Journal of Experimental Social Psychology 113 (2024) 104615

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Most people do not "value the struggle": Tempted agents are judged as less virtuous than those who were never tempted☆

Ryan M. McManus[a,*], Helen Padilla Fong[a], Max Kleiman-Weiner[b], Liane Young[a]

[a] Department of Psychology and Neuroscience, Boston College, Boston, MA, USA
[b] School of Engineering and Applied Sciences, Harvard University, Boston, MA, USA

ARTICLE INFO

Keywords:
Temptation
Virtuosity
Moral judgment
Obligation
Person-level

ABSTRACT

Do people judge those who overcome temptation as more virtuous than those who don't feel tempted in the first place? Because prior research provides conflicting answers to this question, the current paper uses an expanded set of methodological and statistical tools to solve this puzzle. First, we replicated results of prior research showing that agents who overcome temptation are seen as less virtuous than non-tempted agents, with 74–78% of people making this judgment. Second, we used participant-generated stimuli and one measure from each of two published papers to rule out stimulus and measurement sampling as explanations for the previous opposite effects. We replicated our original results: 72–75% of people judged agents who overcame temptation as less virtuous than non-tempted agents. Third, we investigated whether judgments were moderated by relationship context. Again, the majority of people judged agents who overcame temptation–that would harm strangers or close others–as less virtuous than non-tempted agents. Additionally, the following interaction effect was the most common (modal) pattern: While judging tempted agents as less virtuous than non-tempted agents within each relationship context, 39% of people judged agents who were tempted to act in a way that would harm close others as even less virtuous than those agents whose temptations would harm strangers. Together, these results provide a detailed moral psychological account of temptation by: resolving a puzzle in the literature, revealing moderation by relationship context, and documenting the pervasiveness of this effect across stimuli, measures, and persons.

1. Introduction

Imagine the following scenario:

Gabriella and Katy live in different towns. While each one of them is on a walk, they find a wallet on the ground that does not belong to them. The wallets contain cash as well as the owners' ID cards. They could each take the money for themselves, or they could each find the owners and give it all back.

Gabriella feels very conflicted about this decision. She wants to steal the cash, and she is tempted to do so. However, even though she is tempted, she decides not to steal and she gives it all back to the owner.

Katy does not feel conflicted about this decision. She does not want to steal the cash, and she is not tempted to do so. She gives it all back to the owner.

Do you "value the struggle," judging Gabriella as more virtuous, as she was tempted but overcame her temptation? Or do you consider her struggle as revealing her negative moral character, therefore judging Katy as more virtuous? On some philosophical accounts, the most virtuous agents are those who are naturally virtuous–virtue comes easily (e.g., Aristotle, 1985; Taylor, 1981), i.e., non-tempted agents should be seen as more virtuous. On other philosophical accounts, agents deserve credit for working hard to do what is virtuous, and to follow moral rules even and perhaps especially when it doesn't feel effortless or natural (Halfon, 1989; Kant, 1998), i.e., agents who overcome temptation should be seen as more virtuous.

2. Literature review

Empirical work has investigated how engaging in effortful moral behavior affects these third-party moral character judgments. Some

☆ This paper has been recommended for acceptance by Rachel Barkan.

* Corresponding author.
E-mail address: ryan.m.mcmanus.phd@gmail.com (R.M. McManus).

https://doi.org/10.1016/j.jesp.2024.104615
Received 11 June 2023; Received in revised form 11 March 2024; Accepted 17 March 2024
Available online 29 March 2024

work suggests that people judge agents as more praiseworthy, or attribute a better moral character, when those agents exert effort to act morally (just as those who overcome temptation must exert effort). Bigman and Tamir (2016), for example, presented participants with two different situations in which a man is taking a bus to work. In the first situation, a woman is about to get off, and the man realizes that she dropped her wallet. He picks it up and gives it back to her. This situation is contrasted with one in which the woman gets off the bus so quickly that the man must get off the bus and run after her to return the wallet. People judged the man in the second situation as more moral and deserving of reward, effects driven by inferences of agents' differential goal importance to do the right thing. This effect was recently replicated by Berry and Lucas (2022).

Other research suggests that agents who overcome temptation are *less* virtuous than agents who never experience temptation. For example, Critcher, Inbar, and Pizarro (2013) compared a situation in which one agent quickly returned a lost wallet to a situation in which another agent returned a lost wallet after some deliberation. Contrary to other research (Berry & Lucas, 2022; Bigman & Tamir, 2016), Critcher et al. (2013) found that agents who slowly returned the wallet were judged as worse than agents who quickly returned the wallet, arguing that deliberation in this context communicates that the agent at least has a predisposition to consider immoral behavior, and they might eventually act immorally. The absence of deliberation, on the other hand, communicates that the agent may not have such a predisposition.

Additional research is more nuanced on the role of effort in moral judgment, suggesting that judgments of agents who experience temptation are sensitive to contextual features of the situation in which they experience conflict. For example, Everett, Pizarro, and Crockett (2016) found that consequentialists who display mental conflict about making a consequentialist decision are judged more positively than those who display no mental conflict. On the other hand, deontologists who display mental conflict about making a deontological decision are judged less positively than those who display no mental conflict.

Here, we focus on a puzzle in research that has directly investigated the role of temptation in moral character judgments. While young children judge agents who overcome temptation as less morally good than agents who never experience temptation, adults show the opposite pattern (Starmans & Bloom, 2016). For example, after both agents promised their parents they would clean up their toys, adults judged the agent who was tempted to go outside and play with her friends (but ultimately cleaned up her toys) as more morally good than the agent who was not tempted to do so. However, other research suggests that adults judge agents who overcome an immoral temptation as *less* virtuous than agents who never experience that temptation (Berman & Small, 2018). For example, an agent who overcame a temptation to cheat on his wife was judged as less virtuous than an agent who was not tempted to do so. Therefore, not only do philosophical accounts provide opposing answers to the question of who is more virtuous, existing empirical research also provides opposing answers. In this paper, we seek to rule out several methodological differences as a source of the discrepancy between Starmans and Bloom (2016) and Berman and Small (2018), improve the methodological and statistical rigor of studying perceptions of temptation, and extend the theoretical scope of temptation in moral psychology.

3. Rationale for current research

As discussed, extant papers make opposing claims about people's judgments of overcoming temptation. Starmans and Bloom (2016) found that adults judge agents who overcome immoral temptations as "more good" than non-tempted agents, while Berman and Small (2018) adults judged non-tempted agents as "more virtuous" than agents who overcame immoral temptations. However, their methodological approaches varied substantially.

3.1. Discrepancies in descriptions of non-tempted agents

Berman & Small describe tempted agents as having an immoral desire and non-tempted agents as simply lacking that immoral desire (e.g., one man feels a strong desire to sleep with a woman who is not his wife, whereas another man feels no desire to sleep with a woman who is not his wife). Starmans & Bloom, however, describe tempted agents as having an immoral desire and non-tempted agents as actively disliking the activity that could elicit the immoral desire in the first place, e.g., one child is tempted to break their promise to their parent to clean up their toys in order to go outside to play with their friends, whereas the other child does not like playing outside with their friends so is not tempted to break their promise to clean up. Therefore, Berman & Small's stimuli are tightly controlled, varying only the presence or absence of a specific desire and the temptation that arises. In contrast, Starmans & Bloom's stimuli may have led participants to focus on less relevant details of the scenario, leaving it unclear what explains their effect, e.g., a child who does not like playing outside with their friends is atypical and perhaps less warm or virtuous. Moreover, there are many such details that could lead to different inferences about the non-tempted agent, e.g., a child not knowing that their friends are playing outside. While future work will need to map how these action/event features shape inferences of whether and why agents are not tempted, we sidestep questions of how these stimulus details could matter, and we focus *only* on describing agents as either having or lacking an immoral desire. Therefore, in trying to resolve the discrepancy between prior research, we use stimuli modeled after Berman & Small's stimuli (including replicating their work in our Pilot Study). Importantly, however, we assume that Starmans & Bloom's effects would replicate with their stimuli.

3.2. Discrepancies in analytic procedures

Berman and Small (2018) conducted group-level analyses (i.e., comparing sample means with *t*-tests) to make inferences. As has been argued elsewhere (e.g., Richters, 2021), these analyses produce a mismatch between psychological theorizing and the methods used for inference—typical theorizing occurs at the person-level, but then researchers use analytic procedures that operate at the group-level. This has the consequence of answering questions about nebulous parameters, such as population-level means, rather than the (arguably more important) number of persons whose responses match the theorized pattern (see McManus, Young, & Sweetman, 2023, for a demonstration of how we subjected our own work to this critique). However, Starmans & Bloom's research (Starmans & Bloom, 2016) analysis (i.e., comparing the number of participants who made particular judgments) enabled some inference about prevalence. It is therefore possible that Berman & Small's effects occur because a minority of participants show very large effects, while most participants show Starmans & Bloom's effect. To assess whether prior analytical differences could explain the divergent results, we employ multiple analytic techniques, from descriptive to inferential at both the person- and group-level, to demonstrate equivalence (or not) across methods.

