CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	The worst-motive fallacy: A negativity bias in motive attribution
Authors	Walmsley, Joel;O'Madagain Cathal
Publication date	2020-10-21
Original Citation	Walmsley, J. and O'Madagain, C. (2020) 'The worst-motive fallacy: A negativity bias in motive attribution', Psychological Science, 31(11), pp.1430-1438. https://doi.org/10.1177/0956797620954492
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1177/0956797620954492
Rights	© 2020, the Authors. Reprinted by permission of SAGE Publications. This is a post-peer-review, pre-copyedit version of the following article: Walmsley, J. and O'Madagain, C. (2020) 'The worst-motive fallacy: A negativity bias in motive attribution', Psychological Science, doi: 10.1177/0956797620954492. The final authenticated version is available online at: https://doi.org/10.1177/0956797620954492 - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-05 17:34:46
Item downloaded from	https://hdl.handle.net/10468/10750

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

The 'Worst Motive' Fallacy: A negativity bias in motive attribution.

Joel Walmsley[1] & Cathal O'Madagain[2]

(1) Department of Philosophy, University College Cork, Ireland.

Address for Correspondence:

Department of Philosophy, University College Cork,

2-4 Elderwood, College Road,

Cork,

Ireland

T12 AW89

j.walmsley@ucc.ie

(2) Départment d'Études Cognitives, École Normale Supérieure, Paris.

and

School of Collective Intelligence, Université Mohammed VI Polytechnique, Morocco.

cathal.omadagain@um6p.ma

Abstract

In this paper, we describe a hitherto undocumented fallacy—in the sense of 'a mistake in reasoning'—constituted by a negativity bias in the way that people attribute motives to others. We call this the 'Worst Motive Fallacy.' We conducted two experiments to investigate it. In Experiment 1 (n=323), participants expected protagonists in a variety of fictional vignettes to pursue courses of action that satisfy the protagonists' worst motive, and furthermore, participants significantly expected the protagonist to pursue a worse course of action than they would prefer themselves. Experiment 2 (n=967) was a preregistered attempted replication of Experiment 1, including a bigger range of vignettes; the first effect did not replicate for the new vignettes tested but did for the original set. And again we found that participants expected protagonists to be more likely than they were themselves to pursue courses of action they considered morally bad. We discuss the Worst Motive Fallacy's relation to other well-known biases, as well as its possible evolutionary origins and its (meta-)ethical consequences.

Keywords: Cognitive Bias; Motives; Attribution; Meta-ethics; Experimental Philosophy; Moral Intuitions; Moral Judgment

1 Introduction

When we judge the moral status of an action, we routinely take into consideration the motives of the person who performs it. In moral philosophy, otherwise divergent views nonetheless agree that the motives behind an action are crucially relevant to its moral status: the 'Agent-Based' virtue ethics of Slote (1995) takes motives to be the *exclusive* determinants of the goodness of an action, whilst according to Kant's deontological account, an agent cannot perform a moral action without a good 'will.' Even consequentialist views (according to which an action's moral status is to be judged on the basis of its outcomes) leave some room for the agent's motives; Sidgwick (1884, p.200) for example, suggests that "a man who prosecutes from malice a person whom be believes to be guilty does not really act rightly; for, though it may be his duty to prosecute, he ought not to do so from malice."

In order to put these philosophical views into practice, however, we would need to be confident that people are actually *good* at discerning an agent's motives. If we are not very good at impartially identifying the motives of others, or worse, if we are systematically biased in our attribution of motives, then we would need to

be much more cautious of adopting a moral theory that recommends or requires us to do so.

In this paper, we outline the results of two experimental studies suggesting that people *are* subject to such a bias, which we call the 'Worst Motive Fallacy.' We found evidence that people display a tendency to assume that an agent's worst motive for an action is their main motive. We argue that this constitutes a hitherto undocumented cognitive bias

2 The Worst Motive Fallacy: Background

The folk aphorism known as *Hanlon's Razor* states that one should "Never attribute to malice that which is adequately explained by stupidity." The sentiment is often expressed as a necessary warning against concluding without further evidence that an agent acted with bad motives, when their actions could well be explained simply by incompetence. How likely is it that such a bias in fact exists?

One might have predicted that such a bias existed on the basis of several other well-documented psychological effects. First, there is a general bias towards negativity across a wide range of psychological phenomena (see, e.g., Baumeister, Bratlavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001). Emotions, attention, motivation, information processing, and memory are all more strongly influenced by negatively-valenced stimuli than neutral or positive ones. The impact of this negativity bias has been documented for decision-making (Kanouse & Hanson, 1971), and even differences in political ideology (Hibbing, Smith, & Alford, 2014). It is therefore not surprising that we might also pay more attention to negative motives in our assessment of others' actions.

