UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Aiyer, S., Chan, W. Y. (Florence), & Feldman, G. (2024). Outcome Bias in Evaluations of Ethical Decisions: Replication and Extensions of Gino, Moore, and Bazerman (2009). *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.126266

Social Psychology

Outcome Bias in Evaluations of Ethical Decisions: Replication and Extensions of Gino, Moore, and Bazerman (2009)

Sriraj Aiyer[1][a], Wing Yan (Florence) Chan[2][b], Gilad Feldman[2][c]

[1] Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom, [2] Department of Psychology, University of Hong Kong, Hong Kong SAR

Keywords: outcome bias, decision evaluation, ethics, moral decisions, cognitive biases, pre-registered replication

https://doi.org/10.1525/collabra.126266

Collabra: Psychology

Vol. 10, Issue 1, 2024

Outcome bias is the phenomenon whereby decisions which resulted in positive outcomes are rated more favorably than when the same decisions resulted in negative outcomes, ceteris paribus. We conducted a pre-registered replication of Gino, Moore, and Bazerman (2009) Study 1's three scenarios (original's: N = 120) with an extension adding the three scenarios from their Study 2. Our data was collected online with an Amazon Mechanical Turk sample recruited using CloudResearch (N = 402). We tested outcome bias by measuring participants' ratings of how unethical, punishable, and blameworthy a decision maker's behavior was in morally grey scenarios. We partially replicated outcome bias in ratings of punishment (original: η^2_G = .05 [.00, .14]; replication 1-3: η^2_G = .03 [.01, .11]; replication 4-6: η^2_G = .11 [.05, .18]) and blame (original: η^2_G = .12 [.03, .23]; replication 1-3: η^2_G = .06 [.05, .19]; replication 4-6: η^2_G = .16 [.08, .23]), but with support for outcome bias in ratings of unethicality in Scenarios 4-6 (η^2_G = .04 [.01, .08]) but not in Scenarios 1-3 (original: η^2_G = .06 [.004, .16]; replication: η^2_G = .00 [.00, .03]). Similarly, we only found support for the target's finding that ratings of unethicality mediate the relationship between outcome and both perceptions of punishment and blame in Scenarios 4-6. We also added an extension of a control condition and found higher unethicality judgements when a decision resulted in a negative outcome relative to a control condition with no outcome information. Materials, data, and code are available on: https://osf.io/3bz2g/.

Background

Baron and Hershey's (1988) seminal article on outcome bias showed that outcomes tend to affect evaluations of the quality of decisions linked with that outcome. Even when the decision itself and the parameters of that decision were identical, individuals tended to judge decisions resulting in success more favorably than those resulting in failure, even when explicitly told to ignore the outcomes, and claiming to have ignored outcomes when asked whether outcomes impacted their evaluations. If such an outcome bias were not present in decision evaluations, such judgements would be made based only on *ex ante* information (available at the point of the decision) on the part of the decision maker. Hence, as a decision's outcome is only available *ex post* (and not available to the decision maker at the point of the decision), it should not factor into any such judgements. A recent well-powered pre-registered replication found support for Baron and Hershey (1988)'s findings (Aiyer et al., 2023).

Baron and Hershey's work inspired follow-up research that investigated outcome bias in a variety of contexts, including metascience (Emerson et al., 2010), medical decision-making (Savani & King, 2015), and social psychology (Mackie & Ahn, 1998). We were especially interested in the work of Gino, Moore, and Bazerman (2009), which extended outcome bias to the realm of ethical judgements. Their work provided evidence that individuals judged ethically grey decisions with a negative outcome as more unethical than identical decisions with a positive outcome, (η^2_G = .06, 95% CI [.004, .16]).

[a] Equal contribution first author

[b] Equal contribution first author

[c] Gilad Feldman, Department of Psychology, University of Hong Kong, Hong Kong SAR; gfeldman@hku.hk

Table 1. Gino, Moore, and Bazerman (2009) replication and extension: Hypotheses

#	Hypothesis
Original (on scenarios 1-3)	
1a *	People judge a decision in morally questionable situations as more unethical when the outcome is negative compared when the outcome is positive.
1b (deduced)	People judge a decision in morally questionable situations as more blameworthy when the outcome is negative compared when the outcome is positive.
1c (deduced)	People judge a decision in morally questionable situations as more punishable when the outcome is negative compared when the outcome is positive.
2a *	Ethical judgment of a decision mediates the relationship between outcome valence and judgment of punishment deserved for that decision.
2b *	Ethical judgment of a decision mediates the relationship between outcome valence and judgment of blame deserved for that decision.
Extensions	
3a*	People judge a questionable decision as more unethical when the outcome is negative compared to when the outcome is not specified.
3b*	People judge a questionable decision as more blameworthy when the outcome is negative compared to when the outcome is not specified.
3c*	People judge a questionable decision as more punishable when the outcome is negative compared to when the outcome is not specified.
4* (exploratory)	Outcome bias and the impact of outcome on dependent variables would be different in the scenarios from Study 1 and the scenarios in Study 2.
5a (scenarios 4-6)	Ethical judgment of a decision mediates the relationship between outcome valence (positive or negative) and judgment of punishment deserved for that decision.
5b (scenarios 4-6)	Ethical judgment of a decision mediates the relationship between outcome valence (positive or negative) and judgment of blame deserved for that decision.

Note. * indicates preregistered hypotheses.

Chosen Study for Replication: Gino, Moore, and Bazerman (2009)

We conducted an independent pre-registered and well-powered close replication of Gino, Moore, and Bazerman (2009). We also added two extensions to the original work.

Gino, Moore, and Bazerman (2009) examined whether the outcome of decisions extended beyond judgement of the decision's quality to also impact evaluations of ethicality when decisions were made in situations that were ethically grey. In their first study, participants were shown three scenarios depicting questionable decisions, with each resulting in either positive or negative outcomes. Participants then rated the ethicality of the decision and the extent to which the decision-maker deserved blame and punishment (all on a 7-point scale). They hypothesized that: 1) decisions resulting in negative outcomes would be viewed as more unethical than those resulting in positive outcomes, and 2) judgements of ethicality mediated the relationship between outcome and judgements of blame and punishment.

We chose Gino, Moore, and Bazerman (2009) as a target for our replication for a number of reasons. At the time that we initiated the project in 2020, the unpublished manuscript was available as a preprint but had already accrued over 120 citations. We found this to be an unusual impact for an unpublished manuscript, and so were puzzled by why the article had not been published given such strong community interest. In one of the first meta-science examinations of the prevalence of replications in publications, Makel et al (2012) found very low prevalence for replications (estimated at 1.07% of psychology publications) and noted that "if a publication is cited 100 times, we think it would be strange if no attempt at replication had been conducted and published".

When we reached out to the original authors, in 2022, they were surprised to learn that the article was preprinted, and that it had gained so much attention. The second author shared their journey whilst attempting to publish the manuscript (mostly in what are considered top management and psychology journals from around the year 2009) and it became clear that the reasons for rejection were mainly about its perceived contribution beyond the already known outcome bias and the literature on unethicality. Much has changed over the years regarding the perceptions of a "contribution", and there is reason to believe that if the preprint was submitted today (adhering to today's standards of transparency and openness) then it would have likely found a journal home. That the community deemed this unpublished preprint important enough to be cited 120 times, suggests that scholars found this to be a meaningful contribution to the literature.

We should at this point however address that although we embarked on this project in 2020, there have since been developments in our field that have raised some concerns in regard to work conducted by the lead author, with some reflections from the two other authors, especially the last author (Bazerman, 2022b, 2022a). This suggests an urgent need to revisit this work and assess its replicability in light of these developments. The Many Co-Authors' Project (2024) was initiated with the aim of revisiting all published findings that involved said lead author, yet this well-cited preprint has not been included in that project. However, a follow-up article that built on this preprint has been included and flagged as a concern for replicability (https://manycoauthors.org/gino/126). Whilst the preprint has been removed from the SSRN preprint server (https://doi.org/10.2139/ssrn.1099464) "at the request of the author, SSRN, or the rights holder", it can still be viewed on the Internet Archive Wayback Machine (https://web.archive.org/web/20191002030217/https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1099464). Google Scholar has since stopped indexing the preprint and counting its citations, losing crucial information about the impact this preprint has had on the field, yet that can still be deduced indirectly by examining all articles including the phrase "outcome bias in ethical judgments" (134 results as of March 2024).

The article has since led to impactful follow-up work, including the authors' own research on ethical judgments in Gino et al. (2010), as well as the work of Kneer and Machery's (2019) and Lench et al. (2015) exploring outcome biases in morality and attributions. We therefore believe our project would be a unique but timely case of aiming to replicate an impactful unpublished preprint, which is especially useful in a research landscape with pervasive replication and publication bias (in itself a form of outcome bias). From our correspondence with two of the authors, there are currently no replications of this work and assessing its replicability would also aid in gauging the overall generalizability of outcome bias, building off of the seminal work of Baron and Hershey (1988).

Finally, we also saw the potential in addressing some of the weaknesses in the target preprint that would allow for better testing and generalizability. Studies 1 and 2 were conducted with samples of 120 and 58 US undergraduates, fairly small samples of a specific population. The experimental design contrasting positive and negative outcomes could be strengthened by adding a neutral control condition with no-outcome, to allow insights regarding which of the conditions is driving the effect, or perhaps both.

Replication Closeness Evaluation

We conducted a replication of Study 1 from Gino, Moore and Bazerman (2009), with extensions adding the three scenarios from their Study 2. Based on LeBel et al.'s (2018) criteria for the evaluation of replications (Table 1), we classified our experiment as a close replication of the original study. Close replications as per LeBel et al.'s (2018) criteria allow for deviations in procedural details, physical setting and contextual variables.

Deviations from Gino, Moore, and Bazerman (2009)

We deviated from the target study in two main ways. Firstly, we used a different physical setting, as the original study was conducted in-person whereas we conducted the study online on MTurk using CloudResearch, and so our study population was different to that of the original study, which recruited only US undergraduate students. In the target study, the claims made were not about a specific population and therefore broader generalizability to other populations seems reasonable. Secondly, we extended the target study by adding a control condition where no decision outcome was provided to participants for the scenarios, yet given the random between-participants assignment, this should not impact the replication. We also added three more scenarios from the target's Study 2.

Extensions: Neutral Condition and Scenarios from Study 2

We added several extensions and analyses. First, we added a neutral baseline condition to tie the effect to the hindsight bias (Chen et al., 2021; Fischhoff, 1975; Slovic & Fischhoff, 1977) and examine how the impact of positive and negative outcomes on the dependent variables would compare to having information about the outcome. Hindsight bias suggests retrospective judgements of higher probability for a given event happening after having observed that event, leading to inflated knowledge of what an observer would have expected to happen without outcome knowledge. If our results showed evidence of a hindsight bias rather than an outcome bias, decision evaluations would be more swayed by event probabilities rather than the valence (positive or negative) of these events.

Second, we added the three scenarios used in the target article's Study 2 together with the baseline design of Study 1, to test for the robustness of the effect and to allow for a comparison of the different scenarios from the two studies.

Finally, the original paper analyzed whether unethicality mediated the relationship between outcome information and ratings of punishment and blame (see Hypotheses 2a and 2b in Table 1). As an extension, we also investigated the mediating role of unethicality in the three scenarios added from Study 2, such that we can determine if this finding replicates in other scenarios as well (see Hypotheses 5a and 5b in Table 1).

Experimental Manipulations in Target Article

In replicating this study, we note limitations with the manipulation in the target article. Whilst we did not alter the survey materials originally used (see Table 3), we followed the original authors' design in manipulating more of the wording than needed in the scenario texts. In Scenarios 1-3, the authors manipulated other aspects of the scenario apart from the outcome information, explaining their decision that using less unethical actions in the negative outcome condition to "test whether the outcome would overcome the ethicality of the actual decision". In Scenarios 4-6, the authors attempted to emphasize to participants that the outcomes of the depicted decisions were simply a result of probabilities (and hence out of control of the decision makers). This is framed in the target article as decision makers having a lack of 'inside knowledge'. However, it is unclear from the scenario texts whether these probabili-

Table 2. Replication classification: Using Lebel et al (2018) replication taxonomy

Design Facet	Same/ Different to Original Paper	Notes
Effect/Hypothesis	Same	
IV Construct	Same	
DV Construct	Same	
IV Operationalization	Similar with adjustment	We added comprehension checks to ensure that participants read and understood the scenario before proceeding to the DV. This was a needed adjustment for the online target sample. This served as a forced manipulation check.
DV Operationalization	Same	
IV Stimuli	Same	
DV Stimuli	Same	
Procedural Details	Similar	Study conducted via Qualtrics, control ('no outcome') condition added, three scenarios added from their Study 2.
Physical Setting	Different	Study conducted online rather than in-person
Contextual Variables	Different	Original study conducted with US University Students in 2008, current study conducted with American participants in 2020 via MTurk using CloudResearch
Population	Different	Original study recruited undergraduates; we recruited a more diverse population with participants on MTurk using CloudResearch
Overall	**Close Replication**	

ties are known to the decision makers. The probabilities are not kept constant between the positive and negative outcome conditions, in line with the above justification of depicting less unethical actions in the negative outcome condition. The original authors reported verifying that these were indeed interpreted as less unethical via a pilot study. We also note that because Scenarios 4-6 are different situations to Scenarios 1-3, differences between the two groups of scenarios/studies cannot be isolated to be about 'inside knowledge' and could instead be due to differences in the scenarios. While the changes to the ancillary text outside of the manipulation of outcome are explained by the original authors, we feel it important to note the lack of consistency between experimental conditions as a limitation, as it makes the manipulation of outcome less clear. We did not change the original survey materials, and designed our control (with no outcome) condition extension to more closely resemble the negative condition in terms of wording.

Pre-registration and Open-science

We provided all materials, data, and code on: https://osf.io/3bz2g/. We pre-registered the experiment and its analyses were on the Open Science Framework (OSF; https://doi.org/10.17605/OSF.IO/64T5K), data collection began after the pre-registration, and planned data analyses were only conducted after all data had been collected. All measures, manipulations, and exclusions conducted for this investigation are reported. The pre-registration and manuscript were written using a template by Feldman (2023).

Method

Participants

A total of 402 participants were recruited online from Amazon's Mechanical Turk (MTurk) using CloudResearch (Litman et al., 2017). Our pre-registration stipulated that we intended to analyze all data without any exclusions and then apply exclusions as a supplementary analysis. Seven participants met the pre-registered supplementary exclusion criteria. Our analysis of the remaining 395 participants showed that exclusions had little impact on the results, and we provided a summary of the results in the supplemental materials. All questions were forced response and therefore there was no missing data.

Participants first indicated their consent, with two questions confirming their eligibility, understanding, and agreement with study terms, which they had to answer with a "yes" and required responses in order to proceed to the study. The two questions also served as attention checks, with a randomized display order of the options (yes, no, not sure) - 1) "Are you able to pay close attention to the details provided and carefully answer questions that follow?", 2) "Do you understand the study outline and are willing to participate in a survey with attention/comprehension checks?". Failing any of the two attention questions meant that the participants did not indicate consent and therefore could not embark on the study and were asked to return the task. Upon completion of these steps, participants proceeded to begin the survey.

Based on our extensive experience of running similar judgment and decision making replications on MTurk, to ensure high quality data collection, we employed the following CloudResearch options: Duplicate IP Block. Duplicate Geocode Block, Suspicious Geocode Block, Verify Worker Country Location, CloudResearch Approved Participants.

Effect size and confidence intervals were all calculated in R (4.4.0; 2017) with the help of a guide by Jané et al. (2024), and then power analyses were conducted with GPower (version 3.1; Faul et al., 2007). We initially conducted a power analysis of the effects of outcome on punishment, as it was the smallest effect of the three dependent variables reported in the target article (unethicality: η^2_G= .059; punishable: η^2_G = .047; blameworthiness: η^2_G = .117). The required sample size with 0.95 power and 0.05 alpha was 265. One of the changes made to this study compared to the original study was increasing the number of between-participants conditions from two to three (adding a 'no outcome' control condition). This power analysis hence produces a value of 132.5 per condition, meaning that a total of 398 participants (when rounded up) is required across the three conditions.

The six scenarios we ran were either the replication of the three scenarios from Study 1, or the extension with three scenarios adapted from Study 2 of the target article. Participants were randomized as to whether they saw the Study 1 block (Scenarios 1-3) or Study 2 block (Scenarios 4-6) first, and then the scenarios within each block were shown in a randomized order.

We first pretested survey duration with 30 participants to make sure our time run estimate was accurate and adjusted pay as needed, the data of the 30 participants was not analyzed other than to assess technical issues, and survey completion duration and possible needed pay adjustments. These participants were included in the overall analyses. The full description of the scenarios is provided in the supplementary, and available in the Qualtrics exports in the pre-registration.