Specifically, we employ five different analytic techniques across studies. First, as in prior research, we conduct typical group-level analyses (e.g., *t*-tests). Second, we calculate the predicted effect's descriptive pervasiveness (i.e., the proportion of participants' responses that match prediction; see Grice et al., 2020; McManus et al., 2023; Speelman & McGann, 2020). Third, when possible, we conducted randomization tests to assess whether the predicted effect occurs in a proportion of the sample that is unlikely to have occurred via repeated random shuffling of the data (i.e., to rule out physical chance, see Grice, 2021). Fourth, we conducted frequentist prevalence testing to assess whether the predicted effect occurs in a proportion of the sample equal to or greater than a theoretical value of interest, allowing a null hypothesis significance testing inference about population prevalence (see Allefeld, Görgen, & Haynes, 2016; Donhauser, Florin, & Baillet, 2018). Here, we test against

a majority null (i.e., > 0.50, that at least half of persons in the population are likely to show our predicted effects), as recent evidence suggests that laypeople and researchers believe this to be the bare minimum for establishing evidence in favor of a psychological theory (McManus et al., 2023). Fifth and finally, we conducted Bayesian prevalence estimation to obtain a posterior distribution, which not only provides information about the most likely population prevalence value, but also enables calculation of probabilities that the population prevalence value is equal to or greater than a theoretical value of interest (e.g., 0.50, see Ince, Kay, & Schyns, 2022; Ince, Paton, Kay, & Schyns, 2021).[1] Here, we compute probabilities for the majority null ($Pr(\gamma > 0.50)$, and the global null ($Pr(\gamma > 0)$, i.e., that the predicted effect occurs in at least *some* subset of the population). These prevalence methods allow a sample-to-population inference, whereas descriptive pervasiveness and randomization tests do not.

3.3. Discrepancies in sampling of stimuli and measures

In both the papers by Berman and Small (2018) and Starmans and Bloom (2016), an additional concern is differences in stimulus and measurement sampling. As suggested elsewhere (e.g., Yarkoni, 2020), researchers usually intend to generalize over not just participants, but also over instruction sets, measures, and stimuli. In many judgment paradigms, researchers generate (sometimes only one or a few) stimuli and measures, which can sometimes lead to effects for *only* those stimuli or measures. To address this issue in the current paper, rather than relying on published stimulus sets from either of these publications, we ask an independent sample of participants to generate scenarios in which people are tempted but overcome immoral desires, leading to a larger, more diverse, less biased, and potentially more generalizable stimulus set than those in prior research. We then visually represent stimulus variability to catalog the generality (or lack thereof) of effects across stimuli. Measurement-wise, Starmans & Bloom asked participants to judge which agent was "more good" on a binary scale, which prevents judgments that agents are equally good, but Berman & Small asked participants to judge which agent was "more virtuous" on a continuous bipolar scale with a midpoint that allows a judgment that the agents are equally virtuous. Therefore, in Study 1, we use one measure from each paper to test whether differences in measurement explain the discrepant results.

3.4. Extending the theoretical scope of temptation judgments

Finally, beyond evaluating discrepancies of prior work, and in aiming to provide a more theoretically complete picture of the role of temptation in moral psychology, we connect these questions to a burgeoning area of research suggesting that everyday social relationship information moderates third-party moral judgment. Specifically, people believe that they have obligations to close others such as friends or family that they do not have to distant others such as strangers or acquaintances (Marshall et al., 2022; McManus, Kleiman-Weiner, & Young, 2020; McManus, Mason, & Young, 2021). Moreover, these beliefs have downstream consequences on moral judgment, where the violation of relationship-oriented obligations leads to more negative moral judgments (e.g., Everett et al., 2016; Law, Campbell, & Gaesser, 2021). Therefore, consistent with the direction of Berman & Small's effects (Berman & Small, 2018), when agents experience a temptation that could result in harm to a stranger, they might be judged as immoral when compared to non-tempted agents, but not nearly as immoral as agents whose temptation could harm close others. In contrast, consistent with the direction of Starmans & Bloom's effects (Starmans & Bloom, 2016), agents whose temptation could harm close others may be judged as especially moral for overcoming an impulse that would violate their relationship-oriented obligation. Prior research on effort and temptation has not explicitly investigated the role of social relationships; Study 2 addresses these possibilities.

4. Open practices

All studies in this paper were pre-registered via AsPredicted: https://aspredicted.org/M94_2Q4, https://aspredicted.org/S5X_DZL, https://aspredicted.org/KZF_1G9, and https://aspredicted.org/3B4_X4N. All materials, data, code, and analysis output are provided at https://osf.io/bym4t/?view_only=ab62a6698bb84e34a53892e39576623f. This includes stimuli and measures, raw data, organized and commented RMarkdowns, and RNotebook.html files with all visualizations and analysis outputs that can be compared to the results reported here. In these studies, we report all measures, manipulations, and exclusions.

5. Pilot study

Our Pilot Study had three purposes. First, we attempted to replicate Berman and Small (2018) finding that agents who overcome immoral temptation are judged as less virtuous than non-tempted agents. Second, we used all five of the previously described analytical techniques to gain precision in our understanding of the psychology claimed in Berman and Small (2018), allowing us to rule out differences in level of analysis as the source of discrepancy in prior findings. Third, we conducted two replications, each one with a different crowdsourcing platform (i.e., CloudResearch and Prolific), to ensure generalizability across populations (Pilot Study A and Pilot Study B, respectively).

For brevity, we relegate these studies to the SOM as they simply replicate Berman and Small (2018) work and suggest that most people judged agents who overcame immoral temptations as less virtuous than non-tempted agents. Because we used both person- and group-level analytic methods, we are able to rule out differences in level of analysis as a source of discrepancy between Berman and Small (2018) and Starmans and Bloom (2016). However, these findings are still at odds with results from Starmans & Bloom (2016, i.e., overcoming immoral temptation is more virtuous than never being immorally tempted). Study 1 therefore aimed to resolve this discrepancy.

6. Study 1

Beyond differences in level of analysis, an additional explanation for the discrepancy in prior research is that different stimuli lead to different psychological effects (see Yarkoni, 2020). To generate a wide range of stimuli to address this, we conducted a pre-experiment study to create a new stimulus set. Because of our consistent replications of Berman and Small (2018) effect (see Pilot Study), we asked participants to generate scenarios that would favor Starmans and Bloom (2016) findings. Specifically, participants were asked to generate scenarios in which they believed someone overcoming an immoral temptation would be *more* virtuous than someone who never had the immoral temptation in the first place. We used responses to create 10 stimulus bases for Study 1 (see OSF for raw data). These stimuli were created as follows: First, a co-author documented which participants followed instructions. Second, they created a set of stimuli, structurally modeled after Berman and Small (2018) that, among the available responses, maximized variability in content and severity. Finally, all authors of the current paper provided feedback until final versions were agreed upon (see SOM). Therefore, in these new stimuli, if most people still judge overcoming an immoral temptation as less virtuous than never being tempted, this provides strong evidence in favor of Berman & Small's claim, while providing

[1] These analyses attempt to control the false-positive rate at the person-level (i.e., through using separate typical group-level tests on each person's data). Therefore, because we only had low-trial data (i.e., no >10 trials per person per condition), we could not conduct these tests on each person's data, so we must assume person-level patterns are false-positive-controlled at the nominal 0.05 level. See Footnote 4 for more info.

equally strong evidence against Starmans & Bloom's claim (assuming that each paradigm makes similar assumptions about why non-tempted agents are not tempted).[2]

Similar to the above stimulus sampling issue, different dependent measures can also lead to different findings. In Berman and Small (2018), participants assessed the relative "virtuosity" of two agents on a 9-point Likert scale. Starmans and Bloom (2016), on the other hand, had participants assess which of the two agents was "more good" on a binary scale. To address this, rather than choosing our own measure, or choosing one of the measures from previous work, we randomly assign participants to one of the measures from previous work. Specifically, in Study 1, half of our participants respond to the new stimuli with Berman and Small (2018) continuous virtuosity measure, whereas the other half of participants respond using Starmans and Bloom (2016) binary goodness measure. If the same psychological effect emerges for both measures (i.e., most people judge overcoming an immoral temptation as worse than non-temptation), this allows an inference of generalization across measures while simultaneously ruling out measurement differences as an explanation for previous research's discrepancies.