Second, within social psychology, there are known biases when it comes to the attribution of agency *simpliciter* (see, for example, Hewstone, 1983; Shaver, 1985). We are more likely to view events as having been caused by an agent when those events are negative rather than positive (Morewedge, 2009) and even six-month old infants are more likely to attribute agency to a mechanical claw when its actions are 'bad' rather than 'good' (Hamlin & Baron, 2014). Similarly, in the 'side-effect

effect' (Knobe, 2003a, 2003b) we find that an action is more likely to be regarded as *intentional* when it has a harmful side-effect than when it has a beneficial side effect. Pettit and Knobe (2009) further demonstrate that this effect generalises to *desire*; we are more likely to think that an agent *desired* a side-effect, when that outcome is bad.

Third, another closely-related bias is found in the so-called 'actor/observer difference' or the 'Fundamental Attribution Error' of motive attribution (Ross, 1977). An actor's view of their own behaviour emphasizes the role of external situational and environmental factors ('I failed the test because a barking dog kept me awake last night'), whereas their view of another's behaviour emphasizes internal factors such as character and motives ('She failed the test because she's lazy and stupid') (see Jones & Nisbett, 1971; Monson & Snyder, 1977; Nisbett, Caputo, Legant, & Maracek, 1973). We are more inclined to attribute our own failures to negative aspects of our environment while we attribute others' failures to negative aspects of their character.

Bearing these biases in mind, it might not be so surprising were we to systematically commit what we have called the 'Worst Motive Fallacy.' Sometimes an agent can have several *different* reasons for the very same action. In such cases, the reasons can often be ranked, morally speaking: some may be more praiseworthy or noble than others, and the agent may perform the same action for better or worse reasons. Furthermore, the agent—or an observer—may regard one of the reasons for action as the *main* reason for action (where the 'main' motive is not just the strongest *desire*, but rather the motive that would be pursued above all others if the available courses of action could only satisfy *one*). But if an observer were to systematically suppose, all else being equal, that the worst of these motives was the main motive, or that others are primarily motivated by worse motives than they have themselves, this would be a mistake. The study we present here explores whether people have a cognitive bias towards committing the Worst Motive Fallacy, with two experiments.

3 The Worst Motive Fallacy: Two Experiments

3.1 Experiment 1

In our first study, we required participants to read a short story about a protagonist who has two motives for carrying out some action. In each case, one of the motives is good, and one is bad. The protagonist then discovers that they cannot after all pursue the action they had planned: they must choose one of two alternatives, where one satisfies the original 'good' motive, and the other satisfies the original 'bad' motive. The participant was then asked to say which option they thought the character in the story would pick, and asked to rate the goodness or badness of each motive. Finally, participants were asked which action they would prefer to do themselves. We expected that if the Worst Motive Fallacy is a real bias, then the worse a participant rated a motive, the more likely they would be to expect the agent to choose an option that satisfied that motive rather than a competing one. We also expected that participants would expect the characters in the story to choose a worse course of action than they would prefer themselves.

3.1.1 Materials and Methods

Each participant was given one of four different vignettes. Here is an example:[1]

> A politician has some funding left over from her campaign, and she decides to use it to hire a computer engineer that she knows. She does this for two reasons. First, the engineer has recently lost his job and is in need of new work, and the politician wants to help him out. Second, the politician wants the engineer to send misleading messages to her opponent's supporters to send them to vote on the wrong day.
>
> When she describes the work to the engineer, however, the engineer says he will not do it.
>
> The politician has two further options. She could hire the unemployed engineer anyhow, to do ordinary computer maintenance work. This will help the engineer who needs income, but won't help the politician to mislead voters. Or, she could hire a computer hacker who has no

[1] All four vignettes used in Experiment 1 are included as Appendix 1. Experiment 2 used *twelve* vignettes in total, consisting of the four used in Experiment 1, together with eight additional ones (also included in Appendix 1).

problem sending misleading messages. This will help the politician to mislead voters, but will not help out the unemployed engineer.

Which option do you think the politician will pick?

(a) Hire the engineer, to give him work

(b) Hire the hacker, to mislead the voters

After deciding which option the protagonist will pick, participants were then asked to rate the 'goodness' or 'badness' of the motives that were described in the story, on a scale from minus 10 ("very bad") to plus 10 ("very good"). Our hypothesis was that participants would expect the protagonist to pick the option that satisfied the motive they rated as worse.

