.1) = 1.9, p = 0.1) or 2–3 (b = 0.046 [−0.22, 0.33], β = 0.034 [−0.16, 0.24], t(8.6) = 0.35, p = 0.74); a marginal effect for levels between 3 and 4 (b = 0.34 [0.0044, 0.68], β = 0.23 [0.003, 0.46], t(11) = 2.1, p = 0.063); a significant effect for levels between 4 and 5 (b = 0.44 [0.18, 0.7], β = 0.29 [0.12, 0.45], t(370) = 3.4, p < 0.001) and 5–6 (b = 0.27 [0.025, 0.51], β = 0.21 [0.019, 0.4], t(250) = 2.2, p = 0.031); and no effect for levels between 6 and 7 (b = −0.17 [−0.43, 0.09], β = −0.2 [−0.45, 0.059], t(190) = −1.3, p = 0.2).

This nonlinear effect was unexpected and not predicted by any of the theories. There are several potential explanations for it. It could be that the effect of moral justification is actually nonlinear: that simulating performing harm increases the reported likelihood of harming specifically for medium levels of justification (e.g. “undecided” to “justified”) but not high levels (“very justified”). It could also simply be due to a ceiling effect: The harms were rarely rated at the upper end of the justification scale (only 7% of trials were rated between 6 and 7), and nearly half of those trials were near the ceiling (~6.5 out of the 1–7 scale) for reported likelihood of performing the harm. Perhaps imagining performing maximally-justified harms would increase the reported likelihood of performing those harms, were those harms not already at the maximum level of reported likelihood. Finally, the nonlinearity could be explained by some unknown third variable which covaried with trials rated as extremely justified (e.g. perhaps people who rated the harms as highly justified took the task less seriously, or engaged in simulation differently). Future research is needed to adjudicate between these possibilities.

As the nonlinear effect was unexpected and has several possible explanations, we do not interpret it strongly here. It does not appear in our remaining experiments, where we manipulate moral justification directly. Nonetheless, it is an important result which suggests that the schadenfreude perspective does not account for the entirety of our findings.

2.5. How often did people spontaneously justify the harms, as judged by observers?

Experiments 1a-b relied on participants’ own ratings of how justified they construed the harms to be. Participants may, however, have been reluctant to report that they spontaneously found ways to justify performing canonically harmful actions, and these self-reports may be an underestimate of the extent to which people were actually justifying the harms. To further explore the extent to which people had spontaneously justified the harms, we collected an independent sample of “observer” participants. Each observer was shown participants’ simulation descriptions from the imagine trials in Expt. 1a, and asked to judge the extent to which participants in Expt. 1a believed their simulated behavior was morally justified. We averaged across these judgments to obtain an external measure of the extent to which participants had construed the imagined harms as justified.

Notably, observers could only rate the justification for imagine trials (and not control trials). Hence, we could not use their ratings to test whether justification moderated the difference between imagine and control trials. We use this data simply to examine the extent to which participants in Expt. 1a spontaneously justified the harms they were simulating, as judged by observers.
2.5.1. Methods

2.5.1.1. Participants. 328 observer participants were recruited on Amazon Mechanical Turk, using the same qualifications as in Experiments 1a-b. We did not exclude any observer participants.

2.5.1.2. Procedure. Observers were told that, in a previous study, we had asked a different group of subjects to imagine performing various actions. Each observer was presented with 18 randomly selected descriptions from the simulations in Expt. 1a, and asked: “How morally justified did the person believe their imagined behavior was?” They responded on a Likert scale from 1 (“totally unjustified”) to 7 (“totally justified”). We also asked observers to answer dichotomously: “Did the person think their behavior was morally justified?”, with options, “Yes”, “No”, or “The person did not imagine the underlined behavior”. This question produced almost identical results to the continuous rating scale, so we leave it out for the rest of our analysis. (For only a very small percentage of subjects did observers select “The person did not imagine the underlined behavior”, and excluding these subjects did not change our results; see SM for details.)

In Expt. 1a, there were 96 participants, and together they simulated performing 480 total harms. The observers produced a total of 5511 ratings, giving an average of 11 ratings per simulation episode.

2.5.1.3. Statistical analysis. To test whether the participants in Expt. 1a tended to construe the harms as justified (as judged by observers), we averaged the observers’ judgments for each simulation episode and ran a one-sample t-test comparing the mean of these episode justification ratings to 4 (the scale midpoint). We also examined the relationship between the participants’ self-ratings and the observers’ ratings by regressing the participants’ justification rating for each simulation trial on the average observer rating for that trial (using a LMM with random effects for subject and vignette, as described above).

2.5.2. Results

Fig. 1F shows the average justification ratings for the simulation episodes in Expt. 1a. On a majority of trials, what people spontaneously simulated in Experiment 1a was rated as morally justified harm by observers. 61% [57%, 65%] of the simulation episodes from Experiment 1a were rated to have an average justification rating over the scale midpoint (i.e. on the “justified” end of the scale), and the distribution mean was significantly above the midpoint (mean = 4.3 [4.1, 4.4], t (479) = 4.0, p < 0.001). Moreover, participants’ self-ratings for each simulation trial in Expt 1a were strongly related to observers’ average ratings for that trial (b = 0.43 [0.33, 0.54], β = 0.32 [0.25, 0.4], t(470) = 8.1, p < 0.001), suggesting that variation in participants’ self-ratings did capture meaningful variation in their level of moral justification, even if the absolute value of the self-ratings was lower than observers’ ratings.

2.6. Discussion of Experiments 1a-b

Together, Experiments 1a-b showed that, when asked to simulate performing harmful acts, people often spontaneously find ways to construe the harm as justified, and this justification moderates the effect of simulation: When the harms are construed as relatively more justified, simulating performing a harm increases people’s reported likelihood of actually performing the harm. (The effect of justification appeared nonlinear: Simulating harm increased its reported likelihood specifically when the harm was construed as moderately, but not maximally, justified.) These findings are difficult to reconcile with the harm aversion or imagination inflation accounts, and most consistent with the schadenfreude account.