Manipulations

We summarized all the scenarios in Table 3.

Outcome (between-subject)

Participants were randomly assigned to one of the three conditions and answered all six scenarios in that assigned condition: positive outcome, negative outcome, or no outcome. An example scenario in the Replication group (Scenario 1: Pharmaceutical Researcher) begins with the following excerpt in the positive outcome condition:

> "A pharmaceutical researcher defines a clear protocol for determining whether or not to include clinical patients as data points in a study. He is running short of time to collect sufficient data points for his study within an important budgetary cycle within his firm. He believed that the product was safe and effective. As the deadline approaches, he notices that if he had four more data points for how subjects are likely to behave, the analysis would be significant. He makes up these data points, and soon the drug goes to market."

In the negative and no-outcome condition, the text has some differences, which are marked with *italics* (as explained in the 'Experimental Manipulations in Target Article' section):

> "A pharmaceutical researcher defines a clear protocol for determining whether or not to include clinical patients as data points in a study. He is running short of time to collect sufficient data points for his study within an important budgetary cycle within his firm. *As the deadline approaches, he notices that four subjects were withdrawn from the analysis due to technicalities. He believes that the data in fact is appropriate to use, and when he adds those data points, the results move from not quite statistically significant to significant (i.e. the results would support the effectiveness of the product).* He adds these data points, and soon the drug goes to market.

For participants in the no-outcome condition, the above excerpt is the whole scenario that they read. For the other participants, the wording of an extra added sentence on the scenario's outcome was manipulated depending on the condition they were assigned to. This was worded as follows (square brackets indicate the relevant condition):

> "This drug is a profitable and effective drug, and years later shows no significant side effects." [Positive]
> "This drug is later withdrawn from the market after it kills six patients and injures hundreds of others." [Negative]

The full set of scenarios and their wording can be found in the supplemental materials.

Extension: Comparing Target's Study 1 (Scenarios 1-3) to Study 2 (Scenarios 4-6) (within-subject)

We added the three scenarios from Study 2 in the target article with the same design and manipulations, referred to here as Scenarios 4, 5, and 6. This allowed us to compare the three scenarios in Study 1 (Scenarios 1-3) with the three scenarios in Study 2 (Scenarios 4-6).

Comprehension Manipulation Checks

We made an adjustment and deviated from the original study by adding comprehension checks after the scenarios and before moving to the next page to answer the dependent measures. Given concerns about inattentiveness, we wanted to ensure that participants paid attention to the critical information in the scenario. We considered this a more conservative test of the target's design. Participants were required to answer all questions correctly before they could proceed to the next page.

We included two comprehension checks for each scenario that participants had to answer correctly before proceeding. One question asked participants to correctly identify what the decision-maker did in the scenario (e.g., "What did the research/auditor/toy company do?") and the other asked participants what the outcome of the decision was (e.g., "What was the **outcome** of [...]", with possible answers: "the outcome was positive/negative/not specified") .

Table 3. Scenarios 1 to 6 by outcome condition (between-subject design)

Scenario	Positive-outcome condition	Negative-outcome condition	No-outcome condition
Study 1's Scenario 1: Pharmaceutical researcher	A pharmaceutical researcher defines a clear protocol for determining whether or not to include clinical patients as data points in a study. He is running short of time to collect sufficient data points for his study within an important budgetary cycle within his firm. He believed that the product was safe and effective. As the deadline approaches, he notices that if he had four more data points for how subjects are likely to behave, the analysis would be significant (i.e. the results would support the effectiveness of the product). **He makes up these data points**, and soon the drug goes to market. ***This drug is a profitable and effective drug, and years later shows no significant side effects.***	A pharmaceutical researcher defines a clear protocol for determining whether or not to include clinical patients as data points in a study. He is running short of time to collect sufficient data points for his study within an important budgetary cycle within his firm. As the deadline approaches, **he notices that four subjects were withdrawn from the analysis due to technicalities**. **He believes that the data in fact is appropriate to use,** and when he **adds those data points**, the results move from not quite statistically significant to significant (i.e. the results would support the effectiveness of the product). He adds these data points, and soon the drug goes to market. *This drug is later withdrawn from the market after **it kills six patients and injures hundreds of others**.*	A pharmaceutical researcher defines a clear protocol for determining whether or not to include clinical patients as data points in a study. He is running short of time to collect sufficient data points for his study within an important budgetary cycle within his firm. As the deadline approaches, **he notices that four subjects were withdrawn from the analysis due to technicalities**. **He believes that the data in fact is appropriate to use,** and when he **adds those data points**, the results move from not quite statistically significant to significant (i.e. the results would support the effectiveness of the product). He adds these data points, and soon the drug goes to market. *We do not know yet know the outcome of this situation.*
Study 1's Scenario 2: Auditor	An auditor is examining the books of an important client, a client that is not only valuable for their auditing fees, but also buys lucrative advisory services from the auditor's firm as well. **The auditor notices clearly fraudulent practices by their client**. The auditor brings up the issue with the client, who insists that there is nothing wrong with their accounting. The client also threatens to withdraw their business if the auditor withholds their approval. **The auditor agrees to let it go by for one year**, and encourages the client to change their accounting practices over the next year. ***No problems result from the auditor's decision.***	An auditor is examining the books of an important client, a client that is not only valuable for their auditing fees, but also buys lucrative advisory services from the auditor's firm as well. **The auditor notices some accounting practices that are probably illegal, but it would take multiple court cases to be sure about whether the action was legal or not.** The auditor brings up the issue with the client, who insists that there is nothing wrong with their accounting. The client also threatens to withdraw their business if the auditor withholds their approval. **The auditor agrees to let it go by for one year,** and encourages the client to change their accounting practices over the next year. *Six months later, **it is found that the client was committing fraud, their corporation goes bankrupt**, the bankruptcy is connected to the issue that the auditor noticed, and **1,400 people lose their jobs and their life's savings.***	An auditor is examining the books of an important client, a client that is not only valuable for their auditing fees, but also buys lucrative advisory services from the auditor's firm as well. **The auditor notices some accounting practices that are probably illegal, but it would take multiple court cases to be sure about whether the action was legal or not.** The auditor brings up the issue with the client, who insists that there is nothing wrong with their accounting. The client also threatens to withdraw their business if the auditor withholds their approval. **The auditor agrees to let it go by for one year,** and encourages the client to change their accounting practices over the next year. *We do not know yet know the outcome of this situation.*
Study 1's Scenario 3: Toy company	A toy company sells products made by another firm, manufactured in another country. **The toy company knows that the toys contain lead, which can be extremely hazardous to children**. *The toy company successfully sells this product, makes a significant product, and **no children are injured by the lead paint**.*	A toy company **finds out that the products that they were selling**, manufactured by another firm in another country, **contain lead, which can be extremely hazardous to children**. The toy company had **failed to test for lead in the product, since testing is expensive and is not required by U.S. law.** *The lead paint eventually **kills 6 children**, and sends **dozens more to the emergency room** for painful treatment for lead poisoning.*	A toy company **finds out that the products that they were selling**, manufactured by another firm in another country, **contain lead, which can be extremely hazardous to children**. The toy company had **failed to test for lead in the product, since testing is expensive and is not required by U.S. law.** *We do not know yet know the outcome of this situation.*
Study 2's Scenario 4: Sewage treatment	A sewage treatment plant is undergoing remodeling and updating. There is a critical phase to the project during which all the	A sewage treatment plant is undergoing remodeling and updating. There is a critical phase to the project during which all the	A sewage treatment plant is undergoing remodeling and updating. There is a critical phase to the project during which all the

Scenario	Positive-outcome condition	Negative-outcome condition	No-outcome condition
plant	treatment systems are shut off and incoming sewage is diverted to a holding tank until the new systems are activated. This critical phase of the project will last 48 hours. **If there should be substantial rainfall during this 48-hour period, there is a high probability that the holding tanks will overflow into local waterways, with serious negative environmental and health consequences** for the wildlife and people who live in the area. **The company that runs the sewage treatment plant could invest in back-up systems that would eliminate the risk** of overflow even in the case of heavy rain, but this would be expensive. **Historically, the chances of rain during the planned 48 hour period are about 10%.** After extensive consideration, t**he manager in charge of the remodeling decides against instituting the back-up plan.** *In fact, there is* ***no rain during the critical 48-hour period and the plant remodeling is a complete success.***	treatment systems are shut off and incoming sewage is diverted to a holding tank until the new systems are activated. This critical phase of the project will last 48 hours. **If there should be substantial rainfall during this 48-hour period, there is a high probability that the holding tanks will overflow into local waterways, with serious negative environmental and health consequences** for the wildlife and people who live in the area. **The company that runs the sewage treatment plant could invest in back-up systems that would eliminate the risk** of overflow even in the case of heavy rain, but this would be expensive. **Historically, the chances of rain during the planned 48 hour period are about 5%.** After extensive consideration, the **manager in charge of the remodeling decides against instituting the back-up plan**. *It winds up raining a great deal.* ***Twenty people fall ill and wind up in the hospital, many fish die, the water is unsafe for swimming for a week, and fishers are discouraged from eating fish caught downstream.***	treatment systems are shut off and incoming sewage is diverted to a holding tank until the new systems are activated. This critical phase of the project will last 48 hours. **If there should be substantial rainfall during this 48-hour period, there is a high probability that the holding tanks will overflow into local waterways, with serious negative environmental and health consequences** for the wildlife and people who live in the area. **The company that runs the sewage treatment plant could invest in back-up systems that would eliminate the risk** of overflow even in the case of heavy rain, but this would be expensive. **Historically, the chances of rain during the planned 48 hour period are about 5%.** After extensive consideration, the **manager in charge of the remodeling decides against instituting the back-up plan**. *We do not know yet know the outcome of this situation.*
Study 2’s Scenario 5: Natural disaster	A government agency in a developing country finds itself dealing with a **natural disaster in which several thousand poor peasants have been made homeless** during the winter. The agency must decide what sort of short-term housing it will provide for the refugees**. The inexpensive option is tents, which will probably be fine, given the mildness of the local winters—overnight temperatures only fall below freezing once every two years or so, on average.** The more expensive option is to put up temporary shacks that would provide more shelter against the cold. But the shacks would be more expensive and would force the agency to cut funding to other (less urgent) programs. *In the end,* ***the agency's commissioner decides to provide only tents*** *for the refugees. The winter is quite mild, and the* ***tents provide sufficient shelter.***	A government agency in a developing country finds itself dealing with a **natural disaster in which several thousand poor peasants have been made homeless** during the winter. The agency must decide what sort of short-term housing it will provide for the refugees. **The inexpensive option is tents, which will probably be fine, given the mildness of the local winters—overnight temperatures only fall below freezing once every four years or so, on average.** The more expensive option is to put up temporary shacks that would provide more shelter against the cold. But the shacks would be more expensive and would force the agency to cut funding to other (less urgent) programs. *In the end, the* ***agency's commissioner decides to provide only tents*** *for the refugees. The* ***winter is substantially colder than expected, and fifty children among the refugees die of exposure to the cold.***	A government agency in a developing country finds itself dealing with a **natural disaster in which several thousand poor peasants have been made homeless** during the winter. The agency must decide what sort of short-term housing it will provide for the refugees. **The inexpensive option is tents, which will probably be fine, given the mildness of the local winters—overnight temperatures only fall below freezing once every four years or so, on average.** The more expensive option is to put up temporary shacks that would provide more shelter against the cold. But the shacks would be more expensive and would force the agency to cut funding to other (less urgent) programs. *In the end, the* ***agency's commissioner decides to provide only tents*** *for the refugees. We do not know yet know the outcome of this situation.*
Study 2’s Scenario 6: Water supply	There is a river that runs through dry regions in Mexico. **Eighty percent of the time, there has been plenty of water to supply the communities that depend on water from the river.** So when the mayor of a prosperous town that drew its water from the river was asked to **invest in water conservation measures, she assumed they were unnecessary.** It so happened that this town was	There is a river that runs through dry regions in Mexico. **Ninety percent of the time, there has been plenty of water to supply the communities that depend on water from the river.** So when the mayor of a prosperous town that drew its water from the river was asked to **invest in water conservation measures, she assumed they were unnecessary.** It so happened that this town was	There is a river that runs through dry regions in Mexico. **Ninety percent of the time, there has been plenty of water to supply the communities that depend on water from the river.** So when the mayor of a prosperous town that drew its water from the river was asked to **invest in water conservation measures, she assumed they were unnecessary.** It so happened that this town was

Scenario	Positive-outcome condition	Negative-outcome condition	No-outcome condition
	upstream from most of the other communities that depended on the river. Upstream communities naturally have the advantage because they can take what they need first, and downstream communities left without water have little recourse. ***In the end, there was plenty of rain and more than enough water for the communities along the river.***	upstream from most of the other communities that depended on the river. Upstream communities naturally have the advantage because they can take what they need first, and downstream communities left without water have little recourse. ***One year, rainfall is far below expectations and 46 small farms are driven out of business when the lack of water in the river leaves them unable to irrigate their land.***	upstream from most of the other communities that depended on the river. Upstream communities naturally have the advantage because they can take what they need first, and downstream communities left without water have little recourse. *We do not know yet know the outcome of this situation.*

Note. Bold text was as shown to participants, but was not in the original materials as they were provided in the target article, and was therefore a deviation meant to aid participants focus on key elements of the manipulation. We also structured the scenario differently to allow readers to easily compare the conditions. We note that the scenarios were adapted as is, and we do not know why certain elements of the scenarios were changed despite not being related to the manipulation. We show the text that indicates outcome in italics and the main manipulated text is underlined for Scenarios 4-6 where participants are told about the underlying probabilities of the outcomes in the scenarios. Italicized and underlined text was not shown to participants, they are depicted here for ease of readability.

Measures

Unethicality

Participants were asked: "How **unethical** is the [decision-maker's] behavior?" (1 = *Not at all*; 7 = *Extremely*).

Punishment

Participants were asked: "How harshly would you **punish** the [decision-maker's] behavior?" (1 = *Not at all*; 7 = *Extremely*).

Blame

Participants were asked: "How much would you **blame** the [decision-maker's] behavior?" (1 = *Not at all*; 7 = *Extremely*).

Results

We summarized descriptive statistics in Table 4 and the analyses in Table 5, and plotted the results in Figures 1 to 6. All analyses were conducted using the software R (R Core Team, 2017) version 4.2.0.

We conducted an analysis of variance (ANOVA) for each dependent variable with outcome type (positive / negative / no outcome) as a between-participants factor and, for hypothesis 4, scenarios (from Study 1 vs. from Study 2) as a within-participants factor by comparing scenarios 1-3 to scenarios 4-6. In addition, independent-samples two-tailed Student's t-tests were used to test specific hypotheses (Table 1) by outcome type.

Replication Hypotheses

Hypothesis 1 – Outcome Bias on Ethical Judgements

Hypothesis 1, as also predicted in the target article, predicted that participants would judge a decision as more unethical when resulting in a negative outcome compared to a positive outcome. We conducted a series of mixed 3 (within-participants: scenarios – 1,2,3) x 2 (between-participants: outcome – positive, negative) ANOVAs for each dependent variable. We do not find evidence of an effect of outcome on judgements of unethicality (F(1,265) = 1.32, p = .25, η^2_G = .003, 95% CI = [0, 0.02]). We do however find evidence of an effect on both judgements of punishment (F(1,265) = 14.61, p < .001, η^2_G = .027, 95% CI = [0.002, 0.06]). and judgements of blame (F(1,265) = 33.69, p < .001, η^2_G = .063, 95% CI = [0.02, 0.12]). Hence, we observe a similar effect to the original paper for two of the three dependent variables.

Hypothesis 2: Unethicality Mediation

We conducted a mediation analysis using the mediation R package (Tingley et al., 2014) to explore judgements of unethicality mediated the relationship between outcome and both judgements of punishment and blame. First, the outcome type (success vs failure) was regressed on unethicality ratings as a mediator. Then unethicality and outcome type were regressed on the punishment ratings in one model and blame ratings in a second model. Lastly, each estimate of the causal mediation effect (indirect effect: IE) was computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determine the indirect effect at the 2.5% and 97.5% percentiles for the unethicality mediator. We fit a linear model with all data from all scenarios with a positive or negative outcome. We found no support for the effect of outcome (positive vs negative) being mediated by judgements of unethicality. The bootstrapped indirect effect of unethicality on punishment was .05, 95% CI [-.04, .12], Sobel Z = 0.84, p = .40, ACME found to be robust until ρ = 0.7. The bootstrapped indirect effect of unethicality on blame was .03, 95% CI [-.02, .09], Sobel Z = 0.82, p = .42, ACME found to be robust until ρ = 0.5.