6.1. Method

6.1.1. Participants

One U.S. sample ($N = 361$) was recruited via CloudResearch's approved participants list and compensated through Amazon's Mechanical Turk. Importantly, participants who completed our Pilot Study, or our stimulus generation study, were unable to participate in Study 1. As stated in the pre-registration, participants were excluded if they failed a pre-task attention check that was disguised as a task-relevant stimulus, resulting in a final $N = 350$ (Gender: 191 males, 165 females, 2 non-binary, 1 preferred not to disclose; Age: $M = 40.16$, $SD = 11.43$).

6.1.2. Design and procedure

Study 1 used a 2 (Temptation: Non-Tempted vs Tempted) x 2 (Measure Type: Continuous Virtuosity vs Binary Goodness) design in which Temptation was manipulated within-subjects, whereas Measure Type was manipulated between-subjects. Participants were presented with 10 vignettes in total. These vignettes were generated by a pre-experiment task that specified the temptation should be immoral. Therefore, vignettes included in Study 1 were deemed "moral" not by an (additional) independent set of participants' ratings of moral relevance (as was the case in Berman & Small, 2018), but instead by instructing the stimulus-generating participants to create situations that they believed contained immoral temptations. We took this strategy considering that different people will moralize different situations, and therefore not all people will agree on the moral relevance of any one situation. The structure of each vignette mirrored the current paper's opening scenario (and Berman & Small's stimuli), with participants being introduced to two agents encountering the same situation (one tempted and one not) who ultimately behave identically.

After reading each vignette, participants made judgments about virtuosity or goodness. Like Berman & Small's methods (Berman & Small, 2018), virtuosity judgments were made on a 9-point relative bipolar scale (e.g., 1 = Katy is much more virtuous; 5 = Equally virtuous; 9 = Gabriella is much more virtuous). In line with Starmans and Bloom (2016) methods, the other half of participants made binary judgments indicating which of the two agents was "more good."

6.1.3. Statistical power

Each final analyzable dataset (Binary Goodness $N = 177$, Continuous Virtuosity $N = 177$) yielded >80% power to detect $d = 0.20$ for one-tailed one-sample *t*-tests (Faul, Erdfelder, Lang, & Buchner, 2007).[3]

6.1.4. Hypotheses

Per our pre-registration (https://aspredicted.org/KZF_1G9), we had one simple hypothesis:

People will judge agents who overcome immoral temptation as *less* virtuous than agents who are never tempted in the first place.

To prepare data for hypothesis-relevant analyses, we averaged across each participant's multiple vignettes. That is, to get a participant-level value for virtuosity or goodness, we averaged across a participant's 10 judgments.

6.2. Results

6.2.1. Typical group-level tests

Binary Goodness. For binary judgments, the null value tested against was 0.50, as this is the value that would indicate that any one participant chose the non-tempted and tempted agent at similar rates across vignettes. Therefore, a value larger than 0.50 indicates that participants chose the non-tempted agent more often than the tempted agent. Non-tempted agents were judged as significantly "more good" than agents who overcame temptation ($M_{Diff} = 0.72$, $SD_{Diff} = 0.28$), $t(176) = 10.47$, $p < .001$, $d = 0.79$ [0.48, 1.09].

Continuous Virtuosity. For continuous judgments, the null value tested against was 0. Non-tempted agents were again judged as significantly more virtuous than agents who overcame temptation ($M_{Diff} = 0.92$, $SD_{Diff} = 1.22$), $t(172) = 9.94$, $p < .001$, $d = 0.76$ [0.44, 1.07].

6.2.2. Descriptive pervasiveness

Binary Goodness. When making binary judgments, 72% of participants judged non-tempted agents as more virtuous than agents who overcame temptation.

Continuous Virtuosity. When making continuous judgments, 75% of participants judged non-tempted agents as more virtuous than agents who overcame temptation.

6.2.3. Frequentist prevalence tests

Binary Goodness. When making binary judgments, a majority of participants (72%) judged non-tempted agents as more virtuous than agents who overcame temptation, $p < .001$.

Continuous Virtuosity. When making continuous judgments, a majority of participants (75%) judged non-tempted agents as more virtuous than agents who overcame temptation, $p < .001$.

6.2.4. Bayesian prevalence estimation

Binary Goodness. For binary judgments, the most likely population prevalence value is estimated as 71% [96% HPDIs = 63% - 78%]. Using the posterior distribution, we get the following probability for the majority null: $\Pr(\gamma > 0.50) = 1.00$. Therefore, we also get a similarly large probability for the global null: $\Pr(\gamma > 0) = 1.00$.

Continuous Virtuosity. For continuous judgments, the most likely population prevalence value is estimated as 74% [96% HPDIs = 66% -

[2] Although such results would provide strong evidence against Starmans & Bloom's (2016) results being typical across a variety of stimuli, it would not rule out the possibility that their results can occur for a minority of stimuli. Indeed, in their general discussion, they clearly communicate that their results provide evidence that their claimed effects "can" occur. That is, interpretation of their data seems to be more aligned with their providing an existence proof than a general regularity.

[3] We pre-registered sample sizes based on a person-level analysis that we no longer believe is appropriate. For context, while conducting the current research, two of the four authors had a methods and statistics paper under review (McManus et al., 2023) that, over multiple revisions and resubmissions, ultimately led to a reformulation of what ought to be considered appropriate for person-level analyses. Therefore, some of our pre-registrations' original justifications for sample sizes are no longer relevant.

80%]. Using the posterior distribution, we get the following probability for the majority null: $Pr(\gamma > 0.50) = 1.00$. Therefore, we also get a similarly large probability for the global null: $Pr(\gamma > 0) = 1.00$.

6.3. Interim discussion

Results of Study 1 suggest that, across new stimuli and two measures, most people judged overcoming an immoral temptation as less virtuous than never being tempted, suggesting that measurement sampling is an unlikely explanation for previous discrepant results. Additionally, our results suggest that Starmans and Bloom (2016) stimuli may be outliers in the stimulus sampling space, and that perhaps our observed effect is one that will be most typical across new stimuli. We consider this to be especially likely due to 1) our use of participant-generated stimuli that were intended to yield results consistent with those in Starmans and Bloom (2016), and 2) not a single one of these stimuli showed an effect consistent with theirs (see Figs. 1–2). Moreover, another potential discrepancy can be ruled out given our use of stimuli that vary in content. Specifically, Starmans & Bloom's stimuli might be considered low-stakes (e.g., breaking a promise to clean up toys), whereas Berman & Small's stimuli might be considered higher-stakes (e.g., cheating on a spouse), which could explain the opposing effects. However, our stimuli varied in severity (e.g., cheating on a spouse vs criticizing someone's outfit), and, across these stimuli, we still found the effect predicted by Berman and Small (2018).

7. Study 2

Study 2 had the goal of determining whether an often-overlooked factor in moral psychology, namely social relationship information, affects how people tend to judge overcoming immoral temptations. Perhaps people are especially unlikely to "value the struggle" when it is their close others who must do so, as this suggests an impulse to violate special obligations to close others (Marshall et al., 2022; Marshall, Wynn, & Bloom, 2020; McManus et al., 2020; McManus et al., 2021). If this is true, then people may also be especially harsh in their moral judgments when others are overcoming temptations to harm close others—an effect that would be consistent with findings from McManus et al. (2020), in which agents who fail to help family are judged as more immoral than agents who fail to help strangers.

We again conducted a pre-experiment study to create a new stimulus set, asking participants to generate scenarios that would favor Starmans and Bloom (2016) findings. This time, however, we specifically asked participants to generate scenarios in which temptations could harm close others and strangers to ensure that if we found differences as a function of relationship context, the differences could not be attributed to fundamental differences in content between stimuli. For example, being tempted to cheat on a spouse, by its nature, can only affect close others (not strangers); therefore, we did not use this kind of stimulus in Study 2. Along with a subset of Study 1's stimuli, we used this new set of participant-generated scenarios to create 20 stimulus bases that could be manipulated to be about immoral temptations affecting strangers or close others (i.e., close friends and siblings). Measurement-wise, after demonstrating generalization across measures (Study 1), we opted to use only Berman & Small's continuous virtuosity measure for Study 2. We chose this measure because it allows participants to make a judgment that tempted and non-tempted agents are similarly virtuous, whereas Starmans & Bloom's binary goodness measure forces participants to choose one agent over the other.