Experiments 1a-b, however, have two conceptual weaknesses which we address in the rest of our experiments. First, the observed effect of moral justification was only correlational; we did not exogenously manipulate the moral status of people’s simulated behaviors. Thus, it is unclear whether moral justification played a causal role. Second, we did not actually measure people’s affect, and it is thus unclear whether simulating justified harm made people feel good, and whether this played a role in their likelihood reports (as predicted by the schadenfreude account).

To address these issues, in Experiments 2–5 we directly manipulate whether the harm was morally justified, explicitly asking people to construe harms as justified or unjustified. We also collect a measure of affect after each trial, as well as additional exploratory measures of simulation detail and coherence (which, based on past work (Gaesser et al., 2018), we suspected might moderate the effect of imagination). To preview our results, we find that imagining justified harm consistently increases people’s reported likelihood of performing harm; that imagining unjustified harm has a weak-to-null effect; and that the effect of imagining justified harm is consistently mediated by an improvement in affect. In other words, simulating performing harm makes people feel better, which in turn makes them report a higher likelihood of performing harm.

3. Experiment 2

In Experiment 2, we directly manipulated whether the simulated harms are construed as justified or unjustified by adding a sentence to the end of each vignette. In the justified condition, each vignette ended: “Afterward, you think the behavior was justified, and don’t feel bad about it.” In the unjustified condition, each vignette ended: “Afterward, you think the behavior was unjustified, and feel bad about it.” We predicted that imagining justified harm would feel good to people; that it would increase the reported likelihood of performing the harm; and that this increase would be mediated by the improvement in affect.

3.1. Methods

3.1.1. Participants

152 participants were recruited on Mechanical Turk using the same qualifications as Experiments 1a-b (without the CloudResearch “Approved Participants” feature, as this experiment was run before that feature was available). 35 participants (23% of the sample) were further excluded using the same exclusion criteria as in Expts 1a-b. A post-hoc power analysis revealed that this gave us over 80% power to detect our key effects of condition (see SM for details), with one exception: We had very low power to detect an interaction between trial type (imagine vs. control) and justification condition (justified vs. unjustified). Thus, although we report the statistics for this interaction, we do not focus on it in this experiment (or in Experiments 3–4, as they have similar sample sizes); we demonstrate this interaction in a well-powered, pre-registered replication study in Experiment 5 below.

3.1.2. Design and stimuli

Participants were randomly assigned to an unjustified or justified condition. They read the same vignettes as in Experiment 1, with the following addition: In the unjustified (justified) condition, vignettes ended with, “Afterward, you think the behavior was unjustified (justified), and (don’t) feel bad about it.” (Experiment 2 also had a “neutral” condition which did not include a justification instruction. This “neutral” condition was identical to Experiment 1, and produced similar results; we report it in the SM.)

The design was 2 × 2, with justification condition (justified vs. unjustified) as a between-subjects manipulation and trial type (imagine vs. control) as a within-subjects manipulation. The design of imagine and control trials was identical to Experiment 1.

3.1.3. Procedure

The procedure was identical to Experiment 1, with the following
change. After each trial, in addition to the questions from before, we asked: “How did this scenario make you feel?”, “Did you imagine yourself performing the behavior (in the way we asked)?”, “How detailed was the imagined scene in your mind?”, and “How clear and coherent was the imagined scene in your mind?” The first question was meant to gauge affect; the second question was meant to check whether people believed themselves to be following instructions (e.g. that they weren’t refusing to imagine the harms in a justified way despite our instructions); and the third and fourth questions were meant to gauge the vividness of scene imagery of people’s simulations. The latter three questions were only asked on imagine trials. All questions except the second were on Likert scales from 1 to 7; the second question was a binary yes/no.

3.1.4. Statistical analysis

All analyses were performed in the same manner as before, with mixed effects models (starting with random intercepts and maximal random slopes for both subject and vignette, and dropping random slopes if they prevented convergence).

For mediation analyses, we used the R package mediation (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014), which supports mediation analysis with mixed effects models. The package does not allow more than one random effect grouping per model (e.g. subject and vignette random effects), so for the mediation analyses we estimated the models with only subject-level random effects. (In all cases, we verified that redoing the mediation analyses with only item-level random effects did not affect the results; those results are reported in the SM.)

3.2. Results

3.2.1. Manipulation check

As a manipulation check, we tested whether our justification manipulation had a significant effect on people’s reported justification of their simulated behaviors. The mean justification ratings in the unjustified and justified conditions were 3.0 [2.8, 3.3] and 4.5 [4.2, 4.9], respectively. Regressing justification rating on justification condition, we found that this difference was significant (b = −1.4 [−1.8, −0.96], β = −0.78 [−1, −0.54], t(66) = −6.2, p < 0.001). Moreover, people reported a consistent ability to follow the instructions. In all cases, people said “yes” to the “Did you imagine yourself performing the behavior (in the way we asked?)” question on > 90% of trials, with no difference across condition (b = 0.017 [−0.074, 0.11], β = 0.071 [−0.31, 0.45], t (66) = 0.37, p = 0.71).

3.2.2. Effect of imagining harm

We regressed the reported likelihood of harming on trial type (imagine vs control), justification condition (justified vs. unjustified), and their interaction. The results are shown in Fig. 2. There was a main effect of trial type: Imagining performing a harmful act increased the reported likelihood of performing that act (b = 0.64 [0.25, 1], β = 0.34 [0.13, 0.54], t(110) = 3.2, p = 0.0015). Analyzing the simple effects, imagining increased the reported likelihood of harming in both the justified (b = 0.62 [0.23, 1], β = 0.32 [0.12, 0.53], t(35) = 3.2, p = 0.0031) and unjustified (b = 0.37 [0.035, 0.7], β = 0.2 [0.019, 0.39], t (38) = 2.2, p = 0.034) conditions. Although imagination trended towards having a larger effect in the justified condition, there was no significant interaction between justification condition and trial type (interaction, b = −0.28 [−0.78, 0.22], β = −0.15 [−0.41, 0.11], t(110) = −1.1, p = 0.27). The lack of a significant interaction is difficult to interpret as Experiment 2 was highly underpowered to detect this interaction effect; we demonstrate this interaction in a well-powered replication in Experiment 5 below.