Extension Hypotheses

In this series of analyses, we test our extension hypotheses 3 and 4, firstly by looking at changes in ratings when including a control condition where no outcome is provided. We then analyze ratings from Scenarios 4-6.

Table 4. Descriptives for all scenarios and conditions

	1	2	3	1-3 Total	4	5	6	4-6 Total
Unethicality								
Positive Outcome	6.01 (1.32)	5.44 (1.42)	6.47 (1.02)	**5.97** **(1.33)**	3.79 (1.95)	3.34 (1.59)	3.81 (1.84)	**3.64** **(1.81)**
Negative Outcome	6.26 (1.17)	5.57 (1.48)	6.49 (0.95)	**6.11** **(1.28)**	4.71 (1.70)	5.08 (1.77)	4.38 (1.84)	**4.72** **(1.79)**
No Outcome	5.24 (1.69)	5.01 (1.57)	6.18 (1.23)	**5.48** **(1.59)**	4.50 (1.81)	4.00 (1.80)	4.59 (1.62)	**4.36** **(1.76)**
Punishment								
Positive Outcome	5.04 (1.67)	4.62 (1.58)	6.10 (1.18)	**5.25** **(1.61)**	3.02 (1.91)	2.56 (1.52)	3.12 (1.65)	**2.90** **(1.72)**
Negative Outcome	4.94 (1.68)	4.94 (1.68)	6.40 (0.99)	**5.43** **(1.64)**	4.68 (1.73)	4.96 (1.86)	4.11 (1.73)	**4.58** **(1.81)**
No Outcome	4.87 (1.74)	4.52 (1.66)	6.01 (1.26)	**5.13** **(1.69)**	4.43 (1.79)	3.75 (1.78)	4.30 (1.67)	**4.16** **(1.77)**
Blame								
Positive Outcome	5.07 (1.83)	4.66 (1.68)	5.58 (1.67)	**5.10** **(1.76)**	3.44 (2.02)	3.28 (1.87)	3.31 (1.80)	**3.35** **(1.90)**
Negative Outcome	6.26 (1.14)	5.07 (1.64)	6.47 (1.02)	**5.93** **(1.43)**	5.12 (1.70)	5.41 (1.60)	4.75 (1.70)	**5.10** **(1.69)**
No Outcome	5.40 (1.64)	4.77 (1.63)	6.17 (1.12)	**5.45** **(1.58)**	5.04 (1.79)	4.37 (1.77)	4.86 (1.62)	**4.76** **(1.75)**

Note. Format = means (standard deviation).
n: positive outcome = 134, negative outcome = 133, no-outcome = 135.

Table 5. Statistical tests, effects, and comparison to target article

DVs		Positive outcome *M* (*SD*)	Negative outcome *M* (*SD*)	*F*	*df*	*p*	η^2_G	95%CI	Interpretation
Scenarios 1-3 (Replication)									
Unethicality	Original	5.28 (1.13)	5.77 (0.79)	7.35	1,118	.008	.059	[.004, .16]	No signal, inconsistent, weaker effect
	Replication	5.97 (1.03)	6.11 (0.90)	1.32	1,265	.252	.003	[.00, .02]	
Punishment	Original	5.01 (1.10)	5.46 (0.94)	5.81	1,118	.017	.047	[.001, .14]	Signal, consistent, similar effect
	Replication	5.25 (1.18)	5.77 (1.01)	14.61	1,265	<.001	.03	[.002, .06]	
Blame	Original	4.65 (1.40)	5.49 (0.90)	15.60	1,118	<.001	.117	[.03, .23]	Signal, consistent, similar effect
	Replication	5.10 (1.40)	5.94 (0.89)	33.69	1,265	< .001	.06	[.02, .12]	
Scenarios 4-6 (Extension)									
Unethicality	Extension	3.64 (1.81)	4.72 (1.79)	26.51	1,265	<.001	.04	[.01, .08]	Signal (supported)
Punishment	Extension	2.90 (1.72)	4.58 (1.81)	73.39	1,265	<.001	.11	[.05, .18]	Signal (supported)
Blame	Extension	3.35 (1.90)	5.10 (1.69)	89.20	1,265	<.001	.16	[.08, .23]	Signal (supported)

Note. We used the LeBel et al. (2019) paradigm for comparison of original and replication with reference to signal direction and target's effect size overlap with replication's confidence intervals.

Hypothesis 3: No-outcome Condition

As per Hypothesis 3, we predicted that participants would judge a decision as more unethical when resulting in a negative outcome compared to when no outcome information is available. We conducted a series of mixed 3 (within-participants: scenarios – 1,2,3) x 3 (between-participants: outcome – positive, negative, no outcome) ANOVAs for each dependent variable. We find evidence for an effect of outcome on judgements of unethicality ($F(2,399) = 14.54$, $p < .001$, $\eta^2_G = .036$, 95% CI = [0.01, 0.07]). Post-hoc tests showed that the ratings on unethicality in

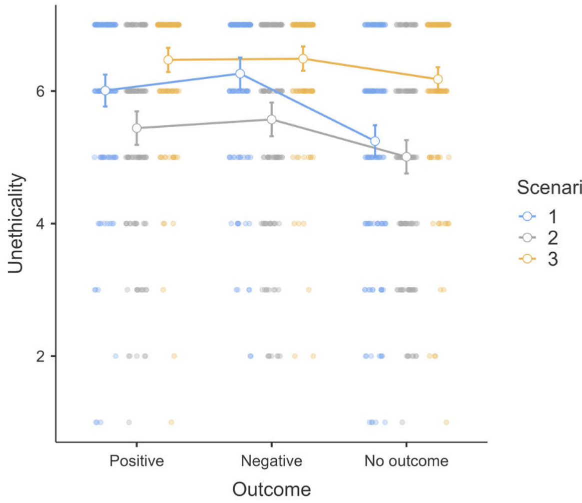


Figure 1. Replication: Scenarios 1-3 unethicality judgments across outcome conditions

Note. Ratings of unethicality across Scenarios 1-3 (replicating the target article's Study 1).
Scale: 1 = *Not at all*; 7 = *Extremely*.

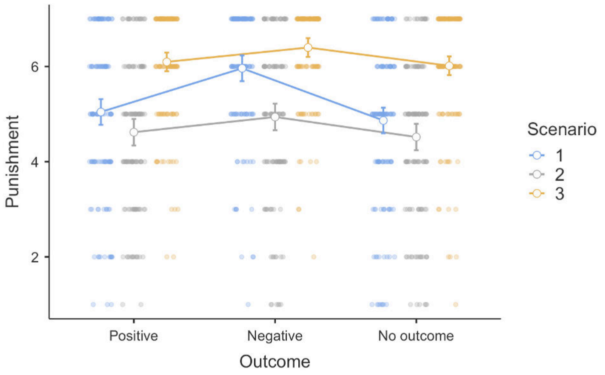


Figure 2. Replication: Scenarios 1-3 punishment judgments across outcome conditions

Note. Ratings of unethicality across Scenarios 1-3 (replicating the target article's Study 1).
Scale: 1 = *Not at all*; 7 = *Extremely*.

negative condition (M = 6.11, SD = 1.28) were higher than in the no-outcome condition (M = 5.48, SD = 1.59), t(771.1) = 6.21, d = 0.44, 95% CI [0.30, 0.58], p < .001. We also found evidence for an effect of outcome on judgements of punishment (F(2,399) = 12.10, p < .001, η^2_G = .029, 95% CI = [0.004, 0.06]). Post-hoc tests reveal higher ratings of punishment in the negative condition (M = 5.77, SD = 1.49) than in the no-outcome condition (M = 5.13, SD = 1.69), t(792.4) = 5.64, d = 0.40, 95% CI [0.26, 0.54], p < .001. Finally, we also find evidence of an effect of outcome on judgements of blame (F(2,399) = 18.34, p < .001, η^2_G = .044, 95% CI = [0.01, 0.08]). Post hoc tests reveal higher ratings of blame in the negative condition (M = 5.93, 1.43) than in the no-outcome condition (M = 5.45, SD = 1.58), t(796.4) = 4.58, d = 0.32, 95% CI [0.18, 0.46], p < .001.

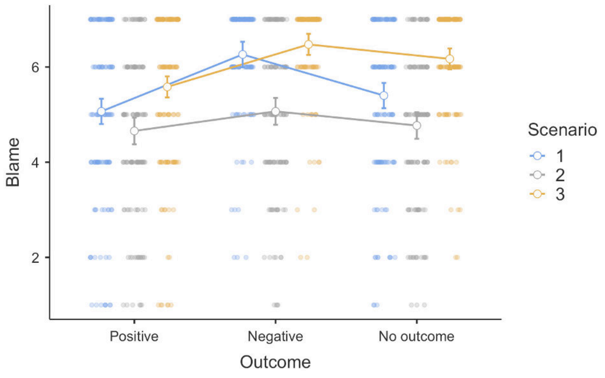


Figure 3. Replication: Scenarios 1-3 blame judgments across outcome conditions

Note. Ratings of unethicality across Scenarios 1-3 (replicating the target article's Study 1).
Scale: 1 = *Not at all*; 7 = *Extremely*.

Hypothesis 4: Scenarios from Study 2

We explored differences in outcome bias comparing the scenarios from the target article's Study 1 and Study 2. We plotted the findings in Figures 4 to 6.

We conducted a series of between-participants 2 (Scenarios 1-3 vs. Scenarios 4-6) x 2 (outcome – positive, negative) ANOVAs on each dependent variable. We observed evidence for a main effect of outcome (F(1,265) = 26.51, p < .001, η^2_G = .037, 95% CI = [0.01, 0.08]) and of study scenarios (F(1,265) = 412.75, p < .001, η^2_G = .348, 95% CI = [0.26, 0.43]) on judgements of unethicality. We also find evidence of an interaction between the two (F(1,265) = 26.39, p < .001, η^2_G = .022, 95% CI = [0.001, 0.06]). In Scenarios 4-6, decisions resulting in negative outcomes (M = 6.11, SD = 0.89) were regarded as more unethical than those resulting in positive outcomes (M = 5.97, SD = 1.03) (t(260.7) = 1.15, d = 0.14, 95% CI [-0.10, 0.38], p = .25). In Scenarios 1-3 decisions resulting in negative outcomes (M = 4.72, SD = 1.45) were again regarded as more unethical than those resulting in positive outcomes (M = 3.64, SD = 1.40) (t(264.6) = 6.16, d = 0.75, 95% CI [0.51, 0.99], p < .001). Hence, participants judged decisions as more unethical overall in Scenarios 4-6.

We also observed evidence for a main effect of outcome (F(1,265) = 73.39, p < .001, η^2G = .108, 95% CI = [0.05, 0.18]) and of scenarios (F(1,265) = 388.73, p < .001, η^2_G = .281, 95% CI = [0.19, 0.36]) on judgements of punishment. We also find evidence of an interaction between the two (F(1,265) = 42.33, p < .001, η^2_G = .031, 95% CI = [0.003, 0.07]). In Scenarios 4-6, decisions resulting in negative outcomes (M = 5.77, SD = 1.01) were regarded as more deserving of punishment than those resulting in positive outcomes (M = 5.25, SD = 1.17), (t(259.9) = 3.82, d = 0.47, 95% CI [0.23, 0.71], p < .001). In Scenarios 1-3, decisions resulting in negative outcomes (M = 4.58, SD = 1.46) were again regarded as more deserving of punishment than those resulting in positive outcomes (M = 2.90, SD = 1.41), (t(264.6) = 9.57, d = 1.17, 95% CI [0.93, 1.41], p < .001).

Finally, we observed evidence for a main effect of outcome (F(1,265) = 89.20, p < .001, η^2_G = .156, 95% CI = [0.08,

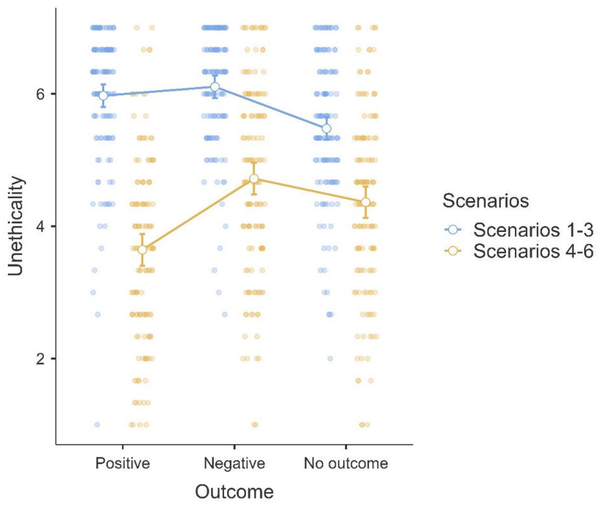


Figure 4. Extension: Unethicality judgments across outcome conditions and comparing Scenarios 1-3 to Scenarios 4-6

Note. Scenarios 1-3 vs. 4-6; by outcome (positive vs. negative vs. no-outcome). Scale: 1 = *Not at all*; 7 = *Extremely*

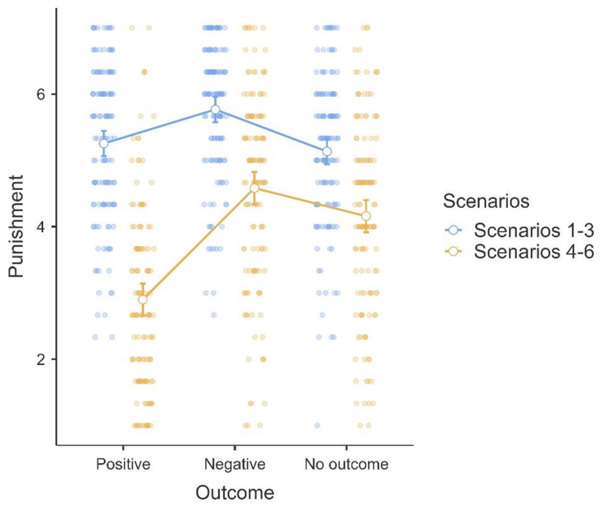


Figure 5. Extension: Punishment judgments across outcome conditions and comparing Scenarios 1-3 to Scenarios 4-6

Note. Scenarios 1-3 vs. 4-6; by outcome (positive vs. negative vs. no-outcome). Scale: 1 = *Not at all*; 7 = *Extremely*

0.23]) and of scenarios (F(1, 265) = 206.39, p < .001, η^2_G = .158, 95% CI = [0.09, 0.24]) on judgements of blame. We also find evidence of an interaction between the two (F(1,265) = 25.73, p < .001, η^2_G = .020, 95% CI = [0.0002, 0.05]). In Scenarios 4-6, decisions resulting in negative outcomes (M = 5.93, SD = 0.89) were regarded as more deserving of blame than those resulting in positive outcomes (M = 5.10, SD = *1.40*), (t(226.5) = 5.81, d = 0.71, 95% CI [0.47, 0.95], p < .001). In Scenarios 1-3, decisions resulting in negative outcomes (M = 5.10, SD = 1.33) were again regarded as more deserving of blame than those resulting in positive outcomes (M = 3.35, SD = 1.62), (t(255.9) = 9.62, d = 1.18, 95% CI [0.94, 1.42], p < .001).

Hypothesis 5: Unethicality Mediation in Scenarios 4-6

We conducted a mediation analysis using the mediation R package (Tingley et al., 2014) using the same analysis process as for replication Hypothesis 2 but for Scenarios 4-6 examining the extent to which unethicality mediated the relationship between outcome and ratings of both punishable and blameworthy the decision maker is in these scenarios. We note that this analysis was not preregistered.

The effect of outcome (positive vs negative) was found to be mediated by judgements of unethicality. The bootstrapped indirect effect of unethicality on punishment was .27, 95% CI [.19, .35], Sobel Z = 4.87, p = <.001, ACME found to be robust until ρ = 0.8. The bootstrapped indirect effect of unethicality on blame was .24, 95% CI [.16, .31], Sobel Z = 4.52, p < .001, ACME found to be robust until ρ = 0.7 (suggesting both mediation effects to be robust to unobserved confounding) .