7.1. Method

7.1.1. Participants

One U.S. sample ($N = 300$) was recruited via CloudResearch's approved participants list and compensated through Amazon's Mechanical Turk. Importantly, participants who completed any of the previous studies were unable to participate in Study 2. As stated in the pre-registration, participants were excluded if they failed a pre-task

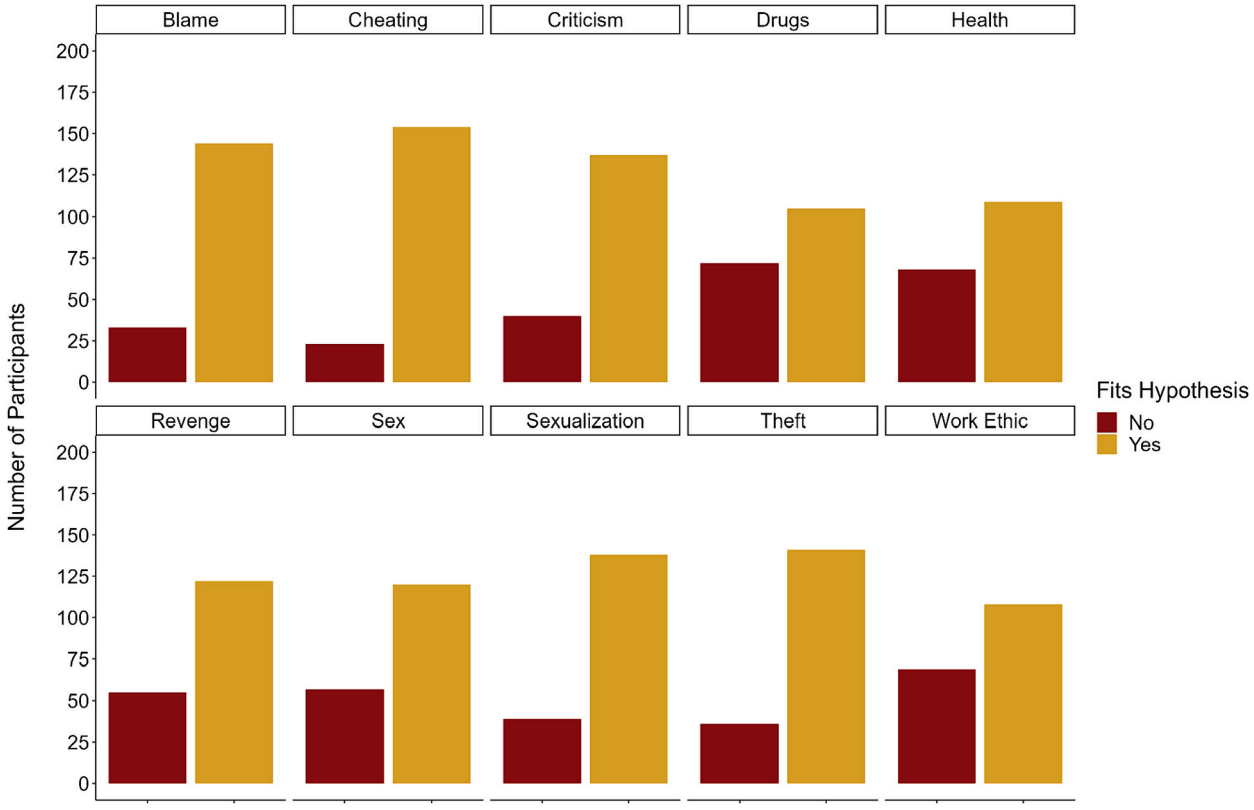

Fig. 1. Study 1 participants whose responses matched hypothesized pattern for binary judgments (by vignette).

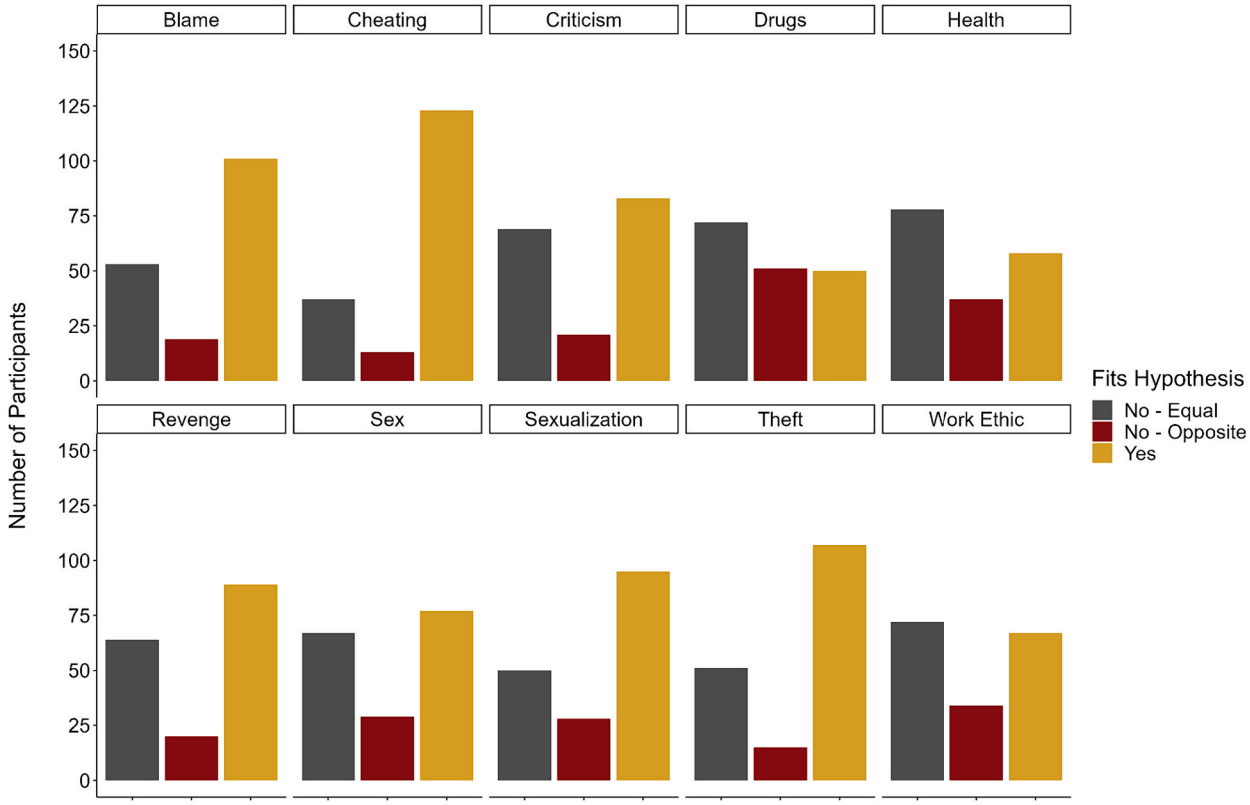

Fig. 2. Study 1 participants whose responses matched hypothesized pattern for continuous judgments (by vignette). Although some stimuli show non-majority effects, it was never the case that there was a majority effect in the opposite direction. That is, each stimulus' modal pattern was either the predicted pattern or no difference between non-tempted and tempted agents (i.e., "equal").

attention check that was disguised as a task-relevant stimulus, resulting in a final $N = 225$ (Gender: 119 males, 104 females, 2 preferred not to disclose; Age: $M = 40.12$, $SD = 11.24$).

7.1.2. Design and procedure

Study 2 used a 2 (Temptation: Non-Tempted vs Tempted) x 2 (Relationship Context: Stranger vs Close Other) design in which both factors were manipulated within-subjects. Here, the Relationship Context factor varied whether giving in to the temptation would have negative consequences for the agents' close friends or relatives versus complete strangers (see Table 1 for an example stimulus with all experimental variants). Unlike previous studies, participants in Study 2 were presented with many more vignettes (20 in total). These vignettes were generated in an identical way to Study 1. As in previous studies, the structure of each vignette mirrored the current paper's opening scenario (and Berman & Small's stimuli), with participants being introduced to two agents who encounter the same situation and ultimately behave identically. Importantly, participants did not see the same vignette across the levels of Relationship Context; specifically, half of participants saw 10 vignettes for the Stranger level and the other 10 vignettes for the Close Other level, whereas the other half of participants saw the opposite list. Which vignette was assigned to which list was accomplished by using a random number generator before creating the study's Qualtrics survey. After reading each vignette, all participants made single judgments of virtuosity on a single 9-point relative scale.

7.1.3. Statistical power

The final analyzable dataset ($N = 225$) yielded >90% power to detect $d = 0.20$ for one-tailed one-sample t-tests, as well as >90% power to detect $dz = 0.20$ for one-tailed paired-samples t-tests (Faul et al., 2007).

7.1.4. Hypotheses

Per our pre-registration (https://aspredicted.org/3B4_X4N), we had two simple hypotheses and one complex hypothesis:

1) For Stranger vignettes, people will judge agents who overcome immoral temptation as less virtuous than agents who are never tempted in the first place.
2) For Close Other vignettes, people will judge agents who overcome immoral temptation as less virtuous than agents who are never tempted in the first place.
3) People's virtuosity judgments for Stranger vignettes will be less extreme than their judgments for Close Other vignettes. They will judge being tempted to harm close others as even less virtuous than being tempted to harm strangers (i.e., the interaction hypothesis).