3.2.3. Mediating role of affect

People rated how the story made them feel on a 1–7 scale from “Very Bad” to “Very Good”. Unsurprisingly, people felt better about the stories in the justified (mean rating = 3.4 [3.2, 3.7]) than unjustified (2.6 [2.5, 2.7]) conditions. More importantly, there was an interaction between justification condition and imagination. We regressed affect on trial type (imagine vs. control), justification condition (justified vs. unjustified), and their interaction, and found that imagining performing harm produced a unique improvement in affect in the justified condition (interaction, b = −0.33 [−0.66, −0.0075], β = −0.26 [−0.52, −0.0059], t (91) = −2, p = 0.048).

To test how this increase in affect in the justified condition related to the perceived likelihood of performing harm, we ran a mediation analysis with trial type as the treatment and affect as the mediator. The improvement in affect partially mediates the effect of simulating justified harm on perceived likelihood of harming (total effect = 0.67, 95% CI = [0.28, 1.1], p < 0.001; indirect effect = 0.28, 95% CI = [0.051, 0.51], p = 0.010; direct effect = 0.39, 95% CI = [0.070, 0.71]; p = 0.024).

3.2.4. Effect of detail/coherence

Detail and coherence ratings were almost identical; across subjects, the average Pearson correlation between them was 0.99. We thus averaged them to obtain a “vividness” score consistent with past work using these measures (Gaesser et al., 2018; Gaesser, DiBiase, & Kenninger, 2017; Gaesser & Schacter, 2014). (We did this throughout the rest of the experiments.) The vividness of imagination, however, was not related to reported likelihood. Restricting our analysis to imagine trials, we regressed the reported likelihood of harming on vividness, justification condition, and their interaction. There was no interaction between vividness and condition (b = 0.031 [−0.43, 0.5], β = 0.016 [−0.22, 0.26], t(360) = 0.13, p = 0.89), and no significant main effect of vividness (b = 0.057 [−0.29, 0.4], β = 0.029 [−0.15, 0.21], t(170) = 0.32, p = 0.75).

Fig. 2. Results of Experiment 2. (A) Reported likelihood of performing the harm, as a function of trial type (imagine vs. control; x axis) and justification condition (justified vs. unjustified; color). (B) Vignette-related affect across justification condition and trial type. (C) The effect of simulating justified harm is partially mediated by its corresponding improvement in affect.
3.3. Discussion

Imagining performing a harmful act again increased people’s reported likelihood of actually performing the act, and this effect was strongest when the harm was construed as morally justified (although not significantly stronger). Moreover, imagining performing justified harm uniquely improved people’s affect, and this improvement mediated the effect of imagination on reported likelihood of harming. These results are again most consistent with the schadenfreude hypothesis.

There are multiple interpretations of the observed improvement in affect after simulating justified harm. On the one hand, it is possible that simulating harm makes people feel good in a hedonic sense. On the other hand, it is possible that simulating justified harm makes people “feel good” in a more abstract sense (e.g. by satisfying their desire for justice) and yet still feels aversive in a hedonic sense. Past work on schadenfreude has used both self-reported affect measures (like the one we employ here) and physiological measures of affect like muscle and neural activation to make the case that people feel genuinely-hedonic enjoyment at others’ misfortune (Gikara & Fiske, 2012; Feather, 2006b). In our experiments, we only collect self-reports of affect, and hence our results cannot adjudicate between these interpretations of schadenfreude. We return to this issue in the General Discussion.

In this experiment, simulating unjustified harm also increased the reported likelihood of harming. This appears at odds with the results of Experiments 1 (where simulating unjustified harm did not increase its perceived likelihood), and with the schadenfreude hypothesis. But, as we show in the rest of our experiments, the apparent effect of simulating unjustified harm is not robust; simulating unjustified harm generally has a weak-to-null effect on the perceived likelihood of performing that harm, and thus we do not strongly interpret the positive result here. Moreover, simulating unjustified harm did not make people feel good; the role of positive affect was unique to justified harm.

The vividness of simulations was not significantly related to their effect on perceived likelihood. This stands in contrast to previous work on simulation and helping behavior, which often observes a robust relationship, suggesting that this association may be stronger for pro-social events (Gaesser et al., 2018; Gaesser & Schacter, 2014). This null result, however, should be interpreted cautiously, as in subsequent experiments we do find a significant relationship. We consider the role of vividness in more depth in the General Discussion.

In Experiments 3 and 4, we replicate this pattern with additional control conditions which rule out various deflationary explanations for the effect of simulating justified harm.

4. Experiment 3

Experiment 3 replicates the result from Expt. 2 with a novel control condition that rules out an important alternate interpretation. When simulating justified harm in Expt. 2, participants reported an improved affect, and, subsequently, a higher likelihood of performing the harmful behavior. We interpreted these results as a direct effect of simulating the harm. But perhaps it is not about simulating harm per se. Instead, the effect could simply be due to possessing the output or semantic knowledge generated by the simulation process. There are numerous differences between simulating an action and possessing the output of that simulation; for instance, simulating the action may be more experiential and vivid (as opposed to engaging more semantic processes), may channel participants to focus on the action itself as opposed to its outcomes (Crockett, 2013; Cushman, 2013; Cushman et al., 2012; Lockwood et al., 2020; R. M. Miller et al., 2014), and may give participants more of a sense of agency or intentionality in their cognition around the scenario. In Experiment 3, we test whether the effect on reported likelihood is directly driven by the simulation process, above and beyond possessing its semantic output.