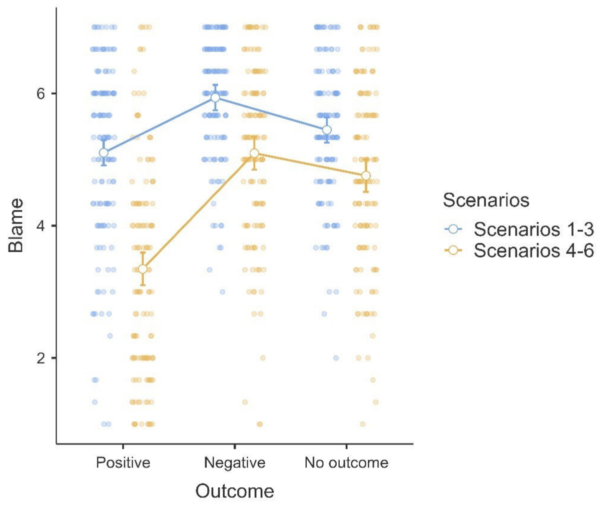


Figure 6. Extension: Blame judgments across outcome conditions and comparing Scenarios 1-3 to Scenarios 4-6

Note. Scenarios 1-3 vs. 4-6; by outcome (positive vs. negative vs. no-outcome). Scale: 1 = *Not at all*; 7 = *Extremely*

Discussion

We conducted a well-powered pre-registered direct replication and extension of Gino, Moore, and Bazerman's (2009) findings on outcome bias in ethical judgements. We concluded a successful replication with signal and direction consistent with that of the target article's findings. Outcomes affected perceived unethicality, in that a decision resulting in positive outcomes was judged as more ethical than a decision resulting in negative outcomes. Moreover, we extended the work by exploring the effect of outcome bias relative to a baseline condition, such that decisions resulting in negative outcome were viewed as more unethical than when no outcome was shown. We also studied the potential effects of scenarios on the part of the decision-maker on the prevalence of outcome bias. We find evidence of an outcome bias, but that its effect depends on both the dependent variables used and the scenarios used. Specifically, punishment and blame were more robust across scenarios, whilst ratings of unethicality were more sensitive to scenario. For the latter, we find that outcome bias was stronger in Scenarios 4-6 compared to Scenarios 1-3.

Replication of Outcome Bias in Ethical Decisions

As per Hypothesis 1, when attempting to replicate the effects of Study 1 in the target article, we found evidence of a similar effect of outcome bias when individuals made judgements of punishment and blame, but found no evidence in support of an effect when making judgements of unethicality. As per Hypothesis 2, we did not find evidence of judgements of unethicality mediating the relationship between outcome and both judgements of punishment and blame.

The relationships between these three variables have been explored in extant literature. Cushman (2008) observed that decisional outcome substantially affected judgments of blame and punishment but only marginally affected judgments of "wrongness". Kneer and Machery (2019) reported that feelings of blame were associated with unethicality but not punishment, and that decisional outcome had a small to medium effect on both judgements of "wrongness" and blame but a medium to large effect on judgments of punishment. Our results correspond with Cushman's (2008) Dual Process Model for moral judgment, according to which punishment judgments correlate with blame judgments and judgments of both deserved punishment and blame are susceptible to outcome bias compared to judgments of unethicality. Hence, the association between these judgements may explain the results we observe in this replication, both in terms of not observing an outcome bias for unethicality judgements and the lack of a mediation effect of unethicality.

Extensions to the Target Article

We extended the target study in two key ways. The first was adding a neutral condition in which outcome information was unknown to participants and testing whether decisions were judged as less ethical when the outcome was negative relative to this neutral condition. The second was adding scenarios to test whether the impact of outcome bias was affected by the scenarios.

For the first extension, we found evidence that ethical decisions were judged as more unethical, punishable and blameworthy when the outcome was negative relative to when outcome information was unknown. This provides strong evidence of an outcome bias when comparing ethical decisions with a negative outcome to those without a known outcome. We also found that ratings of unethicality and punishment (but not blame) were higher in the positive outcome condition than in the no outcome condition for Scenarios 1-3. This could be interpreted as blame being specifically associated with negative outcomes (as it seems intuitively unusual to blame an individual for a positive outcome), but actions can still be judged unfavorably in terms of being unethical or deserving of punishment even if they result in positive outcome.

For the second extension, we found that outcome bias was stronger in Scenarios 4-6. We can only speculate as to the differences between the scenarios, but one main difference was that in Scenarios 1-3 there was more changed in the manipulation than just the outcome, with differences in the circumstances described in the scenario, whereas in Scenarios 4-6 only the outcome was changed. We recommend using Scenarios 4-6 when studying outcome bias, and ensuring that only the outcome is manipulated across scenarios.

Constraints on Generality

We now examine the generality of our results so as to not overclaim about our results (Yarkoni, 2022). We recommend caution when interpreting the results of this research in a wider context. The scenarios used here show a range of possible situations that are ethically ambiguous, but are not entirely representative of all such situations. We had aimed to replicate the original paper's results in a more heterogeneous population than simply university students. The scenarios here may be seen as relatively contrived. However, future work can further explore different types of scenarios, especially those that are more familiar and quotidian to a layperson and with the range of scenarios used here, we feel that some generality can be argued from our results.

Implications, Limitations, and Future Directions

While the work of Baron and Hershey (1988) revealed that we may be biased by outcomes when judging the quality of decisions, the work from Gino et al (2009) extended these findings to the realm of ethical decisions. The results of this replication study suggests that the outcome bias in evaluations of ethical decisions is more nuanced than previously thought, as such evaluations are dependent on the manner in which decision makers are judged for their actions. In morally or ethically grey situations, we observed that judgements of a decision maker's ethicality were less consistent than judgements of how punishable and blameworthy a decision maker was. Part of these differences by dependent variable may stem from differences in the nature of these judgements. While punishment and blame might be more related to the need for accountability for the negative outcomes, the relationship between unethicality and outcome might be more related to the action than to the outcome, and involving perceptions of intent. Future work could hence further investigate the different ways in which individuals are judged in these morally grey situations to understand the nuances differentiating unethicality from blame and punishment.

This work adds to the findings of Baron and Hershey (1988) in their original study, further showing how biased individuals can be by the outcomes of decisions ceteris paribus. In our own replication of Baron and Hershey (Aiyer et al., 2023), we successfully replicated the original findings not only that decisions were judged more favorably when they resulted to successful outcomes relative to those resulting in negative outcomes, but also that individuals exhibited this bias even when they believed that outcomes should not factor into evaluations of decisions. The extent to which these findings hold within the realm of ethical decisions tells us whether these decision evaluations are similar in situations where the criteria for judging a decision as 'good' are more amorphous. By their very nature, ethical decisions are based around differing conceptions of what is right, leading to greater variability in how they may be evaluated from person to person. This variability may explain our observed difference in findings between ethicality and our other dependent variables: individuals agree more on what is punishable and blameworthy but than on what is ethical. Included within these individual differences could hence be different conceptions of whether outcome is used to evaluate decisions from an ethical perspective.

Our extension ties the outcome bias in ethicality judgments to other biases in the literature, namely hindsight bias (Fischhoff, 1975; Slovic & Fischhoff, 1977) and side-effect bias (Knobe, 2003). In outcome bias the contrast is between two known outcomes of different valence, positive vs. negative, that biases the evaluation of the decision. Hindsight bias is about the asymmetries in evaluations when contrasting a known outcome against an unknown outcome. Our design followed the design adopted by Chen et al. (2021)'s Study 3 that combined the experimental design of outcome bias and hindsight bias, showing that people exhibit outcome and hindsight bias regarding the replicability of the study on hindsight bias. Similarly, the side-effect effect (Knobe, 2003) manipulates helpful or harmful outcomes caused as a side-effect to show that people attribute different intentionality depending on the different outcomes, despite it not being directly related to the decision. We see much promise in future research that examines the interplay between outcome bias, hindsight bias, and the side-effect effect.

A limitation of our study that is inherited from the original study (as well as from Baron's and Hershey's (1988) seminal paper on outcome bias) is that participants are presented with hypothetical scenarios involving unnamed decision makers. Judgements of ethics can indeed be probed to some extent using hypothetical situations and thought experiments (e.g. the trolley problem) but this limits the extent to which we can claim that these results definitively capture the mechanisms of ethical judgements on a more day-to-day level. For example, it is unknown whether these results may extend to how individuals judge the ethicality of their own decisions rather than someone else's. Judgements of how we ourselves should act may be different to judgements of how other hypothetical individuals should act in the same situation. A question therefore for future research would be the extent to which outcome bias is either diminished or amplified when individuals judge their own decisions.

We also revisit the limitations in the original study design as discussed earlier (see section 'Experimental Manipulations in the Target Article'). Given that there are inconsistencies in the scenario texts between conditions outside of the intended 'insider knowledge' experimental manipulation, we cannot infer that this is what led to differences in the dependent variables between the conditions. We then compound this with our implementation of a control condition, whereby the wording of the scenarios in this condition resembles the negative outcome condition more than the positive outcome condition. For future studies, we would recommend that researchers keep any ancillary text constant apart from the changes in outcome or indeed any key manipulation. In addition, any extensions or follow-up studies (such as the adding of probability information in Scenarios 4-6) should be the same across all conditions. Finally, when depicting hypothetical scenarios such as these, rather than making assumptions about what participants perceive about the individuals in the scenarios, researchers should make clear what information is known to participants and what is known to the depicted individuals in the scenario (e.g., in the case of the probabilities of outcomes in Scenarios 4-6).

The tendency to evaluate decisions heavily based on the outcome may be over-generalized to inappropriate circumstances, leading to distorted judgments that could have profound implications on many aspects of life, including in legal and clinical settings. Future work can explore specific settings in more detail in order to further study how outcome bias may be lessened or amplified depending on social and contextual factors, rather than the use of hypothetical scenarios as in this study.

Finally, we note that this replication is relatively novel in the sense that it uses a unified design for multiple scenarios and is based on an unpublished target article. This is a notable situation given that despite the multiple rejections from journals for this paper, it has been widely cited as a preprint (up until its aforementioned removal) and has served as the basis for follow-up work. The use of replications for unpublished findings has been relatively underexplored. Such replications could be really beneficial for understanding the quality of unpublished yet influential research and better acknowledges that it is not only published research that informs scientific progress. With more such replications, we can better understand as a field why certain unpublished papers gain notoriety. We also hope that this replication aids the continuing efforts of the Many Co-Authors Project given the ongoing situation (at the time of writing) involving the target article's lead author.

Certain unpublished papers may be just as 'high quality' as their published counterparts and a replication process such as this can be useful for giving them the recognition that they perhaps should have gotten from journals. At the same time, replicating unpublished findings could also reveal where psychology can be deficient in giving certain papers undue credit. Preprints have been very beneficial to the field, but do lack the accreditation that a successful peer review affords. Whilst peer review has its own shortcomings as a process, it is perhaps preferable to having no such process applied to the paper at all. However, when a paper becomes widely cited, we can view it as a 'successful' crowdsourced review from the scientific community. The extent to which citations can be viewed as a reliable metric of scientific merit is of course up for debate: many papers are successfully published but not widely cited. Hence, there is still future work on evaluating the impact of a paper and hence which are most worth conducting a replication for. It could well be however that judging papers by whether they were published or not is a form of outcome bias in of itself.

Conclusion

We found that outcome information impacted judgments of blame and of punishment deserved. However, the impact of outcomes on judgments of unethicality was weaker and less consistent compared to the results of Gino, Moore, and Bazerman (2009). We conclude only partial support for outcome bias in ethical judgements. This was an unusual replication in that we attempted to replicate an unpublished yet impactful manuscript of an interesting phenomenon that deserves more attention. We hope to reignite interest in research on biases in unethicality, answering calls for more independent well-powered direct replications with insightful added extensions. We call for more research on the mechanisms of how outcome information affects judgment of unethicality, punishment, and blame.

Competing Interests

No conflicts of interest.

Funding

This research was supported by the European Association for Social Psychology seedcorn grant.

Author Contributions

Florence conducted the replication, an initial analysis of the paper, experimental design and extensions. She wrote the pre-registration, analyzed the data, and wrote the initial report as part of her thesis.

Gilad supervised the project, conducted the pre-registration, ran data collection, and guided the analyses and the initial report of the findings.

Sriraj followed up on initial work to verify analyses and conclusions, added advanced analyses, tables, and plots, and worked on an initial draft. Sriraj prepared the manuscript for journal submission. Gilad and Sriraj jointly finalized the manuscript for submission and handled revisions.

Citation of the Target Research Article

Gino, F., Moore, D. A., & Bazerman, M. H. (2009). No Harm, No Foul: The Outcome Bias in Ethical Judgments.

Contributor Roles Taxonomy

Role	Sriraj Aiyer	Wing Yan (Florence) Chan	Gilad Feldman
Conceptualization			X
Pre-registrations		X	X
Data curation		X	
Formal analysis	X	X	
Funding acquisition			X
Investigation		X	X
Methodology		X	X
Pre-registration peer review / verification		X	X
Data analysis peer review / verification	X	X	
Project administration			X
Resources			X
Supervision			X
Validation	X		
Visualization	X		
Writing-original draft	X	X	
Writing-review and editing	X		X

Submitted: July 18, 2024 PST, Accepted: October 08, 2024 PST

References

Aiyer, S., Kam, H. C., Ng, K. Y., Young, N. A., Shi, J., & Feldman, G. (2023). Outcomes Affect Evaluations of Decision Quality: Replication and Extensions of Baron and Hershey's (1988) Outcome Bias Experiment 1. *International Review of Social Psychology*. https://doi.org/10.5334/irsp.751

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569. https://doi.org/10.1037//0022-3514.54.4.569

Bazerman, M. H. (2022a). *Behavioral Science Authors Series by Behavior Change For Good Initiative - lecture by Max Bazerman*. https://www.youtube.com/watch?v=AndOnc5c0nk&t=1002s

Bazerman, M. H. (2022b). *Complicit: How we enable the unethical and how to stop*. Princeton University Press.

Chen, J., Kwan, L. C., Ma, L. Y., Choi, H. Y., Lo, Y. C., Au, S. Y., ... Feldman, G. (2021). Retrospective and prospective hindsight bias: Replications and extensions of Fischhoff (1975) and Slovic and Fischhoff (1977). *Journal of Experimental Social Psychology*, *96*, 104154. https://doi.org/10.1016/j.jesp.2021.104154

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Emerson, G. B., Warme, W. J., Wolf, F. M., Heckman, J. D., Brand, R. A., & Leopold, S. S. (2010). Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. *Archives of Internal Medicine*, *170*(21), 1934–1939. https://doi.org/10.1001/archinternmed.2010.406

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feldman, G. (2023). *Registered Report Stage 1 manuscript template*. https://doi.org/10.17605/OSF.IO/YQXTP

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288.

Gino, F., Moore, D. A., & Bazerman, M. H. (2009). *No Harm, No Foul: The Outcome Bias in Ethical Judgments*. https://doi.org/10.2139/ssrn.1099464

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless+ harmless= blameless: When seemingly irrelevant factors influence judgment of (un) ethical behavior. *Organizational Behavior and Human Decision Processes*, *111*(2), 93–101. https://doi.org/10.1016/j.obhdp.2009.11.001

Jané, M., Xiao, Q., Yeung, S., Ben-Shachar, M. S., Caldwell, A., Cousineau, D., Dunleavy, D. J., Elsherif, M., Johnson, B., Moreau, D., Riesthuis, P., Röseler, L., Steele, J., Vieira, F., Zloteanu, M., & Feldman, G. (2024). *Guide to Effect Sizes and Confidence Intervals*. https://doi.org/10.17605/OSF.IO/D8C4G

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194. https://doi.org/10.1093/analys/63.3.190

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, *3*. https://doi.org/10.15626/MP.2018.843

Lench, H. C., Domsky, D., Smallman, R., & Darbor, K. E. (2015). Beliefs in moral luck: When and why blame hinges on luck. *British Journal of Psychology*, *106*(2), 272–287. https://doi.org/10.1111/bjop.12072

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442.

Mackie, D. M., & Ahn, M. N. (1998). Ingroup and outgroup inferences: When ingroup bias overwhelms outcome bias. *European Journal of Social Psychology*, *28*(3), 343–360. https://doi.org/10.1002/(SICI)1099-0992(199805/06)28:3%3C343::AID-EJSP863%3E3.0.CO;2-U

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Many Co-Authors. (2024). https://manycoauthors.org/

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Savani, K., & King, D. (2015). Perceiving outcomes as determined by external forces: The role of event construal in attenuating the outcome bias. *Organizational Behavior and Human Decision Processes*, *130*, 136–146. https://doi.org/10.1016/j.obhdp.2015.05.002

Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 544.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). *Mediation: R package for causal mediation analysis*. http://hdl.handle.net/1721.1/91154

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*. https://doi.org/10.1017/S0140525X20001685

Supplementary Materials

Supplemental Material

Download: https://collabra.scholasticahq.com/article/126266-outcome-bias-in-evaluations-of-ethical-decisions-replication-and-extensions-of-gino-moore-and-bazerman-2009/attachment/254727.docx?auth_token=aRq9tyQYIjJb2GhjQk5D

Peer Review Communication

Download: https://collabra.scholasticahq.com/article/126266-outcome-bias-in-evaluations-of-ethical-decisions-replication-and-extensions-of-gino-moore-and-bazerman-2009/attachment/254728.docx?auth_token=aRq9tyQYIjJb2GhjQk5D

Response Letter

Download: https://collabra.scholasticahq.com/article/126266-outcome-bias-in-evaluations-of-ethical-decisions-replication-and-extensions-of-gino-moore-and-bazerman-2009/attachment/254729.pdf?auth_token=aRq9tyQYIjJb2GhjQk5D

Outcome bias in evaluations of ethical decisions:

Replication and extensions of Gino, Moore, and Bazerman (2009)

Supplementary

Contents

Analysis of the Original research

Methods of the original research

Tables S1 presents the content and procedures of study 1 in the original research.