To calculate a participant-level value for virtuosity in Stranger / Close Other vignettes, we averaged across a participant's 10 Stranger / Close Other judgments.

7.2. Results

7.2.1. Typical group-level tests

Stranger Vignettes. Non-tempted agents were judged as significantly more virtuous than agents who overcame temptation ($M_{Diff} = 0.98$, $SD_{Diff} = 1.20$), $t(224) = 12.18$, $p < .001$, $d = 0.81$ [0.54, 1.09].

Close Other Vignettes. Non-tempted agents were judged as significantly more virtuous than agents who overcame temptation ($M_{Diff} = 1.08$, $SD_{Diff} = 1.15$), $t(224) = 13.98$, $p < .001$, $d = 0.93$ [0.66, 1.21].

Interaction. Virtuosity judgments were significantly different when comparing Stranger vignettes to Close Other vignettes ($M_{Diff} = 0.10$, $SD_{Diff} = 0.65$), $t(224) = 2.33$, $p = .010$, $d_z = 0.16$ [0.02, 0.29], $d_{av} = 0.09$ [0.01, 0.16], suggesting that non-tempted agents were judged more

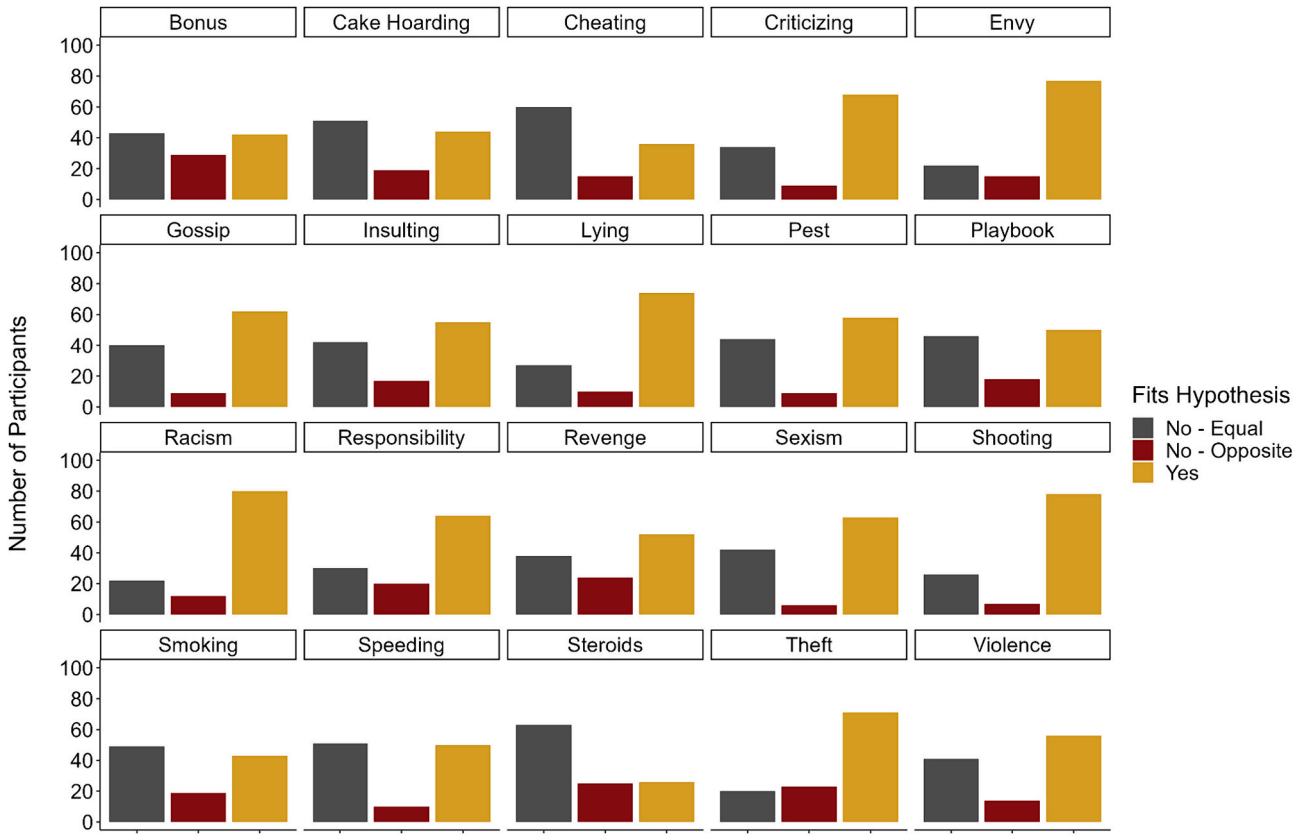

Fig. 3. Study 2 participants whose responses matched the hypothesized pattern for Stranger vignettes (by vignette). Although some stimuli show non-majority effects, it was never the case that there was a majority effect in the opposite direction. That is, each stimulus' modal pattern was either the predicted pattern or no difference between non-tempted and tempted agents (i.e., "equal").

positively in Close Other vignettes. Considering our use of relative virtuosity scales, this confirms our interaction hypothesis: While non-tempted agents were judged as more virtuous than agents who overcame temptation across relationship contexts, being tempted to engage in a behavior that could harm close others was judged as even less virtuous than being tempted to engage in a behavior that could harm strangers.

7.2.2. Descriptive pervasiveness

Stranger Vignettes. 78% of participants judged non-tempted agents as more virtuous than agents who overcame temptation.

Close Other Vignettes. 81% of participants judged non-tempted agents as more virtuous than agents who overcame temptation.

Interaction. Only 39% of participants simultaneously judged non-tempted agents as more virtuous than tempted agents within both Stranger and Close Other vignettes while also having more extreme judgments for Close Other vignettes.

7.2.3. Randomization test

For Study 3's interaction, 1000 random shufflings led to 0(!) datasets yielding a descriptive pervasiveness percentage equal to or greater than the observed descriptive pervasiveness (c-value = 0), suggesting that the original 39% estimate is extremely *un*likely to have occurred via physical chance. This means that the observed descriptive pervasiveness percentage is distinguishable from the possibility that participants were randomly selecting virtuosity values in each condition.

7.2.4. Frequentist prevalence tests

Stranger Vignettes. A majority of participants (78%) judged non-tempted agents as more virtuous than agents who overcame temptation, $p < .001$.

Close Other Vignettes. A majority of participants (81%) judged non-tempted agents as more virtuous than agents who overcame temptation, $p < .001$.

Interaction. A minority (only 39%) of participants simultaneously judged non-tempted agents as more virtuous than tempted agents across Stranger and Close Other vignettes while also having more extreme judgments for Close Other vignettes, $p = .999$.

7.2.5. Bayesian prevalence estimation

Stranger Vignettes. The most likely population prevalence value is estimated as 77% [96% HPDIs = 70% - 82%]. Using the posterior distribution, we get the following probability for the majority null: $Pr(\gamma > 0.50) = 1.00$. Therefore, we also get a similarly large probability for the global null: $Pr(\gamma > 0) = 1.00$.

Close Other Vignettes. The most likely population prevalence value is estimated as 80% [96% HPDIs = 74% - 86%]. Using the posterior distribution, we get the following probability for the majority null: $Pr(\gamma > 0.50) = 1.00$. Therefore, we also get a similarly large probability for the global null: $Pr(\gamma > 0) = 1.00$.

Interaction. The most likely population prevalence value for the predicted interaction pattern is estimated as 35% [96% HPDIs = 29% - 43%]. Using the posterior distribution, we get the following probability for the majority null: $Pr(\gamma > 0.50) < .001$. However, we get a much larger probability for the global null: $Pr(\gamma > 0) = 1.00$.