To test this, we ran our experiment with a different control condition: On control trials, people read narratives of imagined harmful events previously generated by other participants. They thus possessed the semantic knowledge generated by the simulation process, but had not actually done the simulation themselves. If it is the actual act of simulation that increases attitudes and beliefs, then people should still report a higher likelihood on imagine trials.

4.1. Methods

4.1.1. Participants

132 participants were recruited on Mechanical Turk using the same qualifications as Experiment 2. 16 (12% of the sample) were excluded for the same reasons as in prior experiments.

4.1.2. Procedure

Everything was identical to Experiment 3, with one exception: In the control condition, instead of generating rhymes, participants read samples of what prior participants had simulated. For example, consider a participant who gets the elderly-person-on-public-transportation vignette as a control trial. They would read the scenario, and then read three responses from participants in Experiment 2 who had received the vignette as an imagine trial. (The three responses were randomly sampled each time from the set of all imagined responses to the relevant vignette in Experiment 2.)

4.2. Results

4.2.1. Manipulation check

We regressed people’s justification ratings on their justification condition, and found that the condition had a significant effect on people’s justification rating (b = −1.4 [−1.9, −1.0], p = 0.061), but not in the unjustified condition (b = 0.96 [0.48, 1.4], t(42) = −2.4, p = 0.018) and of imagine vs. control (b = 0.38 [0.066, 0.7], β = 0.19 [0.033, 0.36], t(20) = 2.4, p = 0.027). Replicating the pattern of simple effects in Experiment 3, imagining increased the reported likelihood of performing justified harm (although the effect was only marginal; b = 0.38 [0.012, 0.75], β = 0.19 [0.066, 0.37], t(9.9) = 2.1, p = 0.061), but not in the unjustified condition (b = −0.038 [−0.35, 0.27], β = −0.02 [−0.2, 0.16], t(11) = −0.23, p = 0.82). In this experiment, there was a significant interaction between trial type and condition: Imaging performing harm increased the reported likelihood of performing justified harm when the harm was construed as justified (b = −0.44 [−0.85, −0.02], p = 0.042). As Experiments 2–4 were underpowered to detect this interaction, however, we do not interpret it strongly; we demonstrate this interaction clearly in Experiment 5.

4.2.2. Effect of imagination

See Fig. 3. We regressed reported likelihood of harming on trial type (imagine vs. control), justification condition (justified vs. unjustified), and their interaction. There was also a main effect of justified vs. unjustified condition (b = −0.53 [−0.96, −0.099], p = 0.27 [0.48, −0.05], t(74) = −2.4, p = 0.018) and of imagine vs. control (b = 0.38 [0.066, 0.7], β = 0.19 [0.033, 0.36]).

4.2.3. Effect of affect

As in previous experiments, people reported feeling better about the stories in the justified (mean = 3.6 [3.4, 3.8]) versus unjustified (mean = 2.8 [2.6, 2.9]) conditions. We regressed affect on trial type, justification condition, and their interaction. Critically, there was again an interaction between justification condition and imagination: Simulating justified harm improved people’s affect more than simulating unjustified harm (interaction: b = −0.36 [−0.69, −0.03], p = 0.24 [0.46, −0.02], t(110) = −2.1, p = 0.035). Finally, this improvement in affect completely mediated the effect of simulation on reported likelihood of harming (total effect = 0.34, 95% CI = [0.041, 0.64], p = 0.034; indirect
effect = 0.24, 95% CI = [0.049, 0.42], p = 0.014; direct effect = 0.11, 95% CI = [−0.12, 0.33]; p = 0.33).

4.2.4. Effect of vividness

Restricting our analysis to simulation trials, we regressed the reported likelihood of harming on simulation vividness, justification, and their interaction. There was no interaction between vividness and justification condition (β = −0.073 [−0.4, 0.26], β = −0.036 [−0.2, 0.13], t(490) = −0.43, p = 0.66). In Expt. 3, however, there was a main effect of vividness: Across conditions, more vivid simulations of performing harm were associated with a higher reported likelihood of harming (b = 0.29 [0.066, 0.51], β = 0.14 [0.033, 0.25], t(440) = 2.5, p = 0.012).

4.3. Discussion

Experiment 3 replicated the main results from previous experiments – simulating justified harm increased perceived likelihood of harming, mediated by an improvement in affect, while simulating unjustified harm had little to no effect – with an alternate control condition. In the control condition, people read descriptions of other people’s simulations. Critically, Expt. 3 thus demonstrated that the effect of simulating harmful actions is due to the act of simulating itself; it is not due to possessing the output of simulation (i.e. an elaborated story of how the harm was performed). The act of imagining harming someone in a justified way again felt good to participants, and this made them report a higher likelihood of doing it.

In Experiment 3, we found that more vivid simulations had a stronger effect on reported likelihood; this is consistent with previous work on vividness as a key moderator in episodic simulation of helping events (Gaesser & Schacter, 2014). (In Experiment 2, this effect had only been a non-significant trend.) This effect of vividness is confirmed in a well-powered replication in Experiment 5, below.

5. Experiment 4

Experiment 4 again replicates our main result, with another control condition that clarifies the mechanisms underlying the effect. There are at least two further interpretations of our result. On one interpretation, imagining a specific harmful behavior increases positive affect and the perceived likelihood of performing that behavior. On the other, imagining any harm increases the perceived likelihood of performing any other harm – for instance, through a temporary global desensitization to harm-related stimuli. To adjudicate between these views, we asked participants to either imagine an episode involving a specific harmful action and then evaluate the affective valence and the likelihood of performing that harmful action (as in the previous two experiments), or imagine an episode involving a specific harmful action and then evaluate the affective valence the likelihood of performing a different harmful action.