Table S1

Content and procedures of the study 1 in the original research

Scenario	Positive-outcome condition	Negative-outcome condition
Study 1's Scenario 1: Pharmaceutical researcher	A pharmaceutical researcher defines a clear protocol for determining whether or not to include clinical patients as data points in a study. He is running short of time to collect sufficient data points for his study within an important budgetary cycle within his firm. He believed that the product was safe and effective. As the deadline approaches, he notices that if he had four more data points for how subjects are likely to behave, the analysis would be significant. He makes up these data points, and soon the drug goes to market. This drug is a profitable and effective drug, and years later shows no significant side effects.	A pharmaceutical researcher defines a clear protocol for determining whether or not to include clinical patients as data points in a study. He is running short of time to collect sufficient data points for his study within an important budgetary cycle within his firm. As the deadline approaches, he notices that four subjects were withdrawn from the analysis due to technicalities. He believes that the data in fact is appropriate to use, and when he adds those data points, the results move from not quite statistically significant to significant. He adds these data points, and soon the drug goes to market. This drug is later withdrawn from the market after it kills six patients and injures hundreds of others.
Study 1's Scenario 2: Auditor	An auditor is examining the books of an important client, a client that is not only valuable for their auditing fees, but also buys lucrative advisory services from the auditor's firm as well. The auditor notices clearly fraudulent practices by their client. The auditor brings up the issue with the client, who insists that there is nothing wrong with their accounting. The client also threatens to withdraw their business if the auditor withholds their approval. The auditor agrees to let it go by for one year, and encourages the client to change their accounting practices over the next year. No problems result from the auditor's decision.	An auditor is examining the books of an important client, a client that is not only valuable for their auditing fees, but also buys lucrative advisory services from the auditor's firm as well. The auditor notices some accounting practices that are probably illegal, but it would take multiple court cases to be sure about whether the action was legal or not. The auditor brings up the issue with the client, who insists that there is nothing wrong with their accounting. The client also threatens to withdraw their business if the auditor withholds their approval. The auditor agrees to let it go by for one year, and encourages the client to change their accounting practices over the next year. Six months later, it is found that the client was committing fraud, their corporation goes bankrupt, the bankruptcy is connected to the issue that the auditor noticed, and 1,400 people lose their jobs and their life's savings.
Study 1's Scenario 3: Toy company	A toy company sells products made by another firm, manufactured in another country. The toy company knows that the toys contain lead, which can be extremely hazardous to children. The toy company successfully sells this product, makes a significant product, and no children are injured by the lead paint.	A toy company finds out that the products that they were selling, manufactured by another firm in another country, contains lead, which can be extremely hazardous to children. The toy company had failed to test for lead in the product, since testing is expensive and is not required by U.S. law. The lead paint eventually kills 6 children, and sends dozens more to emergency room for painful treatment for lead poisoning.
Study 2's Scenario 4: Sewage treatment plant	A sewage treatment plant is undergoing remodeling and updating. There is a critical phase to the project during which all the treatment systems are shut off and incoming sewage is diverted to a holding tank until the new systems are activated. This critical phase of the project will last 48 hours. If there should be substantial rainfall during this 48-hour period, there is a high probability that the holding tanks will overflow into local waterways, with serious negative environmental and health consequences for the wildlife and people who live in the area. The company that runs the sewage treatment plant could invest in back-up systems that would eliminate the risk of overflow even in the case of heavy rain, but this would be expensive.	A sewage treatment plant is undergoing remodeling and updating. There is a critical phase to the project during which all the treatment systems are shut off and incoming sewage is diverted to a holding tank until the new systems are activated. This critical phase of the project will last 48 hours. If there should be substantial rainfall during this 48-hour period, there is a high probability that the holding tanks will overflow into local waterways, with serious negative environmental and health consequences for the wildlife and people who live in the area. The company that runs the sewage treatment plant could invest in back-up systems that would eliminate the risk of overflow even in the case of heavy rain, but this would be expensive. Historically, the chances of rain during the planned 48 hour period are about 5%. After extensive

Scenario	Positive-outcome condition	Negative-outcome condition
	Historically, the chances of rain during the planned 48 hour period are about 10%. After extensive consideration, the manager in charge of the remodeling decides against instituting the back-up plan. In fact, there is no rain during the critical 48-hour period and the plant remodeling is a complete success.	consideration, the manager in charge of the remodeling decides against instituting the back-up plan. It winds up raining a great deal. Twenty people fall ill and wind up in the hospital, many fish die, the water is unsafe for swimming for a week, and fishers are discouraged from eating fish caught downstream.
Study 2's Scenario 5: Natural disaster	A government agency in a developing country finds itself dealing with a natural disaster in which several thousand poor peasants have been made homeless during the winter. The agency must decide what sort of short-term housing it will provide for the refugees. The inexpensive option is tents, which will probably be fine, given the mildness of the local winters—overnight temperatures only fall below freezing once every two years or so, on average. The more expensive option is to put up temporary shacks that would provide more shelter against the cold. But they shacks would be more expensive and would force the agency to cut funding to other (less urgent) programs. In the end, the agency's commissioner decides to provide only tents for the refugees. The winter is quite mild, and the tents provide sufficient shelter.	A government agency in a developing country finds itself dealing with a natural disaster in which several thousand poor peasants have been made homeless during the winter. The agency must decide what sort of short-term housing it will provide for the refugees. The inexpensive option is tents, which will probably be fine, given the mildness of the local winters—overnight temperatures only fall below freezing once every four years or so, on average. The more expensive option is to put up temporary shacks that would provide more shelter against the cold. But they shacks would be more expensive and would force the agency to cut funding to other (less urgent) programs. In the end, the agency's commissioner decides to provide only tents for the refugees. The winter is substantially colder than expected, and fifty children among the refugees die of exposure to the cold.
Study 2's Scenario 6: Water supply	There is a river that runs through dry regions in Mexico. Eighty percent of the time, there has been plenty of water to supply the communities that depend on water from the river. So when the mayor of a prosperous town that drew its water from the river was asked to invest in water conservation measures, she assumed they were unnecessary. It so happened that this town was upstream from most of the other communities that depended on the river. Upstream communities naturally have the advantage because they can take what they need first, and downstream communities left without water have little recourse. In the end, there was plenty of rain and more than enough water for the communities along the river.	There is a river that runs through dry regions in Mexico. Ninety percent of the time, there has been plenty of water to supply the communities that depend on water from the river. So when the mayor of a prosperous town that drew its water from the river was asked to invest in water conservation measures, she assumed they were unnecessary. It so happened that this town was upstream from most of the other communities that depended on the river. Upstream communities naturally have the advantage because they can take what they need first, and downstream communities left without water have little recourse. One year, rainfall is far below expectations and 46 small farms are driven out of business when the lack of water in the river leaves them unable to irrigate their land.

Note. Imported from the Appendix in the original research.

Hypotheses and results of the original research

For Hypothesis 1 (people judge a questionable decision as more unethical when the outcome is negative rather than positive), the judgments of unethicality, punishment, and blame were subject to a mixed-ANOVA in which outcome information (positive, negative) served as between-subject factors. Supporting Hypothesis 1, a medium effect of outcome information on ethical judgment ($\eta^2 = .059$, 95% CI [.004, .16]) was found in the original research.

Supporting Hypothesis 2a (ethical judgment of a decision mediates the relationship between outcome information and judgment of punishment deserved for that decision), the path between outcome information and judgments of punishment deserved became insignificant ($\beta = -.09$, $p = .52$) when the effect of ethical judgment was included in the regression ($\beta = .75$, $p < .001$), suggesting that ethical judgment fully mediated the relationship between outcome information and judgement on punishment deserved (Sobel test Z = -2.59, $p = .009$). See Figure S1 for the depiction of the mediation effect in Hypothesis 2a in the original research.

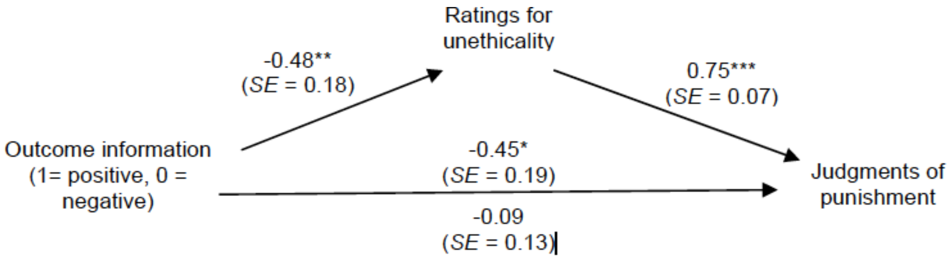


Figure S1. A diagram showing judgment of unethicality as the mediator between judgment of punishment and outcome information extracted from Gino et al. (2009) Page 49

Supporting Hypothesis 2b (ethical judgment of a decision mediates the relationship between outcome information and judgment of blame deserved for that decision), the path between outcome information and judgments of punishment deserved became less significant ($\beta = -.57$, $p = .004$) when the effect of ethical judgment was included in the regression ($\beta = .57$, $p < .001$), suggesting that ethical judgment partially mediate the relationship between outcome and judgement on blame deserved (Sobel test Z = -2.42, *p* = .02). Please see Figure S2 for the depiction of the mediation effect in Hypothesis 2b in the original research.

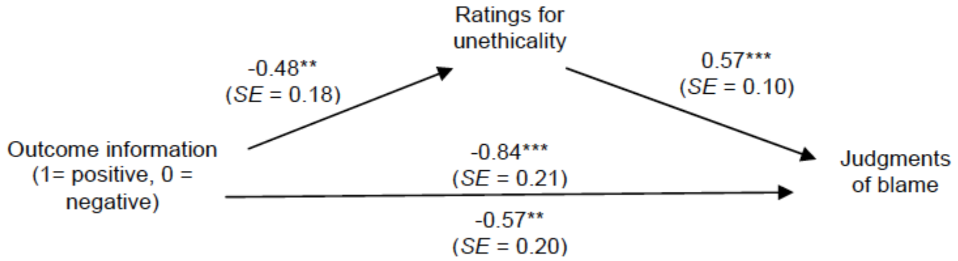


Figure S2. A diagram showing judgment of unethicality as the mediator between outcome information and judgment of blame extracted from Gino et al. (2009) Page 49

Table S2 summaries the results of the original research. The judgments of unethicality, punishment, and blame were subject to a mixed-ANOVA in which outcome information (positive, negative) served as between-subject factors.

Table S2
Gino et al. (2009) Study 1: Findings

IVs	DVs	*M*	*SD*	*F*	*df*	*p*	η^2	95%CI
Positive outcome	Unethicality	5.28	1.13	7.35	1,118	.008	.059	004,.16
Negative outcome		5.77	0.79					
Positive outcome	Punishment	5.01	1.10	5.81	1,118	.017	.047	001,.14
Negative outcome		5.46	0.94					
Positive outcome	Blame	4.65	1.40	15.6	1,118	<.001	.117	.03,.23
Negative outcome		5.49	0.90					

Note. N/A = not provided in the original research.

Power analysis

The required sample size is 398 for study 1 of the original research and extensions with 0.95 power and 0.05 alpha.

We chose the effect of outcome information on the judgment of punishment deserved as the basis of our calculation for the required sample size because it is the smallest effect among the three major effects of outcome bias found in Study 1 of the original research. The calculated required sample size with 0.95 power and 0.05 alpha is 265. However, as there were only two groups (positive and negative outcome conditions) in the original research but we added one group (no-outcome condition) in this replication, we needed a total of 398 participants as only two-third of them were guided into the positive and negative outcome conditions.

Table S2 shows the calculations for the major effect sizes and the calculations for the required sample sizes.

Table S2

Effect sizes calculations of major effects found in study 1 of the original research

	η^2	95%CI	Cohen's *f*	Required sample size with 0.95 power and 0.05 alpha
Effect of outcome information on judgement of unethicality	.059	.004,.16	^[1]0.250	^[4]209
Effect of outcome information on judgement of punishment deserved	.047	.001,.14	^[2]0.222	^[5]265
Effect of outcome information on judgement of blame deserved	.117	0.03,.23	^[3]0.364	^[6]100

Notes. ^[1-6] Please see Table 7 for details of calculations using online calculators and R.

Generalized exclusion criteria

The default generalized exclusion criteria we use in our pre-registration is the following:

> "We will focus on our analyses on the full sample. However, as a supplementary analysis and to examine any potential issues, we will also determine further findings reports with exclusions. In any case, we will report exclusions in detail with results for the full sample and results following exclusions (in either the manuscript or the supplementary).

We excluded 7 participants based on the following criteria (the number in **[square brackets]** indicates the number of participants who met this criteria):

1. Participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale) **[2]**

2. Participants who self-report not being serious about filling in the survey (self-report < 4, on a 1-5 scale). **[1]**

3. Participants who correctly guessed the hypothesis of this study in the funnelling section. **[2]**

4. Participants who have already seen or done the survey before. **[1]**

5. Participants who failed to complete the survey. (duration = 0, leave question blank) **[NA]**

6. (When target sample is online:) Participants not from the United States. **[2]**

The results from our main manuscript do not change with these exclusions.
Filename "knit-no-exclude-vs-exclude.R" allows running both with and without exclusions.
The knitted output analysis file is available on the OSF:
Gino-etal-2009-outcome-bias-ethical-vX-X-Exclusions.html

Comparisons and deviations

Original versus replication

Table S3

Comparisons between the original research and replication

	Original	Replication	Reason for change
Study design	repeated-measure between-subject design		N/A
Procedure	A pilot was done to pre-test the scenarios. Then in the main study, participants were randomly assigned to one of the conditions, read a few scenarios and then rate on unethicality, punishment and blame deserved.	No pilot was done. Participants were randomly assigned to one of the conditions, read a few scenarios and then rate on unethicality, punishment and blame deserved.	We included a "no outcome" condition to address the lack of pre-test.

	Original	Replication	Reason for change
Conditions	2 conditions (positive outcome vs negative outcome)	3 conditions (positive outcome, negative outcome, no outcome)	As we did not pre-test the scenarios, adding a no outcome condition can shed light on whether people really do judge the ethicality of decisions similarly when there is no outcome information, which served as a control condition.
Analyses for hypotheses 2a and 2b	Mediation analysis in Baron and Kenny (1986)	Mediate function in R's psych package	To conduct more robust and up-to-date mediation analyses
Comprehension checks	None reported	2 comprehension questions were added after each scenario	To ensure participants do grasp the main ideas conveyed in the scenarios and increase reliability of results
Content of scenario 1 - pharmaceutical researcher	A pharmaceutical researcher defines a clear protocol for determining whether or not to include clinical patients as data points in a study. He is running short of time to collect sufficient data points for his study within an important budgetary cycle within his firm. He believed that the product was safe and effective. As the deadline approaches, he notices that if he had four more data points for how subjects are likely to behave, the analysis would be significant. He makes up these data points, and soon the drug goes to market. This drug is a profitable and effective drug, and years later shows no significant side effects.	A pharmaceutical researcher defines a clear protocol for determining whether or not to include clinical patients as data points in a study. He is running short of time to collect sufficient data points for his study within an important budgetary cycle within his firm. He believed that the product was safe and effective. As the deadline approaches, he notices that if he had four more data points for how subjects are likely to behave, the analysis would be significant **(i.e. the results would support the effectiveness of the product)**. He makes up these data points, and soon the drug goes to market. This drug is a profitable and effective drug, and years later shows no significant side effects.	Added a simple explanation (please refer to the **bolded** text) to facilitate understanding of statistical terminology
Question on judgment of blame deserved	"How much would you blame the ... behavior **for the observed outcome**?"	"How much would you blame the ... behavior ?"	We added a "no-outcome" condition. We deleted "for the observed outcome" so that we could use the same question across all conditions
Physical setting	Recruited undergraduate students from a university in the United States	Recruited participants in the United States on an online platform (the Amazon Mechanical Turk)	To increase generalizability

Note. Major changes in the texts are bolded.