8. General discussion

The current research aimed to replicate and extend previous research on how perceived temptation shapes third-party moral character judgment. Across studies, people were asked to evaluate two agents simultaneously, one who was tempted but overcame the temptation, and one

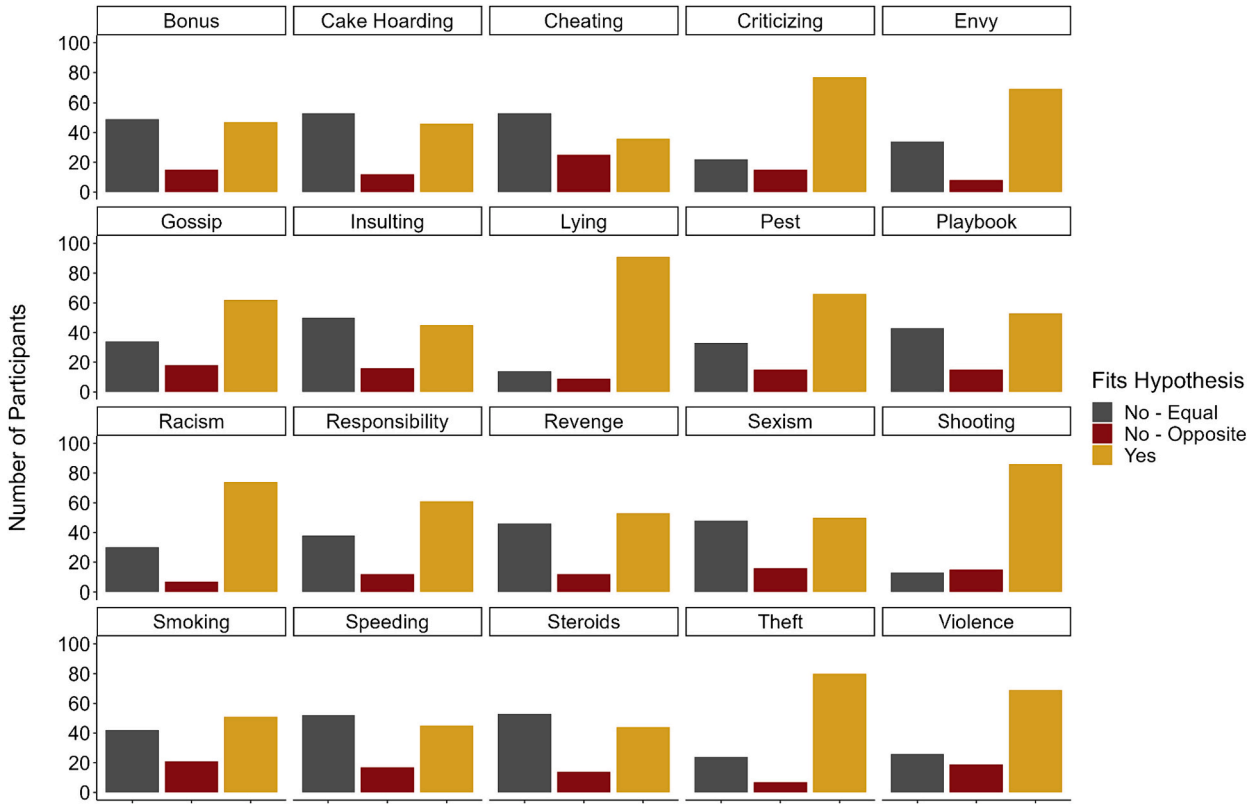

Fig. 4. Study 2 participants whose responses matched the hypothesized pattern for Close Other vignettes (by vignette). Although some stimuli show non-majority effects, it was never the case that there was a majority effect in the opposite direction. That is, each stimulus' modal pattern was either the predicted pattern or no difference between non-tempted and tempted agents (i.e., "equal").

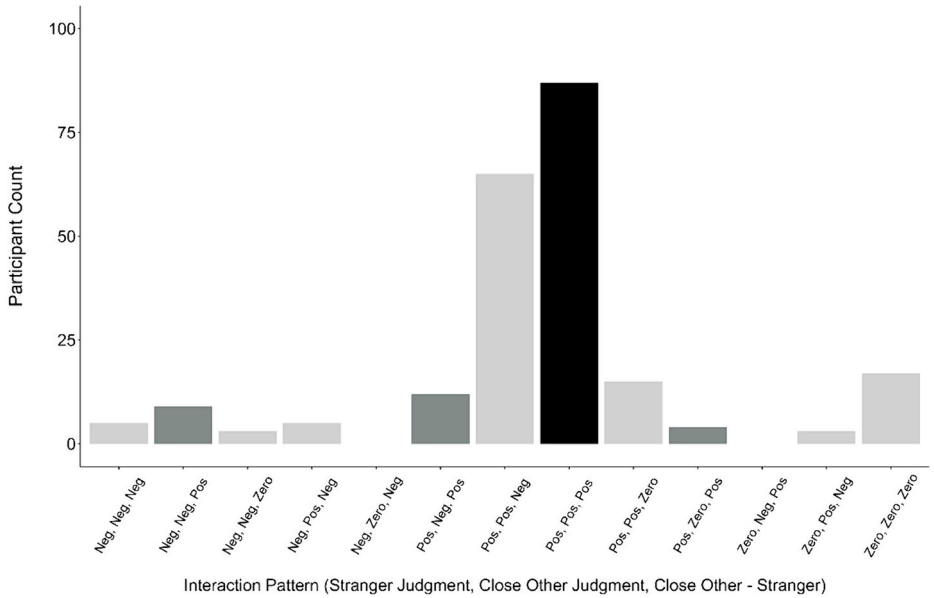

Fig. 5. Possible and empirical interaction patterns in Study 2. Pattern descriptions (e.g., Pos, Neg, Pos) communicate the Stranger judgment, Close Other judgment, and Interaction (Close Other – Stranger), respectively. The black bar represents the hypothesized pattern, whereas dark grey bars represent patterns also yielding an interaction value consistent with the hypothesized pattern (i.e., contributed to the group-level interaction's emergence).

who was never tempted. Specifically, our Pilot Study attempted to replicate recent research showing that agents who overcome immoral temptations are judged as less virtuous than non-tempted agents (Berman & Small, 2018), ruling out level-of-analysis differences as a reason for opposite-signed effects in prior research (Starmans & Bloom, 2016). Study 1 determined that stimulus and measurement sampling discrepancies likely could not account for the opposite-signed effects. Finally, Study 2 expanded the theoretical scope of temptation in moral psychology by examining the moderating effect of social relationship information, showing that moral judgments change as a function of agents

Table 1
Example stimulus ("Lying") and its experimental variants for Study 3.

Stimulus Base	Sara and Chloe are flying on different planes. They are both physicians. One of the passengers on each plane begins to pass out. The flight attendant asks if anybody on board has medical experience to aid the passenger in distress. Sara and Chloe each have the option to lie to avoid the commitment, or they can assist the passenger in need.
Tempted Agent	**Sara** and the passenger are (strangers / close friends). Sara feels very conflicted about this decision. She wants to lie to avoid the commitment of helping the passenger, and she is tempted to do so. However, even though she is tempted, she decides to assist the passenger.
Non-Tempted Agent	**Chloe** and the passenger are (strangers / close friends). Chloe does not feel conflicted about this decision. She does not want to lie to avoid the commitment of helping the passenger, and she is not tempted to do so. She decides to assist the passenger.

Note: All stimuli were presented with the stimulus base, a tempted agent, and a non-tempted agent. Terms inside of parentheses (e.g., strangers / close friends) varied across Stranger and Close Other vignettes (respectively). Therefore, participants always saw the same stimulus when comparing non-tempted to tempted agents, but they saw different stimuli for Stranger versus Close Other vignettes.

being tempted to harm strangers versus close others (i.e., close friends or siblings). See Table 2 for a summary of analyses across studies.

The Pilot Study successfully replicated Berman and Small (2018) finding across group-level and person-level analyses, ruling out level-of-analysis differences as a source of discrepancy between their work and prior work (Starmans & Bloom, 2016): Agents who overcame immoral temptations were judged as less virtuous than agents who never experienced temptation. Indeed, 74–78% of people made this judgment (see SOM). In Study 1, we considered two additional explanations for this discrepancy: stimulus and measurement sampling. Across a new set of participant-generated stimuli, and across two measures, people again judged agents who overcame immoral temptations as less virtuous than non-tempted agents. Indeed, 72–75% of people made this judgment, suggesting it is a general psychological regularity. When investigating stimulus-level variability, there was never a stimulus for which most people made judgments in the opposite direction (and there was never a stimulus in which the opposite pattern was even the modal pattern; see Figs. 1–2). Altogether, Study 1's results are inconsistent with the pattern of results obtained in Starmans and Bloom (2016), suggesting that their stimuli may be outliers in the stimulus sampling space, perhaps because they confound non-temptation with an active dislike of the activity that could give rise to temptation, or that their stimuli contain important but hidden moderators.

In Study 2, we attempted to extend these findings by relying on a recently burgeoning subset of moral psychology research: the impact of close relationships (e.g., Everett et al., 2016; Law et al., 2021; Marshall et al., 2020; Marshall et al., 2022; McManus et al., 2020; McManus et al., 2021). We hypothesized that people's judgments should be moderated when considering whether an agent's giving in to a temptation would negatively affect a stranger versus a close other (i.e., close friend or sibling). When investigating within-relationship judgments, we again found, across group-level and person-level analyses, that people judge overcoming immoral temptations as less virtuous than never being tempted. Indeed, 78% of people made this judgment for situations involving strangers, and 81% of people made this judgment for situations involving close others. Again, there was never a stimulus in which most people made judgments in the opposite direction (and there again was never a stimulus in which the opposite pattern was even the modal pattern; see Figs. 3–4). Moreover, Study 2 used stimuli that varied more in content than Study 1, allowing a stronger test of the possibility that differences in stimulus severity might explain discrepancies between previous research. For example, Study 2 used extreme scenarios such as being tempted to shoot someone (likely killing them) and mundane scenarios such as being tempted to take multiple slices of cake at a wedding (such that other guests might not get cake). Across these stimuli, Starmans and Bloom (2016) pattern never emerged as even the modal pattern.