5.1. Methods

5.1.1. Participants

129 participants were recruited on Mechanical Turk using the same qualifications as Experiments 2–3. 13 (10% of the sample) were excluded using the same criteria as in prior experiments.

5.1.2. Procedure

Everything was identical to Experiments 2–3, with the following exception. Instead of a within-subjects trial type manipulation which compared imagine to control trials, Experiment 4 used a between-subjects manipulation which compared “matched” imagine trials (where participants were asked to imagine performing a harmful act and then report their likelihood of performing the same act) to “mismatched” imagine trials (where participants were asked to imagine performing a harmful act and then report their likelihood of performing a different harmful act). In the “matched” condition, participants completed 5 trials which were identical to the imagine trials in Experiments 1–2: They read a story about performing a harmful behavior, spent a minute imagining themselves performing the behavior, spent 2 min recording what they imagined, re-read the story, and then answered the same series of questions as above. In the “mismatched” condition, participants completed 5 trials where they read a story about one harmful behavior (e.g. taking an elderly person’s seat on a bus), imagined the behavior, and recorded what they imagined; but then they read a story about a different behavior (e.g. lying to a vegetarian guest about the broth in the soup), and answered questions about that behavior.

5.2. Results

5.2.1. Manipulation check

As before, condition had a significant effect on the moral justification of people’s simulated behaviors (b = −1.4 [−2, −0.84], β = −0.75 [−1.1, −0.45], t(51) = −4.9, p < 0.001), and people consistently reported that they were able to follow the instructions across conditions (all means >90%, no significant difference between conditions, b = 0.041 [−0.058, 0.14], β = 0.16 [−0.27, 0.59], t(52) = 0.79, p = 0.43).

5.2.2. Effect of imagination

We regressed the reported likelihood of harming on trial type (matched vs. mismatched), justification condition (justified vs. unjustified), and their interaction (Fig. 4). There was a main effect of trial type: People reported a higher likelihood of performing a harm after simulating performing that harm (“matched” condition), compared to simulating performing a different harm (“mismatched” condition; b = 1 [0.4, 1.6], β = 0.53 [0.21, 0.86], t(110) = 3.2, p = 0.0017). This effect appeared driven by the justified condition. Imagining increased the reported likelihood of harming in the justified condition (b = 0.99 [0.084, 1.9], β = 0.52 [0.045, 1], t(19) = 2.2, p = 0.042), but there was only a non-significant trend in the unjustified condition (b = 0.41 [−0.37, 1.2],
5.2.3. Effect of affect

As in previous experiments, people reported feeling better about the stories in the justified (mean = 3.7 [3.4, 4.0]) versus unjustified (mean = 3.0 [2.8, 3.2]) conditions; there was an overall effect of condition (b = −0.65 [−1.2, −0.1], p = 0.04); direct effect = 0.19 [−0.20, 0.60], p = 0.33; indirect effect = −0.074 [−1.2, −0.25], p = 0.004. This result is expected, given the nature of Experiment 4; the effect of simulating harm on subsequent affect could not be dependent on which harm will be asked about in a future question. Consistent with this logic, affect did not mediate the effect of simulating justified harm on reported likelihood of harming in this experiment (indirect effect = 0.057 [−0.30, 0.45], p = 0.78). (Within the matched condition, however, affect did mediate the effect of justification condition on reported likelihood of harming: total effect = −0.55 [−1.1, 0.04], p = 0.084; direct effect = 0.19 [−0.20, 0.60], p = 0.33; indirect effect = −0.074 [−1.2, −0.25], p = 0.004. This result is consistent with the previous mediation results.)

5.2.4. Effect of vividness

Regressing reported likelihood of harming on condition and vividness of simulation, there was no interaction between vividness and condition (b = 0.12 [−0.41, 0.66], p = 0.061 [−0.21, 0.34], t(160) = 0.44, p = 0.66), and no significant main effect of vividness (b = 0.28 [−0.22, 0.75], p = 0.14 [−0.11, 0.38], t(50) = 1.1, p = 0.26).

5.3. Discussion

Experiment 4 replicated the main results from previous experiments, with another control condition. In the control condition, people also simulated performing harm; but they simulated performing different harms from the ones asked about in the affective valence and perceived likelihood ratings. Expt. 4 thus demonstrated that the effect of simulating harmful actions is specific to the actions being simulated; it is not due to some general cognitive change associated with performing the simulations.

6. Experiment 5

In Experiments 2–4, simulating a justified harm consistently increased positive affect and the perceived likelihood of performing that harm. However, other findings—the effect of simulating unjustified harm, the interaction between justification and imagination, and the effect of simulation vividness—were more ambiguous. To further explore these effects, we conducted a well-powered, pre-registered replication of Experiment 2.

6.1. Methods

6.1.1. Participants

We recruited 438 participants (53% female, mean age 42 years ± s.d. of 12, 80% white) on Amazon Mechanical Turk, using the same qualifications as before (including the CloudResearch “Approved Participants” feature, as this replication was run after that feature became available). 85 participants (19% of the sample) were excluded using the same criteria as before. An a priori power analysis using the results from Experiment 2 revealed that this sample size gave us over 80% power to detect the key effects—including the hypothesized interaction between trial type and justification condition on reported likelihood of harming. There was no detectable effect of age, gender, or race on any of our key effects (see SM), so we do not consider demographic factors further in our analysis.

Note that our pre-registration document indicated a sample size of 400. As in all our experiments, we collected this data by requesting a sample of the desired size (N = 400) on the Turk-interfacing platform CloudResearch—but some participants who started and did not complete the study were not counted towards this sample size. Hence, the 438 participants include people who did not complete the study and were subsequently excluded.

6.1.2. Procedure

The procedure was identical to Experiment 2. The pre-registration document can be found here: https://aspredicted.org/jq39h.pdf.