Pre-registration plan versus final report

Component	**Location**	**Were there deviations? What type?**	**Details of deviation(s)**	**Rationale for deviation**	**How might the results be different if you had/had not deviated**	**Date/time of decision for deviation**	**Any additional notes**
Analyses	p.35	Minor	Additional analyses involving all scenarios 1-6	To provide an additional overall analysis on all scenarios used, and an additional mediation analysis using Scenarios 4-6	No difference	20 May 2020, on the day of data collection	N/A

Peer Review and Communication History

MS Title: Outcome Bias in Evaluations of Ethical Decisions: Replication and Extensions of Gino, Moore, and Bazerman (2009)

Author Names: Sriraj Aiyer, Wing Yan (Florence) Chan, Gilad Feldman

Submitted: Jul 18, 2024

Editor First Decision: Revise & Resubmit
Sep 26, 2024

Dear Gilad Feldman,
I have now received two thorough expert reviews of your manuscript (which are attached again), "Outcome bias in evaluations of ethical decisions: Replication and extensions of Gino, Moore, and Bazerman (2009)", from experts in the field. I also independently read the manuscript before consulting these reviews and after having gone through bot reviews in detail.
The reviewers had mostly very positive reactions to your manuscript and my assessment mirrors this fully. In this regard, I want to especially highlight how incredibly extensive the supplementary materials were, which I very much appreciated. I think the manuscript makes a clear, worthwhile contribution.
Both reviewers still had a few remaining issues they noted; some very minor ones about the writing etc., but also a bit more substantial ones on providing more context and outline to interpret the data collection and results. I have nothing specifically to add to these issue notes.
In my eyes, this substantiates a minor revision. Hence, I highly encourage you to address the reviewers' comments and submit a revised version for further consideration at Collabra: Psychology.
In your resubmission, please include a document with a point-by-point response to the reviewers' comments, outlining each change made in your manuscript or providing a suitable rebuttal; also note any other changes you made in the manuscript, that are not necessarily from addressing the reviewers' comments and explain why the changes are meaningful. Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all necessary copyright permissions have been obtained. This may be the last opportunity for major editing, therefore please fully check your file prior to re-submission.
If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.
We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.
Sincerely,
David Grüning
Reviewer 1
Rating scale questions

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript)				✓	
The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript)				✓	
The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.)	✓				
The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.)		✓			

Open response questions

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

This is a review of the manuscript “Outcome bias in unethicality: Replication and extension of Gino et al. (2009)”. The manuscript reports a replication of two vignette experiments on outcome bias – the effect that acts are judged as less unethical when they lead to positive rather than negative outcomes.

The original studies have several severe flaws – the vignettes in different treatment vary in content beyond the experimental manipulation of the treatment variable. This means that the experimental treatment does not cleanly identify the effect of interest. The present manuscript replicates these studies in their original design. I must say that I am no fan of replications of bad research – if we cannot learn anything from the original research, I don't see how more of it would be better. That said, the authors make a case for this decision, and as a replication, this manuscript is sound apart from some minor issues I enumerated below.

Scope of this review

I read the manuscript and supplemental materials. I also looked at the posted materials, but because the format of the Qualtrics PDF export appears broken, it was hard to investigate the materials fully.

I also checked out the replication files. Note that these files were not included in the OSF preregistration and therefore weren't 'frozen'. However, the files were uploaded on the same day the preregistration was registered. The format of the preregistration made it somewhat difficult to identify decision criteria, but I spotted no undisclosed deviations in the final manuscript.

Since I do not have a Jamovi installation, I did not attempt to replicate the statistical analyses.

Minor issues

p. 1: "positive outcomes were rated" -> are rated

p. 1: "Gino, Moore, and Bazerman (1988)" -> 2009

p. 1: check whether/where zeroes go before the decimal

p. 1: eta^2G should be eta^2_G

p. 8: "Secondly, we extended…" this is repeated in the next paragraph

p. 8: "and examine how positive and negative…" I found this and the following sentence confusingly written

p. 9: "it is unclear from the scenario texts where" -> whether

p. 10: "we provided a summary" -> provide

p. 10: "power analysis of the effect on outcome" -> of outcome

p. 10: "scenarios from Study," -> Study 1

p. 11: "scenario in the replication group" -> pharmaceutical researcher scenario?

p. 11: The excerpt is described as "unchanged in all three conditions", but from Table 3, it differs between the positive and negative treatments.

p. 23: I am not sure I fully grasped the statistical model for the mediation analyses. Is this a mixed-effects model using data from all scenarios? If so, the random effects structure should be mentioned.

p. 23: Relatedly, it would be nice to also report the sensitivity analysis which is provided by the mediation package.

p. 23: Is there a reason not to also report the difference between the positive condition and the no-outcome condition? To me, it seems most sensible to decompose the overall effect into effects of positive and negative outcomes (while acknowledging differences in the phrasing of the vignettes).

p. 24: "ratings on unethicality […] was higher" -> were

p. 28: "Hence, whilst participants judged" This sentence is incomplete

p. 34: "feelings of blame was" -> were

p. 36: "we successfully the original" -> incomplete sentence

I always sign my reviews.

Simon Columbus | MIT
simon@simoncolumbus.com

Reviewer 2

Rating scale questions

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript)				✓	
The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript)					✓
The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.)			✓		
The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.)				✓	

Open response questions

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

Major comments:

I had difficulties finding the actual preregistration of the replication, even though the authors have a “pre-registration” folder in their OSF page. I would suggest to create and refer to the doi of the document in the text so it is easy for readers to access it.

When I read the abstract, introduction, and the results section, I wasn’t entirely sure what was meant by hypotheses 2a & 2b and 5a & 5b. The authors might consider clarifying these predictions a bit further, perhaps by providing an example or offering more detailed explanations. On a related note, I am also missing a detailed justification for adding the mediation analysis and Scenarios 4–6 as an extension to the current replication.

A common criticism of replications based on online samples is that the quality of online testing platforms (such as MTurk) has declined in recent years, and original authors raised concerns about the reliability of (replication) data collected through these platforms.

I am curious whether any additional manipulation checks or quality control measures were performed besides the comprehension checks?

As I understand it, the scenarios presented were classified as ‘grey-area.’ How was this classification determined, and how did the original authors arrive at it? It is crucial for the replication to demonstrate that participants perceive the scenarios as grey-area to the same extent as those in the original study. If no additional manipulation checks were conducted, I would suggest including a paragraph that justifies the adequacy of the current sample’s quality.

Related to that I was wondering what the outcome of the comprehension checks was? This information is crucial for readers to judge the data quality, so I would recommend adding a table showing the outcomes per scenario. If certain individuals show high rates of non-comprehension, I suggest to add an additional exploratory analysis excluding these participants. These can be added to the supplementary material. It makes a strong case for replication critics if the results are robust when retaining only the “highest” quality data points.

Something else that I found interesting and worth to discuss is the fact that ratings of unethicality and punishment (but not blame) were higher in the positive outcome condition than in the no outcome condition for scenarios 1-3. That may be worthwhile to add in the discussion (as interesting exploratory pattern). Do you have any explanations for this finding?

Minor comments:

I really like how the tables clearly state the hypotheses and highlight which aspects of the original studies were the same, similar, or different. However, Table 3, which describes the scenario, might be better suited as an appendix to the manuscript. While it is a valuable addition to both the manuscript and the methods section, it somewhat disrupts the overall flow.

In the introduction, the fraud allegations against Gino are not mentioned, which makes it difficult for the reader to understand what is meant by ‘raised concerns regarding work conducted by the lead author.’ If these allegations are a reason for conducting these replications, it should be clearly stated. Since nothing has been proven yet, you should, of course, address the accusations carefully. Additionally, why was the follow-up article flagged as a concern for replicability? Is there a concern about data manipulation, or have there been unsuccessful replications of this article?

In my opinion, simply stating the number of citations of the preprint as evidence of its impact in the field is insufficient. Why was this paper cited, and how have its findings been applied? What specific impact has it had, and what follow-up research has been conducted based on it?

I didn’t understand the following sentences: ‘Our pre-registration stipulated that we intended to analyze all data without any exclusions and then apply exclusions as a supplementary analysis. Seven participants met the pre-registered supplementary exclusion criteria.’ What exactly were the pre-registered supplementary exclusion criteria?

You mention the limitations related to the experimental manipulations in the target article. I’m curious whether follow-up research by the original authors faced the same limitations or if they maintained the ancillary text consistently in their subsequent work. If they did keep the text constant, the results from this follow-up research could be compared to those from the original study and the replication, with the caveat that these articles have been flagged.

An interesting point worth discussing is that ratings of unethicality and punishment (but not blame) were higher in the positive outcome condition compared to the no outcome condition for scenarios 1-3. It might be valuable to include this as an interesting exploratory pattern in the discussion. Do you have any explanations for this finding?

Author Response
Oct 4, 2024

See a supplemental file “2458234-reply-to-decision-letter-annex.pdf”

Editor Final Decision: Accept
Oct 8, 2024

Dear Dr. Feldman,

I have now had a chance to read over your revised manuscript "Outcome bias in evaluations of ethical decisions: Replication and extensions of Gino, Moore, and Bazerman (2009)”, along with the response letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and I raised, especially clearing up all technical questions regarding the supplementary materials.

I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point.

You will be receiving separate correspondence regarding any production and technical comments, data deposits, as well as publication charges. We work with the Copyright Clearance Center to process any applicable APC charges.

You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely,
David Grüning

Reply to Collabra decision letter reviews: Gino et al. (2009) replication

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response to each item. We also provide a summary table of changes. Please note that the editor's and reviewers' comments are in bold with our reply underneath in normal script.

A track-changes manuscript is provided with the file: Collbara-RNR-Gino-etal-2009-replication-extensions-main-manuscript-v3-G-trackchanges .docx

Summary of changes

Section	Actions taken in the current manuscript
General	We addressed the minor grammatical issues throughout the manuscript raised by Professor Columbus.
Introduction	We added context from the original authors regarding their implementation of the manipulation and our own explanation around keeping the materials the same for replication purposes. We added a brief explanation of our choice of target article to address comments from Reviewer 2.
Methods	We addressed vignette text changing between outcome conditions when describing the methods. We added details on our participant recruitment process and manipulation checks to address comments from Reviewer 2 on data quality.
Results	No changes.
Discussion	We addressed Reviewer 2's point on an exploratory finding of differences between the positive and no outcome conditions.
Reporting	No changes.
Supplementary materials	We added explanations of our materials on OSF to address comments about the preregistration. We added more information on our exclusion criteria used for supplementary analysis, and updated our Rmarkdown output code and uploaded outputs with and without exclusions.

Reply to Editor: Dr./Prof. David Grüning

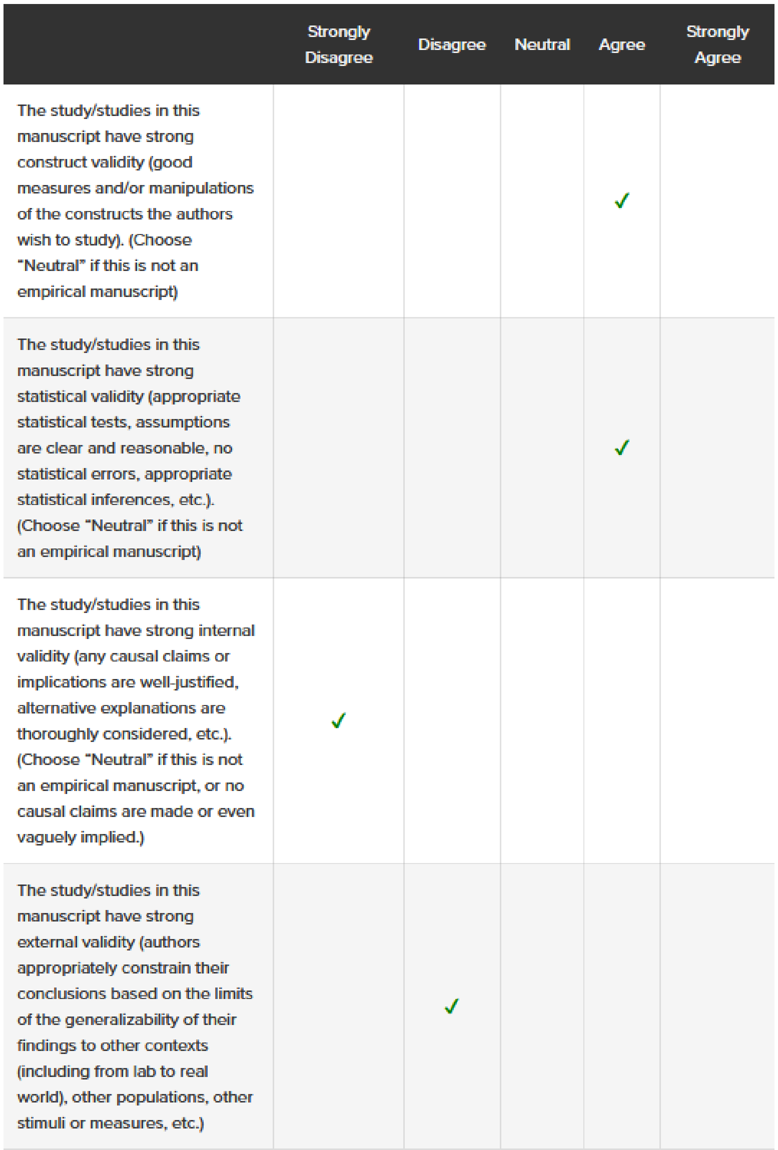
> **I have now received two thorough expert reviews of your manuscript (which are attached again), "Outcome bias in evaluations of ethical decisions: Replication and extensions of Gino, Moore, and Bazerman (2009)", from experts in the field. I also independently read the manuscript before consulting these reviews and after having gone through bot reviews in detail.**
>
> **The reviewers had mostly very positive reactions to your manuscript and my assessment mirrors this fully. In this regard, I want to especially highlight how incredibly extensive the supplementary materials were, which I very much appreciated. I think the manuscript makes a clear, worthwhile contribution.**

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit.

> **Both reviewers still had a few remaining issues they noted; some very minor ones about the writing etc., but also a bit more substantial ones on providing more context and outline to interpret the data collection and results. I have nothing specifically to add to these issue notes. In my eyes, this substantiates a minor revision. Hence, I highly encourage you to address the reviewers' comments and submit a revised version for further consideration at Collabra: Psychology.**
>
> **In your resubmission, please include a document with a point-by-point response to the reviewers' comments, outlining each change made in your manuscript or providing a suitable rebuttal; also note any other changes you made in the manuscript, that are not necessarily from addressing the reviewers' comments and explain why the changes are meaningful. Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all necessary copyright permissions have been obtained. This may be the last opportunity for major editing, therefore please fully check your file prior to re-submission.**

We respond to each point below. In summary, we added context regarding the design of the original study, explained some of the comments on our preregistration on OSF, and detailed some justifications for our research decisions.

Reply to Reviewer #1: Dr./Prof. Simon Columbus



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose "Neutral" if this is not an empirical manuscript)				✓	
The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose "Neutral" if this is not an empirical manuscript)				✓	
The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose "Neutral" if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.)	✓				
The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.)		✓			

This is a review of the manuscript "Outcome bias in unethicality: Replication and extension of Gino et al. (2009)". The manuscript reports a replication of two vignette experiments on outcome bias – the effect that acts are judged as less unethical when they lead to positive rather than negative outcomes.

The original studies have several severe flaws – the vignettes in different treatment vary in content beyond the experimental manipulation of the treatment variable. This means that the experimental treatment does not cleanly identify the effect of interest. The present manuscript replicates

> **these studies in their original design. I must say that I am no fan of replications of bad research – if we cannot learn anything from the original research, I don't see how more of it would be better. That said, the authors make a case for this decision, and as a replication, this manuscript is sound apart from some minor issues I enumerated below.**

Thank you for the positive and supportive opening note and the constructive feedback. We agree and wrote that there are issues with the original design, particularly around differences in the scenarios' text between conditions outside of the key manipulation of outcome.We marked these differences in Table 3, and used bolding to aid participants in focusing on the key details for each scenario.

We should note that the original authors wrote the following explanation regarding their manipulation choices in Scenarios 1-3 in Study 1:

> "In creating the materials, we matched a less unethical action with the negative outcome, and a more unethical action with a positive action, in order to test whether the outcome would overcome the ethicality of the actual decision. In the positive-outcome condition, the descriptions included elements of unethical or questionable practices but a positive outcome…In the negative-outcome condition, instead, the descriptions of unethical or questionable practices were less extreme from an ethical standpoint, but the outcome for each scenario was negative."