We interpret participants' answers to the question about which option the agent will pick to indicate their assessment of the protagonist's *main* motive in the original action: since the protagonist can only satisfy one of the stated original reasons, whatever action they choose would presumably be the one that satisfies the motive that is most important to them. Of course, one might worry that merely *mentioning* the bad motive might prejudice participants against the protagonist; the mere fact that the protagonist could even entertain such bad motives might be taken as evidence that they are a bad person generally. But this is precisely the point; participants know that the protagonist has both good and bad motives since they are both stated explicitly. If our hypothesis about the Worst Motive Fallacy is correct, participants will indeed be biased towards thinking that the bad motives are the main ones.

Our measure of goodness or badness of a motive also allowed us to control for what we will call the 'extremeness' of a motive. Suppose you were told that someone had two motives in mind when they were setting out on some course of action: they wanted to end world hunger, and also they wanted to mildly annoy a co-worker. If you were told this person found they could only do one of these things, it is obvious that you would expect them to pursue the action that would satisfy the motive of ending world hunger. But this is because the motive is so much more 'extreme' than its competitor. Mildly annoying a co-worker is a bit bad, but ending world hunger is *extremely* good. All else being equal, we should expect that if someone has

such an extreme motive in mind they will likely prefer an action that would satisfy it, over an action that would satisfy a competing but not very extreme motive. If the good motives in our vignettes were, without our realizing it, all considered as more extreme than the bad ones, then this would create a reason for participants to expect the agents in the story to follow the good motives. And if the bad motives were all rated as more extreme, that would create a reason to expect the agent to follow the bad motives. Any bias to expect the agents in the story to follow the bad actions over the good could therefore be revealed only if the extremeness of the motives were controlled. To do this, we treated the *distance from zero* in either direction as a measure of the 'extremeness' of the motive. If the motive were considered extremely bad, we expected it would be rated as -10, but if it were just mildly good, it might be rated as +2. The 'extremeness' of the first motive would therefore get a score of +10 (ten points from zero) and the extremeness of the second as +2. Overall, we wanted to make sure that any difference in extremeness was not explaining any bias we found to favour the bad over good motives, by adding this measure as a control.

Notice that participants are not simply asked 'Which of the motives described do you think is the protagonist's main motive?' This is to avoid alerting the participants to the purpose of the study. Once participants know what a study is exploring, there is a risk that they will adjust their answers to give what they think is a favourable representation of themselves, thereby disguising their real attitudes (Krumpal, 2013). Asking participants what the protagonist in the story will likely do next allows us to explore participants' assumptions about the protagonist's motives without asking them about this directly.

Finally, we asked participants which of the options they would choose themselves if they were in the position of the protagonist. We predicted that participants' responses to this question would either fail to match the course of action they expect of the agent in the story, or else would negatively correspond to that action. This allowed us to rule out the possibility that participants were simply ascribing to the protagonists the course of action that the participants would prefer themselves.

In each vignette, the motives ascribed to the character were counterbalanced so

that in half the vignettes the 'bad' motive was described first, and in the other half the 'good' motive was described first. The number of words used to describe the good and bad motives and actions was the same, so that overall we gave participants no reason to suppose that one of the motives described was the primary motive. Similarly, the contexts described are very different—we used four different vignettes: a politician (described above); a man who must decide whether to take the bus to town to buy his friend a present, or take the train to rob a pensioner; a child going to a party who must decide whether to wear a dress that will embarrass the host, or a pair of jeans that will make her mother happy; and a college student who has to decide whether to go to France for the summer where he expects he can cheat on his girlfriend, or go to stay with cousins in Argentina where he will learn Spanish to improve his studies (vignettes can be read in Appendix 1).[2] In spite of the diversity of the contexts described, and the counterbalancing of the presentation of the motives, we predicted that participants would be more likely to expect the character to pursue the action that satisfies the worst of the two motives.

The study was run using the Qualtrics online survey platform, while participants were collected using Amazon's Mechanical Turk. Participants were paid 25c for their participation, which took about 1-2 minutes. We piloted 52 participants and found a significant difference between the rating of motives behind actions participants expected to be chosen versus not chosen (n=52, t=-2.132; p=0.037). A power analysis indicated that 75 participants would produce a significant difference with alpha = 0.01. Since we designed four vignettes and counterbalanced the order in which the motives and options were mentioned in the vignettes in eight ways, we aimed to collect data from 80 participants for each vignette, which would allow the eight counterbalanced versions to be distributed evenly over the group. As a result we aimed to collect data from 320 participants. We ran three 'attention' questions at the beginning of each experiment, and excluded participants who did not correctly answer all of those questions. We also excluded participants who did

[2]Experiment 2 (vignettes for which are also included in Appendix 1), examines an even wider range of hypothetical scenarios: a man deciding between working for an NGO or a travel agency; a bar owner deciding how to treat a regular customer; a shop owner trying to decide where to open a second branch; an unprepared professor trying to decide whether to cancel his class; a man deliberating over which car to buy; an inventor deliberating between two labour-saving devices; a country's president deciding how best to spend his funds; and an aspiring teacher deciding which city to move to.

not complete the whole test. Finally, if the same participant took the survey more than once, we excluded the second response. In total, we collected 408 responses and excluded 85, resulting in 323 responses.