6.2. Results

6.2.1. Effect of imagination

We regressed the reported likelihood of harming on trial type (imagine vs. control), justification condition (justified vs. unjustified), and their interaction. Confirming the pattern of results in Experiments 2–4, we found a significant interaction: Imagining harm increased the reported likelihood of performing harm significantly more when the harm was construed as justified (b = −0.53 [−0.75, −0.3], β = 0.26 [−0.38, −0.15], t(21) = −4.6, p < 0.001). There was also a significant main effect of imagining harm (b = 0.71 [0.51, 0.92], β = 0.36 [0.26, 0.46], t(16) = 7.1, p < 0.001) and of condition (b = −0.65 [−0.96, −0.35], β = −0.33 [−0.48, −0.17], t(39) = −4.3, p < 0.001). The simple effects showed a similar pattern: Imagining justified harm increased the reported likelihood of harming (b = 0.71 [0.48, 0.94], β = 0.36 [0.24, 0.48], t(10) = 6.3, p < 0.001), and imagining unjustified harm had a marginal and much smaller effect (b = 0.19 [−0.018, 0.39], β = 0.098 [−0.0043, 0.2], t(9.4) = 1.8, p = 0.099).

6.2.2. Effect of affect

We regressed affect on trial type (imagine vs. control), condition (justified vs. unjustified), and their interaction. As in Experiments 2–4, the interaction was significant: Simulating justified harm produced a stronger improvement in affect than simulating unjustified harm (Fig. 5C; interaction, b = −0.58 [−0.74, −0.42], β = −0.37 [−0.48, −0.27], t(350) = −7.2, p < 0.001). This pattern was mirrored in the simple effects: Simulating justified harm improved affect (b = 0.52 [0.38, 0.67], β = 0.33 [0.24, 0.42], t(11) = 7.3, p < 0.001), while simulating unjustified harm had no detectable effect (b = −0.051 ± 0.025, β = −0.03 [−0.14, 0.06], t(110) = 0.30, p = 0.76). (Within the matched condition, however, the effect was not significant in the justified condition (b = 0.024 [0.067, 0.068], p = 0.58) while it was in the unjustified condition (b = 0.105 [0.067, 0.149], p = 0.01)).

6.3. Discussion

The key new result here is the lack of a justified–unjustified interaction in Experiment 5. This absence of a main effect of condition likely stems from the reduced sample size compared to Experiments 2–4. However, this interaction is a replication of Experiment 4's original finding. The present pattern is consistent with the interpretation that imagining harm changes affective responses in a fashion that is specific to the degree of moral justification.
confirmed and clarified the pattern of results from Experiment 2. Simulating justified harm was partially mediated by the improvement in affect; and more than the effects of simulating unjustified harm, which had no significant effects. Finally, the improvement in affect partially mediated the relationship between simulating justified harm and the reported likelihood of harming (total effect \( \beta = 0.73 \), \( p < 0.001 \)); the effect of simulating justified harm on the reported likelihood (total effect \( \beta = 0.65 \), \( p < 0.001 \)) and no detectable effect in the unjustified condition ( \( \beta = 0.25 \), \( p = 0.12 \)).

6.2.3. Effect of vividness
Restricting our analysis to imagine trials, we regressed the reported likelihood of harming on simulation vividness, condition (justified vs. unjustified), and their interaction. As suggested by the trends in Experiments 2–4, there was a significant relationship between simulation vividness and reported likelihood of harming: More vivid simulations were associated with a higher reported likelihood of harming ( \( \beta = 0.42 \) [0.28, 0.56], \( t(680) = -0.19, p = 0.049 \)).

6.3. Discussion
Experiment 5 – a pre-registered replication of Experiment 2 – confirmed and clarified the pattern of results from Experiments 2–4. Simulating justified harm improved affect and increased the reported likelihood of performing harm; these effects were significantly stronger than the effects of simulating unjustified harm, which had no significant effects; the effect of simulating justified harm on reported likelihood of harming was partially mediated by the improvement in affect; and more vivid simulations of justified harm produced a stronger change in reported likelihood. These results are again most consistent with the schadenfreude account.

7. General discussion
When people imagine performing a commonplace harmful action, they subsequently report a higher likelihood of performing the harm, and this happens in part because people often spontaneously construe the harm as morally justified (Experiment 1) and because imagining justified harm improves affect (Experiments 2–5). These results suggest that episodic simulation may serve to increase people’s inclination towards performing commonplace harms. Together these studies take an important first step in understanding how episodic simulation shapes our attitudes towards harming others.

Our findings support and expand on these perspectives in important ways. Consistent with research on schadenfreude, we find that imagining justified harm improves people’s affect, and that this improved affect contributes to the increase in reported likelihood of performing harm (Cikara, Bruneau, & Saxe, 2011). Put simply: Imagining performing mild, commonplace harmful acts makes people feel better about them, which in turn makes them believe they’d be more
likely to actually do them. Moreover, our results suggest that episodic simulation may contribute to producing “virtuous violence”; by simulating performing harms they believe to be morally justified, people may become more positive about enacting them. This result is consistent with a perspective where most harm is, from the mindset of the perpetrator, motivated for ethical reasons (Fiske & Rai, 2014).

7.1. Limitations and qualifications

Reported vs. actual behavior and affect. Our findings show an effect of episodic simulation on people’s self-reported affect and likelihood of harming, and do not speak directly to people’s actual choices and/or affect. Understanding people’s perceived likelihood of harming is an important first step in uncovering the relationship between episodic simulation and harm (Gaesser & Schacter, 2014), and people’s post-simulation reports have been shown to track their subsequent actual behavior in other contexts (D’Argebeau & Jimenez, 2020; Weiler, Suchan, & Daum, 2010). Moreover, one common concern with self-report measures – demand effects/desirability bias – is unlikely to have strongly affected our results. People’s responses in our experiments run counter to the socially desirable responses: People are more well-liked when they report being averse to harm (Everett, Pizarro, & Crockett, 2016), and yet our participants report an increase in their likelihood of performing harmful behaviors. Finally, people’s pattern of responses is difficult to explain by them simply reporting what they believed was expected of them. For instance, in Experiments 1a-b, people were not told anything about morally justifying the harms, and yet often spontaneously construed the harms as justified; and in Experiments 2-5, people in the “unjustified” condition were explicitly told to construe the harms in a way that they felt bad about, and yet simulating those harms did not produce any worsening of affect.