This partly explains the reason for the differences in the wordings. It is tricky to gauge whether this achieved the desired goal and whether it was needed at the cost of cleaner manipulation of outcome. Given that the authors designed the study with this justification (and ran a pilot study to test it), for the replication it made more sense to keep the materials the same for a more direct replication. These are ultimately issues with the target preprint and we believe there is value in a simple replication and alerting researchers to this issue so that future studies could try and test for that or address it.

We added this context to the Experimental Manipulations in Target Article section:

> In replicating this study, we note limitations with the manipulation in the target article. Whilst we did not alter the survey materials originally used (see Table 3 below), we followed the original authors' design in manipulating more of the wording than needed in the scenario texts. In Scenarios 1-3, the authors manipulated other aspects of the scenario apart from the outcome information, explaining their decision that using less unethical actions in the negative outcome condition to "test whether the outcome would overcome the ethicality of the actual decision". In Scenarios 4-6, the authors attempted to emphasize to participants that the outcomes of the depicted decisions were simply a result

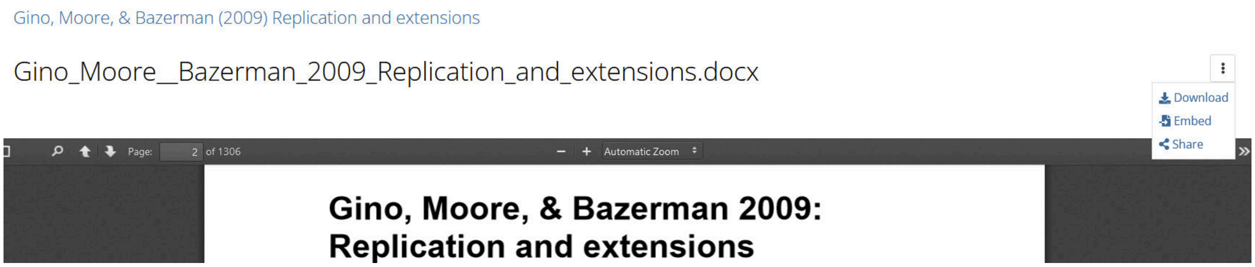
of probabilities (and hence out of control of the decision makers). This is framed in the target article as decision makers having a lack of 'inside knowledge'. However, it is unclear from the scenario texts whether these probabilities are known to the decision makers and the probabilities are not kept constant between the positive and negative outcome conditions, in line with the above justification of depicting less unethical actions in the negative outcome condition. The original authors reported verifying that these were indeed interpreted as less unethical via a pilot study. We also note that because Scenarios 4-6 are different situations to Scenarios 1-3, differences between the two groups of scenarios/studies cannot be isolated to be about 'inside knowledge' and could instead be due to differences in the scenarios. While the changes to the ancillary text outside of the manipulation of outcome is explained by the original authors, we feel it important to note the lack of consistency between experimental conditions as a limitation, as it makes the manipulation of outcome less clear. We did not change the original survey materials, and designed our control (with no outcome) condition extension to more closely resembe the negative condition in terms of wording.

.1. I read the manuscript and supplemental materials. I also looked at the posted materials, but because the format of the Qualtrics PDF export appears broken, it was hard to investigate the materials fully.

The Qualtrics exported file is actually a Word document, and not a PDF, and so we believe that this reflects a misunderstanding regarding the use of the OSF, which tries to automatically convert Word DOCX files to PDFs for display, and gets it very wrong when it comes to Qualtrics exports.

Therefore, if we visit the frozen pre-registered Qualrics survey Word DOC file on https://osf.io/gf8wd you should be able to download it and open the Word file. Additionally, you can import the .QSF file to any Qualtrics account, including the free version (which only limits data collection, but allows survey viewing), and browse everything about the survey.

Screenshot (click Download on the menu with three dots):



Gino, Moore, & Bazerman (2009) Replication and extensions

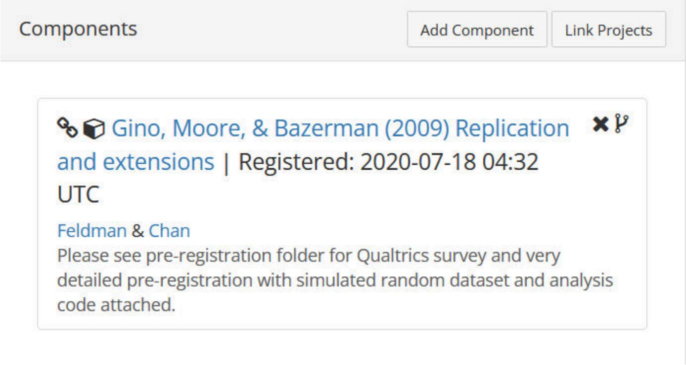
Gino_Moore__Bazerman_2009_Replication_and_extensions.docx

Gino, Moore, & Bazerman 2009:
Replication and extensions

> **.2. I also checked out the replication files. Note that these files were not included in the OSF preregistration and therefore weren't 'frozen'. However, the files were uploaded on the same day the preregistration was registered. The format of the preregistration made it somewhat difficult to identify decision criteria, but I spotted no undisclosed deviations in the final manuscript.**

We believe this represents a misunderstanding about how OSF works. The preregistration is indeed frozen, and has its own DOI: https://doi.org/10.17605/OSF.IO/64T5K.

The overarching project can be found on the sidebar indicating "Registered":.

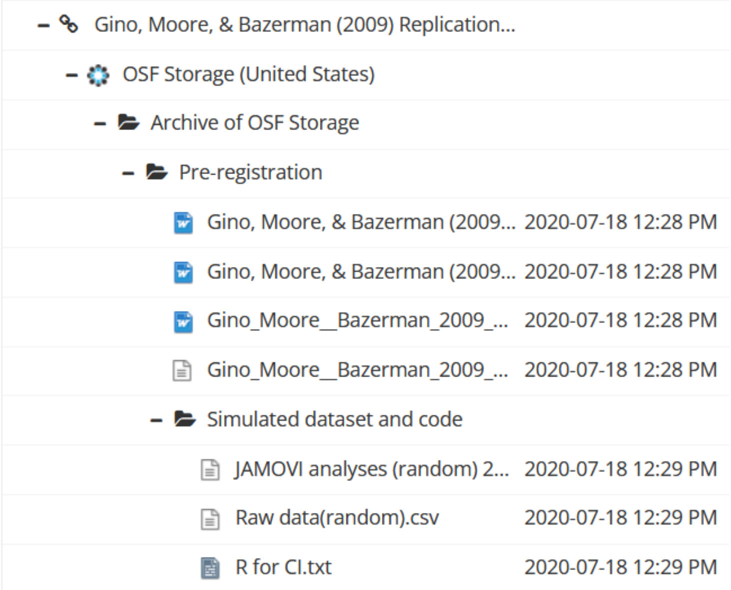


Gino, Moore, & Bazerman (2009) Replication and extensions | Registered: 2020-07-18 04:32 UTC

Feldman & Chan

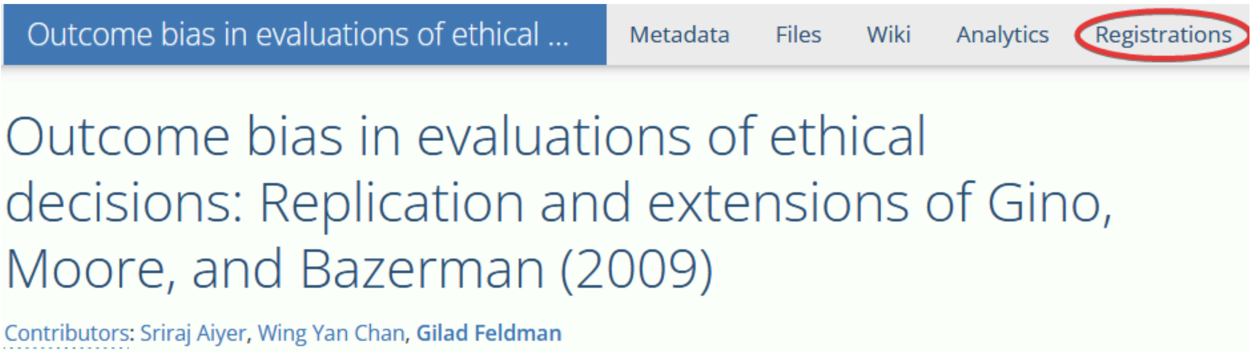
Please see pre-registration folder for Qualtrics survey and very detailed pre-registration with simulated random dataset and analysis code attached.

It also appears in the file navigator under "Archive of OSF Storage" linked to that component, all the files are frozen.



- Gino, Moore, & Bazerman (2009) Replication...
  - OSF Storage (United States)
    - Archive of OSF Storage
      - Pre-registration
        - Gino, Moore, & Bazerman (2009... 2020-07-18 12:28 PM
        - Gino, Moore, & Bazerman (2009... 2020-07-18 12:28 PM
        - Gino_Moore__Bazerman_2009_... 2020-07-18 12:28 PM
        - Gino_Moore__Bazerman_2009_... 2020-07-18 12:28 PM
        - Simulated dataset and code
          - JAMOVI analyses (random) 2... 2020-07-18 12:29 PM
          - Raw data(random).csv 2020-07-18 12:29 PM
          - R for CI.txt 2020-07-18 12:29 PM

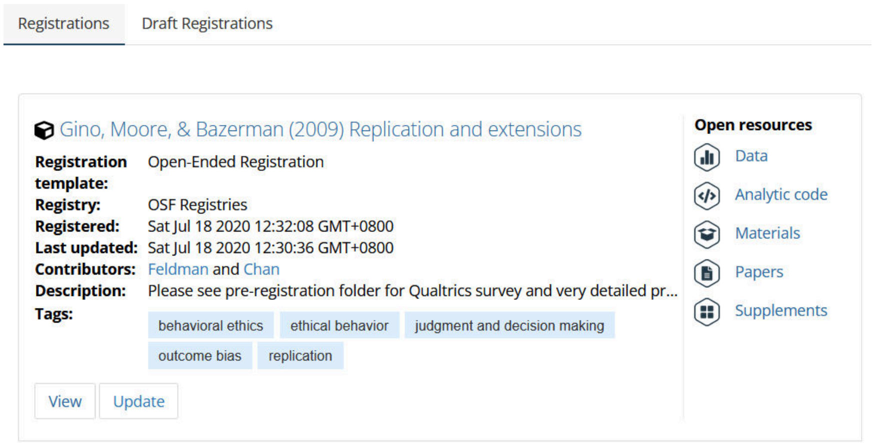
Pre-registrations on OSF can also be found using the “Registrations” tab on the top of the project



Outcome bias in evaluations of ethical decisions: Replication and extensions of Gino, Moore, and Bazerman (2009)

Contributors: Sriraj Aiyer, Wing Yan Chan, Gilad Feldman

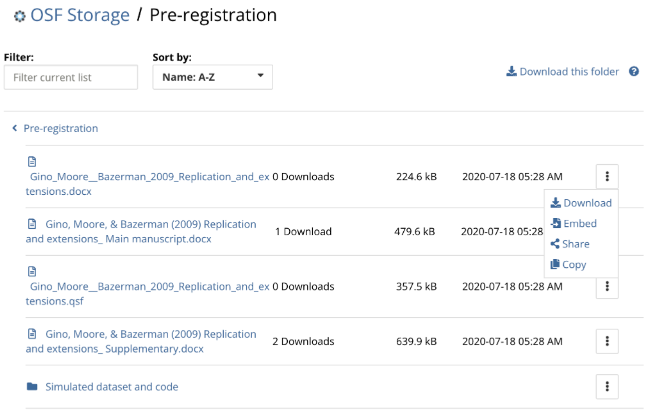
Which will take you to the following page:



Gino, Moore, & Bazerman (2009) Replication and extensions

Registration template: Open-Ended Registration
Registry: OSF Registries
Registered: Sat Jul 18 2020 12:32:08 GMT+0800
Last updated: Sat Jul 18 2020 12:30:36 GMT+0800
Contributors: Feldman and Chan
Description: Please see pre-registration folder for Qualtrics survey and very detailed pr...
Tags: behavioral ethics, ethical behavior, judgment and decision making, outcome bias, replication

To navigate the frozen pre-registration you will need to click on the files tab which will show you the following, where you could download all the files :



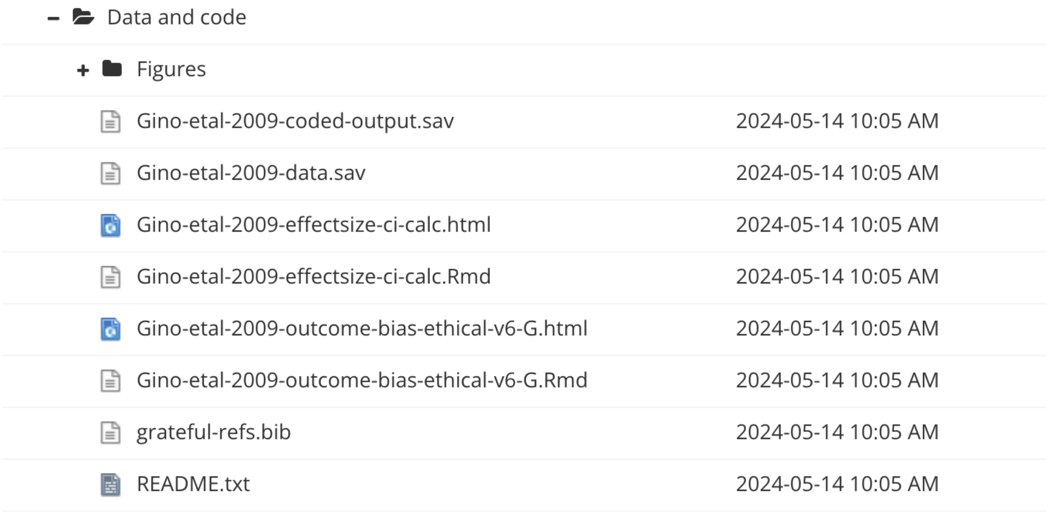
OSF Storage / Pre-registration

Pre-registration

Gino_Moore__Bazerman_2009_Replication_and_extensions.docx	0 Downloads	224.6 kB	2020-07-18 05:28 AM
Gino, Moore, & Bazerman (2009) Replication and extensions_ Main manuscript.docx	1 Download	479.6 kB	2020-07-18 05:28
Gino_Moore__Bazerman_2009_Replication_and_extensions.qsf	0 Downloads	357.5 kB	2020-07-18 05:28 AM
Gino, Moore, & Bazerman (2009) Replication and extensions_ Supplementary.docx	2 Downloads	639.9 kB	2020-07-18 05:28 AM
Simulated dataset and code			

.3. Since I do not have a Jamovi installation, I did not attempt to replicate the statistical analyses.

We conducted and shared all analyses on R/Rmarkdown with knitted output HTML files that allow you to see each of the analyses and their associated R code. These were available on the OSF under the folder "Data and code" (both named as 'outcome-bias-ethical'):

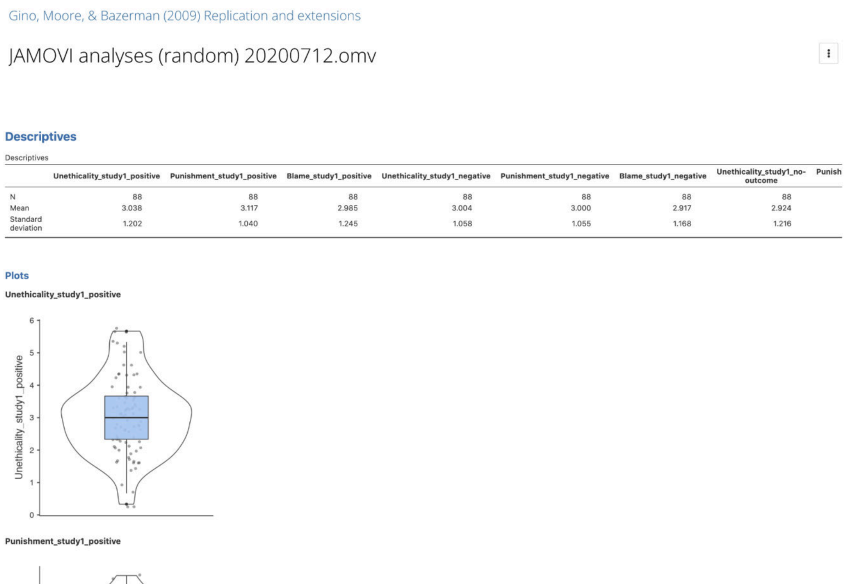


– Data and code	
+ Figures	
Gino-etal-2009-coded-output.sav	2024-05-14 10:05 AM
Gino-etal-2009-data.sav	2024-05-14 10:05 AM
Gino-etal-2009-effectsize-ci-calc.html	2024-05-14 10:05 AM
Gino-etal-2009-effectsize-ci-calc.Rmd	2024-05-14 10:05 AM
Gino-etal-2009-outcome-bias-ethical-v6-G.html	2024-05-14 10:05 AM
Gino-etal-2009-outcome-bias-ethical-v6-G.Rmd	2024-05-14 10:05 AM
grateful-refs.bib	2024-05-14 10:05 AM
README.txt	2024-05-14 10:05 AM

(Note that the updated files after this revision are *-v8-G.Rmd/html

JAMOVI runs on R, so what we basically did was to use JAMOVI for the pre-registration, and then later exported the R code to the final use in Rmarkdown. In addition, JAMOVI .OMV files are just ZIP files that include browseable HTML files. You do not need to know JAMOVI or have JAMOVI to browse them, you simply unzip the .OMV file and click on the .HTML file.