3.1.2 Analysis

We analysed the results using a Generalized Linear Model (GLM) in R version 3.4.3, using package lme4. This allows us to model the main predictor along with control predictors all at once, and avoid the problems associated with running multiple analyses. In our full model we included choice of option (good or bad) as the dependent variable. The main predictor was the difference between the participant's ratings of the motives behind the two options. We expected that the worse a participant rated the bad motive relative to the good one, the more likely they would be to expect the protagonist to pursue the action that satisfied that motive. To control for 'extremeness' we included the difference between the absolute ratings of the motives (the distances from zero in either direction). We included this as a further predictor in the model as a random effect, along with the interaction between extremeness and rating.

We identified significance levels using likelihood ratio tests to compare the fit of different models. The full model included terms for 'rating', 'extremeness' (the absolute value of rating), and their interaction, along with a measure of the participant's *own* preference (which choice they said they would prefer themselves), the vignette read by the participant, and duration (how long the participants took to complete the test). The null model included only duration and vignette.

3.1.3 Results

The full-null comparison was significant (n=323, χ^2 = -19.16, df=4, p<0.001). To isolate the source of the effect, we compared a model including the interaction between the absolute rating and the basic ratings, but found no effect of the interaction (χ^2=-0.541, df=1, p = 0.4616). We therefore removed the interaction from the model. We now compared models that differed only by the inclusion of the basic rating, but where both featured the absolute rating and the participant's

own preference. We found that including the basic rating in the model significantly improved the fit (χ^2 = -5.8005, df=1, p=0.016). This outcome confirmed our hypothesis: participants expected the agents in the story to follow the course of action that satisfied the motive they considered relatively worse, and that in spite of any effect of the extremeness of the action.[3] There was indeed such an effect of extremeness: the more extreme the motive, the more likely the participants were to expect the agent to follow it (χ^2=8.7769, df=1, p=0.003). But since this is included in the same model as the 'basic rating', it cannot account for the bias toward the bad motives by itself.[4] These results are plotted in the left hand side of Figure 1.

Finally we found a significant effect of the participant's *own* preference: participants predicted that they themselves would have chosen the opposite of what they expected the protagonist to choose (χ^2 = -6.607, df=1, p= 0.01), as can be seen in the right hand side of Figure 1. This is consistent with findings that we generally expect others to act in worse ways than we would prefer to act ourselves.

There was no effect of vignette and there was no effect of duration of the experiment. All data and scripts, and the results of our models can be viewed at https://osf.io/mjrpy/.

3.2 Experiment 2

The editor of this journal suggested that we conduct a pre-registered replication of this study, to test its robustness. We therefore conducted a pre-registered replication using the same set-up as Experiment 1 (see https://osf.io/vae8g/), in which we ran a total of twelve vignettes: the original four, as well as eight further vignettes

[3] Another way to look at this is by inspecting the full model (called 'full1' in our script, viewable here: https://osf.io/96eb3/); here you can see that even including the effect of extremeness in the model, the effect of the difference in ratings is still clearly significant (z= 2.483, p=0.013))

[4] For the purposes of Experiment 1, we treated vignette as a random effect in the analysis. As emphasized by the editor, treating vignette as random *intercept* would allow us to make a stronger claim about the generalizability of the claim (see Judd, Westfall, & Kenny, 2012). We therefore ran a Generalized Linear Mixed Model (GLMM) and included vignette as a random intercept rather than a random effect. The model failed to converge, and so we removed removed the term 'duration'. The model converged. We now compared models with and without the basic ratings term, and including vignette as a random intercept. The inclusion of the basic rating showed a significant improvement of fit (χ^2= 4.434, df=1, p=0.035). However, we are advised that a model including a random intercept with as few as four levels will give unreliable results, converging or not. Such an analysis is therefore perhaps unreliable for the present study, with only four levels of stimuli. Our pre-registered replication (see Experiment 2) tested twelve vignettes for this reason, and the analysis includes vignette as a random intercept.