Nonetheless, in many contexts self-reports do not accurately predict behavior (Baumeister, Vohs, & Funder, 2007), and it is unknown the extent to which these simulation-induced effects predict objective changes in affect or harm behavior. It is possible that, in real situations, other motives (such as harm aversion) would be stronger and would overcome any simulation-related shifts in beliefs/intentions. Testing the effect of episodic simulation on actual harmful behavior is a critical area for future research.

Mild vs. extreme harms. Another major limitation of our paradigm is our use of relatively mild harms, like lying or littering. In contrast, many psychological studies – and, indeed, many important real-world decisions – revolve around more drastic and violent harms or norm violations. Examining mild harms was an important first step: Mild harms are much more common, and people are much more likely to find themselves in situations where they could perform them. This lends our approach more external validity. Moreover, people’s willingness to perform more serious harms would likely be at floor; using mild harms allows more sensitive measurement. Nonetheless, it is very possible that the effects of simulation would be different for harms with more viscerally negative consequences. Future research should adapt our paradigm to include these more extreme behaviors.

7.2. Relationship to prior work on moral simulation

7.2.1. Episodic simulation and helping behavior

Parallel to our findings on harmful behaviors, prior research found that simulating helpful behaviors increases the perceived likelihood of performing those behaviors. After imagining helping a person in need, participants reported a higher willingness to help (Gaesser et al., 2018; Gaesser et al., 2019; Gaesser, DiBiase, & Kensinger, 2017; Gaesser, Dodds, & Schacter, 2017; Gaesser, Horn, & Young, 2015; Gaesser & Schacter, 2014). This prosocial effect is driven by multiple mechanisms. As participants experience the scene imagery of the helping episode more vividly, they report that they are more likely to help. The affective signal that arises from simulating helpful events also contributes to this effect, as people are more willing to help after imagining helping episodes that elicit positive affect (Gaesser, DiBiase, & Kensinger, 2017). Our work here shows that a similar effect applies to harmful behaviors – so long as those harms are perceived as morally justified. Recent work on the prosocial imagination effect has revealed that manipulating the location (spatial context) in which the helping episode takes place increased the vividness of scene imagery and subsequently increases willingness to help (Gaesser et al., 2018). An intriguing possibility is that manipulating location of imagined harm may have a similar consequence on judgments to harm others.

In addition to scene imagery and affect, the impact of episodic simulation is also mediated by perspective-taking. In several studies, imagining helping increased perspective-taking ratings, which then increased willingness to help (Gaesser et al., 2018; Gaesser et al., 2019; Gaesser et al., 2020). These findings suggest that episodic simulation in the prosocial domain may recruit and augment considerations of the thoughts and feelings of the person in need. Indeed, there may be an important psychological link between episodic and mentalizing systems: The episodic representation and the phenomenological qualities of past and future events in which an imagined person is embedded can affect whether and how the person’s mind is experienced (see Gaesser, 2020 for review of this “episodic mindreading hypothesis”). Whether and how perspective-taking for a potential target of harm is impacted by episodic simulation currently remains unknown. Intriguingly, imagining harming others as punishment for moral transgressions led to increased humanizing of the target as a means of motivating moral sanctions (Rai et al., 2017), suggesting there may be an association. What’s more, our present results suggest that people may not only humanize targets of harm as a means of motivating that harm, but that doing so may in fact bring the person imagining the harm even more pleasure.

7.2.2. Episodic simulation and doing harm

Our results dovetail with two prior findings about episodic simulation and decisions to harm. The first comes from studies of moral decision-making in amnesic patients, a population with severely impoverished episodic simulation ability (Maguire & Hassabis, 2011). When asked whether they would harm someone for the greater good (e.g. in sacrificial moral dilemmas), amnesic patients are more averse to harming others compared to healthy controls (McCormick, Rosenthal, Miller, & Maguire, 2016). This finding suggests that episodic simulation, which healthy controls have but amnesic patients lack, decreases rather than increases harm aversion.

The second comes from Rai et al. (2017), who found that, when people imagine harming someone for morally justified reasons, they choose to humanize the target of harm. Rai and colleagues suggest that people do this because it feels good to commit morally justified harm to a deserving, humanized target. This interpretation matches our finding that imagining doing morally-justified harm improves people’s affect.

There are two prior studies on mental visualization of harmful actions that appear to conflict with our results. Caruso and Gino (2011) asked people to contemplate performing immoral actions (e.g. over-reporting the number of hours you’ve worked), with their eyes either open or closed. People judge the actions to be worse, and say they are less likely to do the actions, when their eyes are closed, and this effect is mediated by their use of more extensive visual simulations. This finding suggests that simulating immoral behaviors decreases the perceived likelihood of acting morally – which conflicts with our results.

In a related study, Kroell and Greene (2012) investigated the role of visualization in trolley-type moral dilemmas. Participants with a more visual cognitive style were more averse to harming another person for the greater good, and interfering with participants’ visual simulations reduced this aversion. This finding suggests that the capacity to simulate harmful actions (like pushing the large man in front of the trolley) is associated with an increased aversion to those actions – which, again, conflicts with our results.

This conflict may be more apparent than real because our
experiments differed from those of Caruso and Gino (2011) and Amit and Greene (2012) in several crucial ways. First, the role of simulation in our task was more transparent to participants. Unlike these previous experiments, we explicitly instructed participants to perform a mental simulation and describe what they had imagined. This ensures that people were performing episodic simulations, rather than indirectly inferring that they were (e.g. from eye closure). Moreover, we instructed people to make their simulations as detailed and episodic as possible (i.e. located in a specific time and place, with lots of sensory information), and gave them an extended time period in which they had nothing to do but perform the simulation. It is possible that simulation has an inverted-U-shaped relationship with harm aversion, where mild simulation leads to more harm aversion but extensive simulation leads to less.