When considering Study 2's interaction hypothesis (i.e., in addition to judging agents who overcame temptation as less virtuous within each relationship context, judgments would be more extreme for temptations to harm close others), group-level analyses showed the predicted effect. However, person-level analyses showed that only 39% of people indeed made such judgments (see Fig. 5). Importantly, though, a randomization test suggested that this percentage was distinguishable from random responding, providing evidence in favor of the robustness of the pattern. While this means that it may not be a general psychological regularity in the majority sense (capturing only 39% of participants' responses), it is a psychological regularity in another important sense. That is, if we were to randomly sample one additional person from the population and have them engage in Study 2's judgment task, the pattern we should predict is this pattern, as its occurrence is distinguishable from random responding, with the highest observed probability of all possible patterns. We note that this may not be as compelling as finding a majority effect, but it is an empirical reality.

Overall, our results reveal the importance of perceived temptation on moral judgment. Theoretically, our studies provided tentative explanations for why previous research showed conflicting results, showing that, across analytic strategies, stimuli, and various ways of measuring moral judgment, people tend to judge agents who overcome immoral temptation as less virtuous than agents who never experience temptation in the first place. We also extended the literature on temptation, close relationships, and moral judgment by showing that the most common psychological experience is believing that being tempted to harm close others is even less virtuous than being tempted to harm strangers. These results add to a growing literature aimed at resituating moral psychology in everyday relational contexts (see Bloom, 2011). Methodologically, our findings corroborate recent calls to adopt analytic procedures that enable inferences about person-level psychology (e.g., McManus et al., 2023; Richters, 2021; Speelman & McGann, 2020). We hope the current research is one of the first of many to adopt these approaches and therefore resituate moral psychology not just in everyday relational contexts, but also in moral psychologists' intended object of study: individual persons.

8.1. Situating the current research within the broader moral psychology literature

The current studies connect to work assessing how motive inferences shape moral judgment. For example, donors who give anonymously in a double-blind way versus not, or those who behave prosocially in private versus in public, are judged as more charitable and more virtuous (De Freitas, DeScioli, Thomas, & Pinker, 2018; Kraft-Todd, Kleiman-Weiner, & Young, 2024), effects likely and empirically driven by differences in perceived reputational benefits. Additional research has documented how stated motives affect judgments of prosocial actors, with those motivated by empathy being judged more positively than those motivated by not wanting to feel distress (Erlandsson, Wingren, & Andersson, 2020), and those motivated by warm-glow emotions being judged more positively than those who are motivated by material rewards (Barasch, Levine, Berman, & Small, 2014). Extensive work shows similar effects on the relation between inferred motives, stated motives, and moral judgment (see Berman & Silver, 2022, for an overview). This all suggests that the reasons for why agents make prosocial or selfless decisions have strong effects on assessments of their underlying character. In relating these findings to the current work, participants in our paradigm were likely making inferences about what kind of person would be tempted to behave immorally (e.g., cheat on a spouse) and therefore infer their likelihood of eventually engaging in that behavior.

More relevant to our current studies, high effort prosocial actors are judged more positively than low effort prosocial actors (Berry & Lucas,

Table 2
Analysis summary across studies.

		Analysis				
		Group Level	**Descriptive Pervasiveness**	**Randomization Test**	**Frequentist Prevalence**	**Bayesian Prevalence**
Study	**Effect**					
Pilot A						
	Moral (NT > T)	✓	✓	-	✓	✓
Pilot B						
	Moral (NT > T)	✓	✓	-	✓	✓
Study 1						
	Binary (NT > T)	✓	✓	-	✓	✓
	Continuous (NT > T)	✓	✓	-	✓	✓
Study 2						
	Stranger (NT > T)	✓	✓	-	✓	✓
	Close Other (NT > T)	✓	✓	-	✓	✓
	Interaction (NT > T in S) and (NT > T in CO) and (CO > S)	✓	☐	✓	☐	☐

Note: A green check communicates that the pre-registered effect occurred, whereas a red X communicates that the pre-registered effect did not occur. Pre-registered effects are specified in parentheses with the following labels: NT = Non-Tempted agent, T = Tempted agent, S = Stranger, CO = Closer Other. The "Group-Level" column shows whether predicted effects occurred using typical group-level tests (e.g., t-tests). The "Descriptive Pervasiveness," "Frequentist Prevalence," and "Bayesian Prevalence" columns show whether predicted effects occurred in a majority of the sample or can be expected to occur in the majority of the population. The "Randomization Test" column shows whether the pre-registered interaction effect occurred in a percentage of the sample that was unlikely to occur via repeated random shuffling of the data. Due to the nature of our experimental designs (i.e., using relative scales to measure differential goodness or virtuosity), randomization tests could be conducted only for our interaction predictions; that is, shuffling of the simple effect data on their own (i.e., a single column of judgments) would continually result in the same descriptive pervasiveness percentage.

2022; Bigman & Tamir, 2016), as effort is a signal of morality's importance. However, if we consider our studies as manipulating mental effort, our findings are inconsistent with Berry & Lucas, 2022 and Bigman and Tamir (2016). Our findings are most consistent with Critcher et al. (2013), where slow and deliberative (i.e., high mental effort) prosocial behavior results in worse character judgments when compared to quick (i.e., low mental effort) prosocial behavior. We suggest that these patterns are due to inferences of the motives of tempted agents. As argued in Critcher et al. (2013), the presence of an immoral temptation communicates that the agent has multiple competing motives (i.e., to do good and bad). As hypothesized earlier, the existence of these competing motives might taint the agent's moral character and lead to predictions that they will eventually behave immorally, should additional opportunities arise. The absence of an immoral temptation, on the other hand, communicates that the agent does not have competing motives.

8.2. Limitations and future directions

The current research has several important limitations. First, we assumed throughout this research that, to test for a valid judgment difference between tempted and non-tempted agents, non-tempted agents must be described as simply lacking the desire of the tempted agents. This assumption is in stark contrast to Starmans & Bloom's stimuli (Starmans & Bloom, 2016), where non-tempted agents actively disliked what would generate a temptation in the first place. We therefore used the stimulus-wording from Berman and Small (2018) to make a claim about a non-confounded comparison between tempted and non-tempted agents.

However, Starmans & Bloom's comparison could indeed be valid in a different sense, i.e., if their goal was to uncover a moderator leading to tempted agents being judged as more virtuous than non-tempted agents. For example, one might be interested in whether physical capacities are an important moderator for how people think about tempted versus non-tempted agents. Consider two school-aged boys, a big/strong one who can physically harm their bully and another small/frail boy who could not harm their bully. If the strong boy resists the urge to retaliate, people might judge him as more virtuous, as he could have acted differently. We grant this possibility, but this is not a simple case that compares non-tempted and tempted agents. Instead, this case makes (implicit) assumptions about the definition of "non-tempted" (e.g., being coupled

with a physical incapacity to act). This example mirrors Starmans & Bloom's original stimuli in which a child's non-temptation to play outside with friends was coupled with a dislike of playing outside. Therefore, our current studies, as well as prior studies, occupy only a small portion of the space in which temptation-related stimuli can be conceptualized and manipulated.

Second, we gave participants explicit access to mental state information (internal conflict or lackthereof) that is rarely available in plain form outside of the lab. In the real world, this information must typically be inferred via behavior (e.g., decision speed, see Critcher et al., 2013). Therefore, the current research might be most analogous to instances in which moral judgments are made only after others have revealed their thoughts directly or indirectly through gossip. In general, future research should consider the robustness of the observed effects across more ecologically valid manipulations of temptation. For example, consider two men, John and Tony, out at two separate bars. They both are in committed relationships, and both happen to run into ex-girlfriends from college. When the bars close, the ex-girlfriends ask the men to go home with them. When asked, John decides to take a taxi back to his ex-girlfriend's apartment. However, when he gets out of the taxi, he paces around and ultimately gets back in the taxi and goes home. Tony, on the other hand, declines the offer and calls a taxi to go home. Here, temptation is manipulated by varying how each agent behaves in a multi-step plan before ultimately making the same decision to go home.

Third, we measured only third-person judgments of temptation. However, as has been shown in other moral judgment research (e.g., Hirschfeld-Kroen, Jiang, Wasserman, Anzellotti, & Young, 2021), there may be interesting and important first- versus third-person differences. Specifically, because people often have stronger priors about themselves compared to others, perhaps they generate situational attributions for their own experiences of temptation and therefore do not discount their own virtuosity. However, people may be more willing to generate dispositional attributions when they infer that others have experienced temptation, leading to discounting. This possibility is consistent with classic social psychological theorizing (e.g., Reeder & Brewer, 1979) and recent research on motivated versus rational belief maintenance from a third-party perspective (Kim, Mende-Siedlecki, Anzelotti, & Young, 2020; Kim, Park, & Young, 2020). Future research can determine whether similar processes occur when comparing first- to third-person attributions in general, and in the realm of temptation specifically.