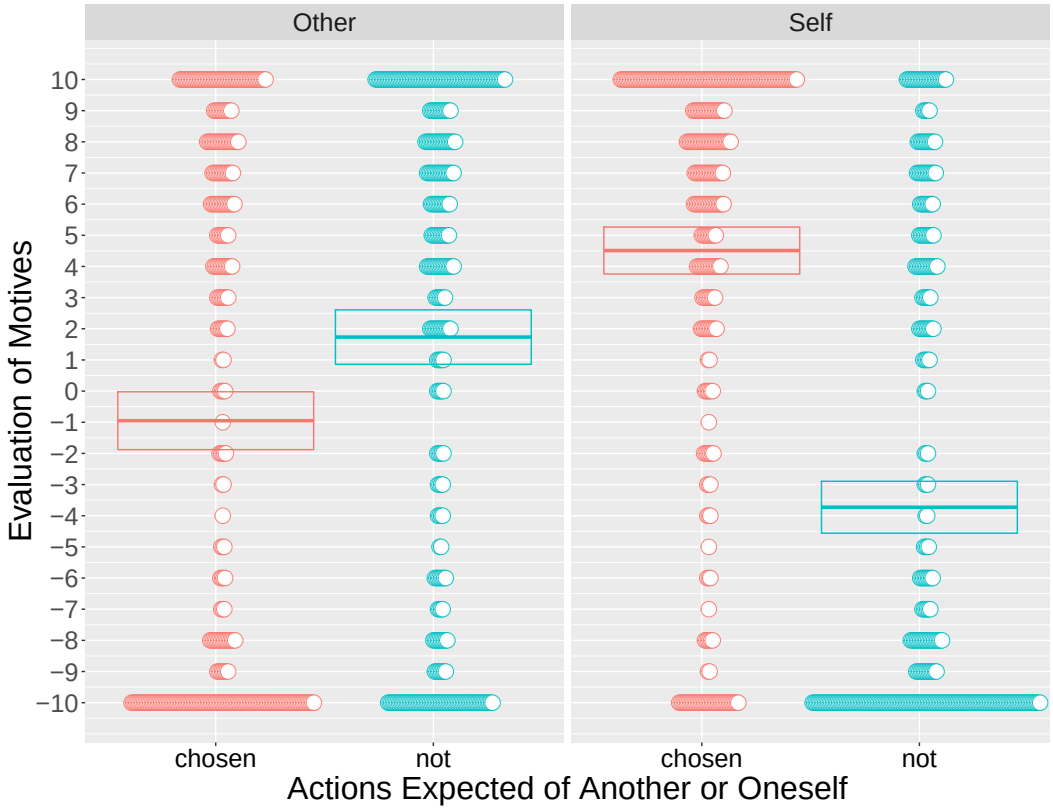


Figure 1: Experiment 1. The X axis shows the options that participants expected would be 'chosen' or 'not chosen' by the protagonist (on the left), or by themselves (on the right). The Y axis shows how participants rated the motives behind those actions. Disks represent individual participants' ratings of a given motive, boxes represent mean and 95% confidence intervals. On the left are the the options the participants expected the protagonist to choose. Overall, it is clear that participants rated the motives behind the actions they expected to be chosen as worse than the motives behind the actions they expected not to be chosen. On the right are the options that participants stated they would prefer themselves. Here it is clear that although participants expected protagonists to make choices driven by the motives they rated as worst, they also expected the protagonists to pursue actions worse than they would prefer themselves.

(reproduced in Appendix 1: one of these was based on a recommendation by a reviewer, and the other seven were written by the authors). To analyse the new data we used a Generalised Linear Mixed Model in R version 3.4.3, using package lme4 and the logit link function. This allowed us to include a random intercept of vignette in our analysis, thereby ensuring that any effects are generalizable to the full set of vignettes. In total we collected 1657 responses and excluded 690 (41%), resulting in 967 responses. These were excluded for not passing attention questions (25.7%), taking the test more than once (14.6%) or not finishing the test (1.7%). The high drop rate raises some suspicions about the data quality, and the results should be considered in that light.

3.2.1 Results

The full-null comparison was highly significant (N=967, χ^2 =141.98, df=4 $p<$ 0.0001). We found no significant effect of the rating: as can be seen in the plot on the left hand side of Figure 2 the 95% confidence intervals are overlapping. The first main effect of Experiment 1 did not, therefore, replicate as we thought it might. We did find an effect of 'extremeness': the more extreme an option, the more likely participants were to follow it (χ^2=39.712, df = 1, p<0.0001). However, this time the good motives were treated as more extreme than the bad motives. As we discuss below, this may explain why our prediction that participants would pick bad over good motives did not replicate. And again we found a significant overall effect of the participant's own choice across all vignettes. Participants expected the protagonist in the story to pursue actions that they rated as worse than the actions they would have preferred to pursue themselves (N=967, χ^2 =97.141, df=1 p<0.0001). This is depicted in Figure 2.