Second, the content being simulated in our task was different from prior tasks. In prior work, participants presumably visualized the behavior being performed. But in our task, participants had to imagine themselves performing the harmful behavior. It is possible that imagining oneself performing harm biases towards justified harm, and is thus more likely to produce “virtuous violence” simulations.

### 7.2.3. Remembering doing harm

While we focused on episodic simulation, our findings may also extend to research on episodic memory. Given that episodic simulation relies on many of the same cognitive and neural processes as episodic memory (Schacter & Addis, 2020; Atance & O’Neill, 2001; Schacter et al., 2017; Schacter & Addis, 2007), the effects of imagined harmful events examined here may extend to remembered harmful events. The same affective and scene imagery mechanisms explored in the present studies likely extend to remembering past morally-justified harms. In fact, given that remembered events tend to be associated with greater sensory and contextual information than imagined events (Bernstsen & Bohn, 2010; D’Argembeau & Van der Linden, 2004; Gamboz et al., 2010; Gilmore, Nelson, & McDermott, 2016; McDonough & Gallo, 2010; Weiler et al., 2010), it is possible that remembering (compared to imagining) may have a stronger effect on judgments of harm as a result of differences in the strength of the sensory detail. The implication, then, is that remembering harming somebody in the past may increase the likelihood of repeating the harm in the future. However, given that individuals tend to experience emotions less intensely when they reflect on past events than when they anticipate future events (Caruso, 2010; Van Boven & Ashworth, 2007), it is an open question how closely remembering past harms will mirror our findings on imagining future harms.

### 7.3. Questions for future research

#### 7.3.1. What aspect of the harm are people feeling good about?

Although participants in our task report improved affect after simulating justified harm, it is unclear exactly which aspects of the harms they feel better about. People’s tendency to experience pleasure at other’s justified misfortunes – i.e. schadenfreude – has been linked to multiple motivations, such as correcting norm violations (Feather, 2006b), improving one’s relative social standing or personal gain (Cikara, 2015; Cikara, Botvinick, & Fiske, 2011; Smith et al., 2006), and strengthening one’s social identity against an outgroup (Cikara, 2015; Cikara et al., 2014; Cikara, Botvinick, & Fiske, 2011; Cikara & Fiske, 2012). Since our participants decided for themselves how to construe the harm as justified, we don’t know which form of schadenfreude may have applied to each simulation. An important question for future research is how exactly simulation-induced schadenfreude operates to increase people’s reported likelihood of harming.

A related question is: What is the nature of the post-simulation improvement in affect? There are different interpretations of this affective improvement. Concretely, our measure asked participants to report how the vignette “made them feel” (on a scale from ‘Very Bad’ to “Very Good”). In line with past work on schadenfreude, participants could have interpreted “feeling good” as hedonic enjoyment (Cikara & Fiske, 2012; Feather, 2006b). But they could also have interpreted it as a more abstract goodness (e.g. satiating their desire for justice), while still feeling hedonically averse to the harmful act. Past work on schadenfreude has employed both self-reported and physiological measures of affect to make the case that people feel genuinely-hedonic enjoyment at others’ misfortune (Cikara & Fiske, 2012; Feather, 2006b). In our experiments, we only collect self-reports of affect, and hence our results cannot adjudicate between these interpretations. Adjudicating between these interpretations (using, e.g., physiological measures of hedonic affect) is a critical area for future research.

#### 7.3.2. What aspect of the simulations is driving their effects: episodic scene construction or generation of semantic knowledge?

Following past work (Gaesser et al., 2018; Gaesser & Schacter, 2014; Sawczak et al., 2019; Vollberg et al., 2021), we’ve argued that the episodic component of the simulations – i.e. constructing a vivid, depictive representation of the imagined harm – is what is driving participants’ improved affect and increased reported likelihood of harming. This conclusion is bolstered by the fact that (a) more vivid simulations produced a stronger effect in Experiment 5, (b) merely reading the output of the simulations did not produce an effect in Experiment 4, and (c) excluding the small percentage of participants who did not actually imagine performing the behavior (as judged by observers) did not change our results (see Section 2.5 and SM for details). Nonetheless, when performing simulations people also likely generated semantic knowledge, and this semantic knowledge may also have played a role in increasing people’s reported likelihood of harming. An important task for future research is to tease apart the mechanisms of simulation-induced increases in the reported likelihood of performing harm.

#### 7.3.3. Does simulating performing harm induce people to justify it?

We focused on one interpretation of the schadenfreude hypothesis: that simulating harm which is construed as justified will make people feel good and increase their reported likelihood of performing harm. An additional possibility is that simulating harm induces people to justify the harm. We found inconclusive evidence on this possibility; aggregating data from Experiments 1a-b (where people did not receive instructions on how to construe the harm), participants rated the harms as a little more justified on simulation trials (average of 3.2 out of 7) than on control trials (average of 3.1 out of 7), but the difference was small and only statistically marginal (p = 0.07). Future research should more carefully examine whether simulating performing harm can make people justify that harm.

### 8. Conclusion

We examined the role of episodic simulation in people’s decisions to harm others. We found that simulating justified harm makes people feel good and increases the perceived likelihood of performing that harm in the future. This result accords with the growing consensus that causing harm is not always viewed in a negative light; moral psychologists should seek to understand when imagining harming others feels good, and how that process impacts people’s decisions to harm.

### Acknowledgements

We thank members of the Moral Psychology Research Lab for their feedback on this work, and three anonymous reviewers for their critiques and suggestions. This research was supported by Grant #61061 from the John Templeton Foundation.

### Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105104.