Luckily, even if you don't feel like unzipping the .OMV file, OSF actually does that for you, so this is what it looks like on the OSF:



Gino, Moore, & Bazerman (2009) Replication and extensions

JAMOVI analyses (random) 20200712.omv

Descriptives

Descriptives

	Unethicality_study1_positive	Punishment_study1_positive	Blame_study1_positive	Unethicality_study1_negative	Punishment_study1_negative	Blame_study1_negative	Unethicality_study1_no-outcome	Punish
N	88	88	88	88	88	88	88	
Mean	3.038	3.117	2.985	3.004	3.000	2.917	2.924	
Standard deviation	1.202	1.040	1.245	1.058	1.055	1.168	1.216	

Plots

Unethicality_study1_positive

Punishment_study1_positive

.4. Minor issues

p. 1: “positive outcomes were rated” -> are rated

p. 1: “Gino, Moore, and Bazerman (1988)” -> 2009

p. 1: check whether/where zeroes go before the decimal

p. 1: eta^2G should be eta^2_G

p. 8: “Secondly, we extended…” this is repeated in the next paragraph

p. 8: “and examine how positive and negative…” I found this and the following sentence confusingly written

p. 9: “it is unclear from the scenario texts where” -> whether

p. 10: “we provided a summary” -> provide

p. 10: “power analysis of the effect on outcome” -> of outcome

p. 10: “scenarios from Study,” -> Study 1

p. 11: “scenario in the replication group” -> pharmaceutical researcher scenario?

Thank you for catching those, we addressed all of these issues.

p. 11: The excerpt is described as "unchanged in all three conditions", but from Table 3, it differs between the positive and negative treatments.

We changed the description of the scenarios to reflect that the text does differ between outcome conditions.

p. 23: I am not sure I fully grasped the statistical model for the mediation analyses. Is this a mixed-effects model using data from all scenarios? If so, the random effects structure should be mentioned.

We added that this is a linear model.

p. 23: Relatedly, it would be nice to also report the sensitivity analysis which is provided by the mediation package.

We added a summary of sensitivity analyses for Hypotheses 2 and 5 (with the full information now available in the code/html output). See below from the Hypothesis 5 section on page 36:

> "We fit a linear model with all data from all scenarios with a positive or negative outcome. We found no support for the effect of outcome (positive vs negative) being mediated by judgements of unethicality. The bootstrapped indirect effect of unethicality on punishment was .05, 95% CI [-.04, .12], Sobel Z = 0.84, p = .40, ACME found to be robust until = 0.7. The bootstrapped indirect effect of unethicality on blame was .03, 95% CI [-.02, .08], Sobel Z = 0.82, p = .41, ACME found to be robust until = 0.5."

p. 23: Is there a reason not to also report the difference between the positive condition and the no-outcome condition? To me, it seems most sensible to decompose the overall effect into effects of positive and negative outcomes (while acknowledging differences in the phrasing of the vignettes).

For the mediation analysis, we aimed to replicate and therefore simply follow the original paper's analyses for which there was no neutral (no-outcome) condition.

p. 24: "ratings on unethicality […] was higher" -> were

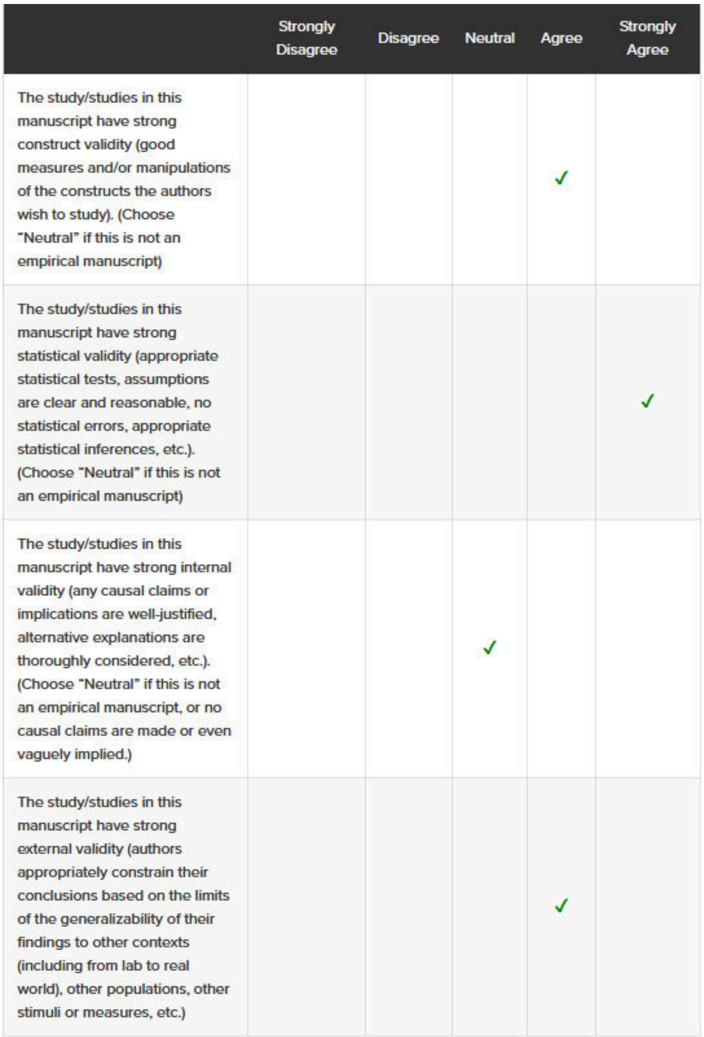
p. 28: "Hence, whilst participants judged" This sentence is incomplete

p. 34: "feelings of blame was" -> were

p. 36: "we successfully the original" -> incomplete sentence

Thank you for spotting these issues. We revised accordingly.

Reply to Reviewer #2



	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose "Neutral" if this is not an empirical manuscript)				✓	
The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose "Neutral" if this is not an empirical manuscript)					✓
The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose "Neutral" if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.)			✓		
The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.)				✓	

Thank you for the constructive feedback.

Major comments:

.1. I had difficulties finding the actual preregistration of the replication, even though the authors have a "pre-registration" folder in their OSF page. I would suggest to create and refer to the doi of the document in the text so it is easy for readers to access it.

Thank you. We updated the link to use the DOI instead.

> **.2. When I read the abstract, introduction, and the results section, I wasn't entirely sure what was meant by hypotheses 2a & 2b and 5a & 5b. The authors might consider clarifying these predictions a bit further, perhaps by providing an example or offering more detailed explanations.**

The hypotheses were summarized in Table 1. We added a reference to Table 1 just before the results in the relevant sessions to make this clearer.

> **.3. On a related note, I am also missing a detailed justification for adding the mediation analysis and Scenarios 4–6 as an extension to the current replication.**

We added justification of this to the section "Extensions: Neutral condition and scenarios from Study 2":

> "Finally, the original paper analyzed whether unethicality mediated the relationship between outcome information and ratings of punishment and blame (see Hypotheses 2a and 2b in Table 1). As an extension, we also investigated the mediating role of unethicality in the three scenarios added from Study 2, such that we can determine if this finding replicates in other scenarios as well (see Hypotheses 5a and 5b in Table 1)."

> **.4. A common criticism of replications based on online samples is that the quality of online testing platforms (such as MTurk) has declined in recent years, and original authors raised concerns about the reliability of (replication) data collected through these platforms.**
>
> **I am curious whether any additional manipulation checks or quality control measures were performed besides the comprehension checks?**

Thank you for raising that. We realized we could have done better in reassuring readers about how we addressed this concern.

We did not use MTurk directly, but rather MTurk via CloudResearch that ensures data quality and is comparable to platforms like Prolific. This allowed us to put basic data quality measures in place through CloudResearch. In addition, we ran simple attention checks in the consent form before embarking on the study. We added an explanation to the methods section:

> "Participants first indicated their consent, with two questions confirming their eligibility, understanding, and agreement with study terms, which they had to answer with a "yes" and required responses in order to proceed to the study. The two questions also served as attention checks, with a randomized display order of the options (yes, no, not sure) - 1) "Are you able to pay close attention to the details provided and carefully answer

> questions that follow?", 2) "Do you understand the study outline and are willing to participate in a survey with attention/comprehension checks?". Failing any of the two attention questions meant that the participants did not indicate consent and therefore could not embark on the study and were asked to return the task. Upon completion of these steps, participants proceeded to begin the survey.
>
> Based on our extensive experience of running similar judgment and decision making replications on MTurk, to ensure high quality data collection, we employed the following CloudResearch options: Duplicate IP Block. Duplicate Geocode Block, Suspicious Geocode Block, Verify Worker Country Location, CloudResearch Approved Participants."

This mirrors our many other replications using a similar setup, which have shown consistent support for classic findings, including outcome bias. In this study, our findings are not random noise, as we found support for some things and not for others, which also help address concerns regarding data quality.

> **.5. As I understand it, the scenarios presented were classified as 'grey-area.' How was this classification determined, and how did the original authors arrive at it? It is crucial for the replication to demonstrate that participants perceive the scenarios as grey-area to the same extent as those in the original study. If no additional manipulation checks were conducted, I would suggest including a paragraph that justifies the adequacy of the current sample's quality.**

We referred to the scenarios as 'ethically grey' more as a turn of phrase, rather than a formal classification. Given that the scenarios display a range of responses in terms of ethicality (see Figure 4), the scenarios are not interpreted as unequivocally ethical or unethical. Using scenarios that are clearly interpreted as one or another would be very limiting to use when asking participants to rate their perceived ethicality.

> **.6. Related to that I was wondering what the outcome of the comprehension checks was? This information is crucial for readers to judge the data quality, so I would recommend adding a table showing the outcomes per scenario. If certain individuals show high rates of non-comprehension, I suggest to add an additional exploratory analysis excluding these participants. These can be added to the supplementary material. It makes a strong case for replication critics if the results are robust when retaining only the "highest" quality data points.**

We noted on page 21:

> "Participants were required to answer all questions correctly before they could proceed to the next page."

This means that all participants were required to pass the comprehension checks before they could see and answer the dependent measures. This is one of the steps that we took to ensure attentiveness and comprehension and reduce noise and need for analyses regarding exclusions which could get confusing to interpret.

> **.7. Something else that I found interesting and worth to discuss is the fact that ratings of unethicality and punishment (but not blame) were higher in the positive outcome condition than in the no outcome condition for scenarios 1-3. That may be worthwhile to add in the discussion (as interesting exploratory pattern). Do you have any explanations for this finding?**

Thank you. We added the following to page 37:

> "We also found that ratings of unethicality and punishment (but not blame) were higher in the positive outcome condition than in the no outcome condition for scenarios 1-3. This could be interpreted as blame being specifically associated with negative outcomes (as it seems intuitively unusual to blame an individual for a positive outcome), but actions can still be judged unfavourably in terms of being unethical or deserving of punishment even if they result in positive outcome."

> **.8. Minor comments:**
>
> **.8.1. I really like how the tables clearly state the hypotheses and highlight which aspects of the original studies were the same, similar, or different. However, Table 3, which describes the scenario, might be better suited as an appendix to the manuscript. While it is a valuable addition to both the manuscript and the methods section, it somewhat disrupts the overall flow.**

We included Table 3 for transparency in terms of the survey materials we use. Whilst this may come at a cost of flow of the manuscript, we especially think the scenarios are important to include in full given that we discuss the design of the original vignettes earlier on the "Experimental Manipulations in Target Article' section. As a result, we would like to keep the table in the main manuscript. We are open to revising this if given clear editorial guidelines.

> **.8.2. In the introduction, the fraud allegations against Gino are not mentioned, which makes it difficult for the reader to understand what is meant by 'raised concerns regarding work conducted by the lead author.' If these allegations are a reason for conducting these replications, it should be clearly stated. Since nothing has been proven yet, you should, of course, address the accusations carefully. Additionally, why was the follow-up article flagged as a concern for replicability? Is there a concern about data manipulation, or have there been unsuccessful replications of this article?**

The project was started before the allegations came to light, meaning that they do not relate to why we conducted the replication in the first place. We do not have specific concerns about data manipulation for this study and this is the first replication attempt of this study that we know of. Given the allegations, all of the target article's lead author's published articles were flagged as part of the Many Co-Authors Project, including the follow-up article we cited. We only discussed the allegations and this project involving her previous co-authors to mention the added importance of this replication that was not apparent to us when we had started the project (as it seems it would be odd not to mention the allegations at all). We would rather not go into more details in an academic article.

> **.8.3. In my opinion, simply stating the number of citations of the preprint as evidence of its impact in the field is insufficient. Why was this paper cited, and how have its findings been applied? What specific impact has it had, and what follow-up research has been conducted based on it?**

We previously briefly mentioned some of the follow-up literature and its impact in the introduction:

> The article has since led to impactful follow-up work, including the authors' own research on ethical judgments in Gino et al. (2010), as well as the work of Kneer and Machery's (2019) and Lench et al. (2015) exploring outcome biases in morality and attributions.

We aimed this as a replication focusing on the specific article rather than a review of the literature, and would rather not go too deep into reviewing this literature. Citations are, for better or worse, one of the most widely used indicators for impact, or engagement. To motivate this a bit further, we added the following:

> In one of the first meta-science examinations of the prevalence of replications in publications Makel et al (2012) found very low prevalence of replications (estimated as 0.2%, 1 in 500 articles are publications) and noted that "if a publication is cited 100

times, we think it would be strange if no attempt at replication had been conducted and published".

.8.4. I didn't understand the following sentences: 'Our pre-registration stipulated that we intended to analyze all data without any exclusions and then apply exclusions as a supplementary analysis. Seven participants met the pre-registered supplementary exclusion criteria.' What exactly were the pre-registered supplementary exclusion criteria?

We appreciate the feedback. The exclusion criteria was on page 22 of the preregistration's supplementary file, indicating this criteria:

> Participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale)
> Participants who self-report not being serious about filling in the survey (self-report < 4, on a 1-5 scale).
> Participants who correctly guessed the hypothesis of this study in the funnelling section.
> Participants who have already seen or done the survey before.
> Participants who failed to complete the survey. (duration = 0, leave question blank)
> (When target sample is online:) Participants not from the United States.

We adjusted our code to run our Rmarkdown both with and without exclusions using the file "knit-no-exclude-vs-exclude.R" which now also produces "Gino-etal-2009-outcome-bias-ethical-v8-G-exclusions.html".

We added the following to the supplemental materials:

> The results from our main manuscript do not change with these exclusions. Filename "knit-no-exclude-vs-exclude.R" allows running both with and without exclusions. The knitted output analysis file is available on the OSF: Gino-etal-2009-outcome-bias-ethical-vX-X-Exclusions.html

.8.5. You mention the limitations related to the experimental manipulations in the target article. I'm curious whether follow-up research by the original authors faced the same limitations or if they maintained the ancillary text consistently in their subsequent work. If they did keep the text constant, the results from this follow-up research could be compared to those from the original study and the replication, with the caveat that these articles have been flagged.

This is tricky, given the issues raised regarding the lead author's work, and so we suggest caution in looking at the author's follow-up work as a reference for comparison. The mentioned follow-up article (Gino, Shu, & Bazerman, 2010) showed positive results and seems to implement the manipulation of outcome in a cleaner manner, yet we again face the issue of that article being flagged for the Many Co-Authors Project. Our main aim here was to focus on the target study and to provide guidance for future researchers (keeping all other aspects constant outside of the manipulation is important for any research). Future replication work could revisit the target article's authors' follow-up articles.

> **.8.6. An interesting point worth discussing is that ratings of unethicality and punishment (but not blame) were higher in the positive outcome condition compared to the no outcome condition for scenarios 1-3. It might be valuable to include this as an interesting exploratory pattern in the discussion. Do you have any explanations for this finding?**

Thank you. See our response to point #7.