To explore more closely why our main prediction did not replicate, we subset the data in the replication to distinguish the original vignettes (that were also used in Experiment 1) from the new vignettes (that were not). We found that in the original vignettes, the main effect of rating was again significant (N=332; χ^2=5.9072, df=1 p= 0.015) and therefore in this sense, the precise outcome of Experiment 1 was indeed replicated. Participants reading the vignettes from the first study again

expected the characters in the stories to be more likely to act on the bad rather than the good motive. The effect size is stronger, in fact, than the effect in the first study (χ^2= 4.434, df=1, p=0.035) in spite of the sample size being almost identical (323 participants in Experiment 1, 332 participants in Experiment 2 reading the same vignettes). There was also an effect of extremeness: as before, participants expected the agent to follow the action they rated as more extreme (χ^2=18.589, df=1 p<0.0001). But again, since this is included in the model that displays the effect of 'basic rating', it cannot explain away that effect.

In the new vignettes considered on their own, on the other hand, there was no effect of rating. There was however again a significant effect of extremeness (N=635; χ^2=20.034, df=1 p< 0.0001). This could mean that the participants may have treated the 'extremeness' of the good motives in the new vignettes as stronger than the bad ones, and this may explain why the tendency to rate the bad motives as more likely to be acted upon was not found in the new vignettes. To check this we looked at the mean values of the good versus bad motives in the original versus the new vignettes from the replication. Sure enough, for the original vignettes featured in the replication, the bad motives were considered more extreme than the good motives (mean absolute bad: 8.2; mean absolute good: 7.45), but for the new vignettes, the good motives were considered more extreme than the bad motives (mean absolute bad: 5.63; mean absolute good: 7.38).

We think this explains why the effect was not found for the new vignettes; any bias the participants may have had to consider the bad motives as more likely to be acted upon was counteracted by their consideration of the good motives as being more extreme than their competitors. Recall that absolute rating is included in the model that reveals the effect of rating in Experiment 1 and for the original vignettes in Experiment 2, which means that absolute rating or extremeness could not explain the effect of rating that was found in those analyses. The effect of 'badness', whereby participants expect the agents in the story to act on the worst motive, *goes beyond* the effect of extremeness, even though they both go in the same direction. But where there is an effect of extremeness in the opposite direction, this could explain why we found *no effect* of rating in the new vignettes, since considering the good motives as

more extreme than the bad motives will have pushed participants' intuitions about which motive was more likely to be acted upon in the direction of the good motives, thus cancelling out any effect of 'badness'. We note that these exploratory analyses were not included in the preregistration. Further details on methods in Experiment 2 can be found in Appendix 2.

Overall, then, we can say that we have found clear evidence of a tendency for participants to expect others to be more likely to act on the worst motive attributed to them, all else being equal. We were not able to replicate the effect in our new vignettes, but because the good motives were rated as far more extreme than the bad motives in those cases, this failure does not cast doubt on the finding. However, future studies, involving new vignettes with motives that are more evenly balanced, will be needed to make the case that the effect will generalize indefinitely to new cases.

4 General Discussion

The results of our experiments suggest that the tendency against which Hanlon's razor warns is in fact a real tendency in our judgments of others' motives. Across a range of contexts, we found evidence that people were inclined to expect that agents are motivated primarily by the worst of the reasons that they have for a given action, and that people expect others to be motivated by worse reasons than they are motivated by themselves. Whilst we were unable to replicate this effect for a broader range of vignettes than we considered in Experiment 1 (the first four from Experiment 1 plus an additional 8), it did replicate for the original vignettes when these were analysed separately. Additionally, it is clear that there was a difference in the new vignettes which could explain this failure: participants rated the good motives as more extreme than the bad motives in the new vignettes. This difference may have counteracted any tendency to be biased to expect the agent to act primarily on the worse motive. We also found that in all cases participants expected the agents in the story to be more likely to act on the motives they found more extreme, as we suspected we might when we set up our study. And finally, the second effect we reported, whereby participants expected others to be more