Fourth, while our work expanded on others' research on the consequences of perceiving temptation for moral judgments (Berman & Small, 2018; Starmans & Bloom, 2016; Zhao & Kushnir, 2022), it cannot provide inferences about downstream behavioral consequences. Given that prior research suggests that trustworthiness is the most valued trait in others (Cottrell, Neuberg, & Li, 2007), understanding how inferred temptation affects perceived trustworthiness and actual trusting behavior is important. A modified two-stage economic exchange could accomplish this.

Finally, recent calls have been made for researchers to communicate constraints on the generality of their findings (see Simons, Shoda, & Lindsay, 2017; Yarkoni, 2020). Importantly, the current studies surveyed only U.S. participants using various crowdsourcing platforms. Future research is needed to determine the universality or uniqueness of the effects reported here. Relatedly, the methodology used throughout this paper (i.e., examining person-level responses) makes clear another sampling issue, even within cultures. Specifically, who are the people who show these effects or not, and more generally, who are the people being sampled? While robustness checks of our primary effect (i.e., that overcoming immoral temptation is judged as less virtuous than never being tempted) suggest that it occurred across all levels of all demographic factors collected (see the "Robustness Plots" in each RMarkdown on our OSF page: https://osf.io/bym4t/?view_only=ab62a6698bb84e34a53892e39576623f), the proportions of certain levels of some demographic factors were worrisome if our goal–or psychology's goal more generally–is to make universal claims. Therefore, although our statistical methodology allowed sample-to-population inferences about the prevalence of our effects, it is an important and open question whether these prevalence estimates change as demographic diversity increases.

9. Conclusion

Expanding the theoretical and methodological reach of previous research, the current work showed that people view agents' temptations harshly and especially so when temptations can affect close others. Specifically, most people (72–81%) judge agents who overcome temptations as less virtuous than agents who do not experience temptation. Moreover, this pattern is especially robust for judgments of agents who overcome temptations to harm close others (i.e., best friends or siblings) versus strangers. Together, our methodological approaches and empirical findings add to two exciting and growing literatures: the importance of resituating moral psychology into everyday relational contexts, and the importance of the individual person (rather than the population mean) in the study of psychology broadly. The continued interplay between these fields may end or rejuvenate classic debates in moral psychology, a prospect that, no matter the outcome, we are eager to witness.

CRediT authorship contribution statement

Ryan M. McManus: Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Formal analysis. **Helen Padilla Fong:** Data curation, Formal analysis, Visualization, Writing – review & editing, Investigation. **Max Kleiman-Weiner:** Supervision, Writing – review & editing, Methodology. **Liane Young:** Funding acquisition, Supervision, Writing – review & editing, Methodology.

Declaration of competing interest

The work described has not been published previously. It is not under consideration for publication elsewhere. It is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out. If accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder.

Data availability

Our data are already shared at: https://osf.io/bym4t/ (an anonymized link is in the main text)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2024.104615.

References

Allefeld, C., Görgen, K., & Haynes, J. D. (2016). Valid population inference for information based imaging: From the second-level t-test to prevalence inference. *Neuroimage, 141*, 378–392.

Aristotle. (1985). *Nicomachean ethics* (T. Irwin, Trans.). Indianapolis, IN: Hackett

Barasch, A., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or selfless? On the signal value of emotion in altruistic behavior. *Journal of Personality and Social Psychology, 107*, 393–413.

Berman, J. Z., & Silver, I. (2022). Prosocial behavior and reputation: When does doing good lead to looking good? *Current Opinion in Psychology, 43*, 102–107.

Berman, J. Z., & Small, D. A. (2018). Discipline and desire: On the relative importance of willpower and purity in signaling virtue. *Journal of Experimental Social Psychology, 76*, 220–230.

Berry, Z., & Lucas, B. J. (2022). How much is enough? The relationship between prosocial effort and moral character judgments. *Personality and Social Psychology Bulletin*, Article 1461672221135954. Advance online publication.

Bigman, Y., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General, 145*, 1654–1669.

Bloom, P. (2011). Family, community, trolley problems, and the crisis in moral psychology. *The Yale Review, 99*(2), 26–43.
Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology, 92*, 208–231.
Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*(3), 308–315.
De Freitas, J., DeScioli, P., Thomas, K. A., & Pinker, S. (2018). Maimonides' ladder: States of mutual knowledge and the perception of charitability. *Journal of Experimental Psychology: General, 148*, 158–173.
Donhauser, P. W., Florin, E., & Baillet, S. (2018). Imaging of neural oscillations with embedded inferential and group prevalence statistics. *PLoS Computational Biology, 14*(2), Article e1005990.
Erlandsson, A., Wingren, M., & Andersson, P. A. (2020). Type and amount of help as predictors for impression of helpers. *PLoS One, 15*, 1–23.
Everett, J. A. C., Pizarro, D., & Crockett, M. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772–787.
Faul, F., Erdfelder, E., Lang, A., & 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.
Grice, J. W. (2021). Drawing inferences from randomization tests. *Personality and Individual Differences, 179*, Article 110963.
Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science, 3*(4), 443–455.
Halfon, M. (1989). *Integrity: A philosophical inquiry*. Philadelphia: Temple University Press.
Hirschfeld-Kroen, J., Jiang, K., Wasserman, E., Anzellotti, S., & Young, L. (2021). When my wrongs are worse than yours: Behavioral and neural asymmetries in first-person and third-person perspectives of accidental harms. *Journal of Experimental Social Psychology, 94*, Article 104102.
Ince, R., Kay, J. W., & Schyns, P. G. (2022). Within-participant statistics for cognitive science. *Trends in Cognitive Sciences, 26*(8), 626–630.
Ince, R., Paton, A. T., Kay, J. W., & Schyns, P. G. (2021). Bayesian inference of population prevalence. *eLife, 10*, Article e62461.
Kant, I. (1998). *The groundwork for the metaphysics of morals* (M. J. Gregor, Trans.). Cambridge, England: Cambridge University Press (Original work published 1785).
Kim, M., Mende-Siedlecki, P., Anzelotti, S., & Young, L. (2020). Theory of mind following the violating of strong and weak prior beliefs. *Cerebral Cortex, 31*(2), 884–898.
Kim, M., Park, B., & Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences, 24*(2), 101–111.
Kraft-Todd, G., Kleiman-Weiner, M., & Young, L. (2024). *Virtue discounting: Observability reduces moral actors' perceived virtue*. Open Mind.
Law, K. F., Campbell, D., & Gaesser, B. (2021). Biased benevolence: The perceived morality of effective altruism across social distance. *Personality and Social Psychological Bulletin, 48*(3), 426–444.
Marshall, J., Gollwitzer, A., Mermin-Bunnell, N., Shinomiya, M., Retelsdorf, J., & Bloom, P. (2022). How development and culture shape intuitions about prosocial obligations. *Journal of Experimental Psychology: General, 151*(8), 1866–1882.
Marshall, J., Wynn, K., & Bloom, P. (2020). Do children and adults take social relationship into account when evaluating other peoples' actions? *Child Development, 91*, 1395–1835.
McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science, 31*(3), 227–242.
McManus, R. M., Mason, J. E., & Young, L. (2021). Re-examining the role of family relationships in structuring perceived helping obligations, and their impact on moral evaluation. *Journal of Experimental Social Psychology, 96*, Article 104182.
McManus, R. M., Young, L., & Sweetman, J. (2023). *Psychology is a property of persons, not averages or distributions: Confronting the group-to-person generalizability problem in experimental psychology*. Advanced in Methods and Practices in Psychological Science.
Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86*(1), 61–79.
Richters, J. E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology, 43*(6), 366–405.
Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123–1128.
Speelman, C. P., & McGann, M. (2020). Statements about the pervasiveness of behavior require data about the pervasiveness of behavior. *Frontiers in Psychology, 11*, 1–16.
Starmans, C., & Bloom, P. (2016). When the spirit is willing, but the flesh is weak: Developmental differences in judgments about inner moral conflict. *Psychological Science, 27*(11), 1498–1506.
Taylor, G. (1981). Integrity. In *, 55. Proceedings of the Aristotelian Society (supplementary volume)* (pp. 143–159).
Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences, 45*, E1.
Zhao, X., & Kushnir, T. (2022). When it's not easy to do the right thing: Developmental changes in understanding cost drive evaluations of moral praiseworthiness. *Developmental Science, e13257*.