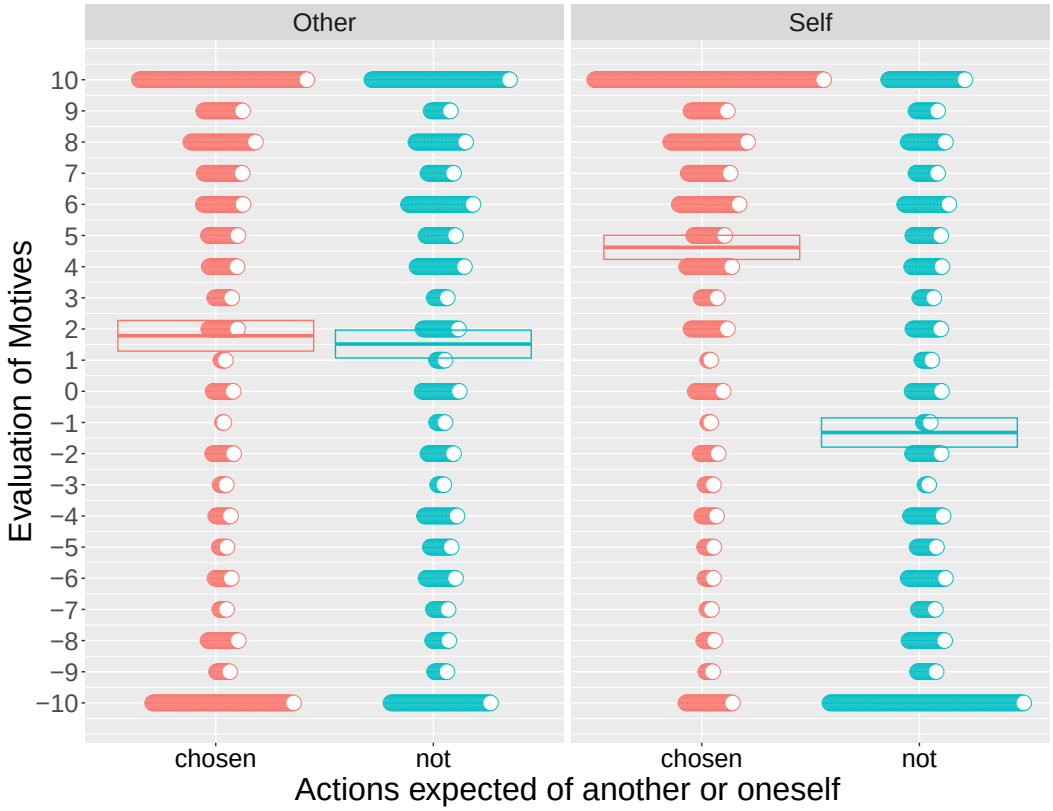


Figure 2: Experiment 2. The X axis shows the options that participants expected would be 'chosen' or 'not chosen' by the protagonist (on the left), or by themselves (on the right). The Y axis shows how participants rated the motives behind those actions. Disks represent individual participants' ratings of a given motive, boxes represent mean and 95% confidence intervals. Although the effect of rating on the option that participants expected the protagonist to choose did not replicate, participants did significantly expect the protagonist in the story to have a worse primary motive than they would have themselves.

likely to pursue actions they consider bad than they would prefer to do themselves replicated for all of the vignettes in Experiment 2.

The Worst Motive Fallacy fits naturally within the family of general negativity biases mentioned in Section 2 since it suggests, in effect, that we are also negatively biased in our moral evaluation of others' *motives*. Plausibly, we consider the worst reasons for actions to be the main motives because of our more general tendency to place greater focus on negative stimuli rather than positive, coupled with a more pronounced tendency to evaluate others' characters more negatively than we do our own.

We think that the Worst Motive Fallacy may arise due to the adaptive advantages that are gained from paying more attention to negative rather than positive aspects of others' behaviour, the wisdom of which is recommended by *another* common folk aphorism:[5] 'Hope for the best, but prepare for the worst'. A cognitive bias may be selected for when the errors in which it results are less costly than erring in the opposite direction (Haselton & Nettle, 2006). In general, the evolutionary story goes, it is more advantageous to pay attention to negative aspects of our environment and thereby avoid harm, even if that means failing to notice positive aspects and thereby missing good opportunities. In the context of the Worst Motive Fallacy, although being overly suspicious of others' motives may incur the cost of failing to take up co-operative opportunities, the cost of naïvely entering co-operative partnerships with malicious actors may be higher. It will therefore be more advantageous to err on the side of falsely believing that others have bad motives than to risk falsely believing that they have good motives.

Given the evolutionary account proposed here, several follow-up studies using similar methods could be conducted in order to explore the 'boundary conditions' of the effect we have identified. For example, one might expect to find an in-group/out-group effect that leads to more benign interpretations of others' motives when they are relatives, or in cases where more is known about the protagonist's history, prior behaviour or decision-making context. If we are generally predisposed to treat others with suspicion, then it seems plausible that this would extend to

[5]Usually attributed to Benjamin Disraeli.

an increased negativity bias when it comes to out-group attribution of motives. Such an effect would likely manifest along depressingly familiar prejudicial lines of gender, race, nationality, class and so on.

5 Conclusion

What about those philosophical theories, considered at the outset, that appeal to an agent's motives in the assessment of the morality of actions? The present study suggests that we should be cautious about appealing to our assessment of others' motives to judge the morality of their actions. The negativity bias we have uncovered casts doubt on the practicalities of any (meta-)ethical theory that recommends that our moral evaluation of others' actions should be rooted in our assessment of their motives. Similarly, the robust effect of 'extremeness' (i.e., the fact that people are more likely to attribute a motive when it is further away from morally 'neutral' in *either* direction) gives us a further reason to be wary. Given two competing motives we have reason to believe an agent to have in mind, it seems we are likely to consider not only the bad motive to be the main motive, but also the more morally extreme motive to be the main motive. Whilst such meta-ethical theories could still be correct that, objectively, an actor's motives play an essential role in the goodness of their actions, they should nonetheless carry a user-warning, as it were, that our subjective assessment of those motives may be far less reliable than is generally supposed.

These are matters for future investigation beyond the scope of this paper. Our present focus has been to formally identify this un-noticed fallacy and to demonstrate for the first time that there is a tendency for people actually to commit it. Of course, the reader might suspect that our *main* motive in writing the present paper was something else again: to publish in a top-ranking peer-reviewed journal for the purposes of fame, glory and career advancement. We suggest, however, that to suppose this would be to commit a fallacy, whose cause is a demonstrably commonplace cognitive bias.

Author Contributions

Joel Walmsley: conceptualization; methodology; writing: original draft, review and editing; resources; investigation.

Cathal O'Madagain: conceptualization; methodology; writing: original draft, review and editing; investigation; formal analysis; data curation.

Funding

This work was supported by a grant from the Department of Philosophy, University College Cork (JW), as well as the grant FrontCog: ANR-17-EURE-0017, held by the Institute d'Etudes Cognitives, Paris Sciences et Lettres (COM)

References

Baumeister, R. F., Bratlavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323-370.

Hamlin, J. K., & Baron, A. S. (2014). Agency attribution in infancy: Evidence for a negativity bias. *PLoS ONE*, *9*(5), 1-8. Retrieved from `http://dx.doi.org/10.1371%2Fjournal.pone.0096112` doi: 10.1371/journal.pone.0096112

Haselton, M., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, *10*((1)), 47-66. doi: 10.1207/s15327957pspr1001_3

Hewstone, M. (Ed.). (1983). *Attribution theory: Social and functional extensions*. Oxford: Blackwell.

Hibbing, J., Smith, K., & Alford, J. (2014). Differences in negativity bias underlie variations in political ideology. *Behavioral and Brain Sciences*, *37*(3), 297-307.

Jones, E., & Nisbett, R. (1971). The actor and the observer: Divergent perception of the causes of behavior. In E. Jones, D. Kanouse, H. Kelley, R. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior*. N.J.: General Learning Press.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69.

Kanouse, E., & Hanson, L. (1971). Negativity in evaluations. In E. Jones, D. Kanouse, H. Kelley, R. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior.* N.J.: General Learning Press.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*(2), 309–325.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47*(4), 2025–2047. doi: 10.1007/s11135-011-9640-9

Monson, T., & Snyder, M. (1977). Actors, observers, and the attribution process: Toward a reconceptualization. *Journal of Experimental Social Psychology*, *13*, 89–111.

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General*, *138*(4), 535-545.

Nisbett, R., Caputo, C., Legant, P., & Maracek, J. (1973). Behavior as seen by the actor and as seen by the observer. *Journal of Personality and Social Psychology*, *27*, 154–165.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind and Language*, *24*(5), 586–604.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173-220.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296-320.

Shaver, K. G. (1985). *The attribution of blame: Causal, responsibility and blameworthiness.* New York: Springer.

Sidgwick, H. (1884). *The methods of ethics (3rd edition).* London: MacMillan and

Co.

Slote, M. (1995). Agent-based virtue ethics. *Midwest Studies in Philosophy*, *20*(1), 83–101.

—◇◇—

Appendix 1 (SOM-R): Vignettes

Note: each vignette is constructed so that the different motives are presented using the same number of words (to rule out the possibility of participants simply selecting—or avoiding—the course of action that was quickest to read). Each vignette presents the competing actions and motives three times, so the experiment used eight versions of each vignette, with all the variations of the order of presentation (to rule out the possibility of participants selecting whichever option was presented first, or last). Experiment 1 used only the first four vignettes listed here; Experiment 2 used all twelve.

"Vacation"

> Patrick is studying Spanish in university, and making plans for the summer. He wants to go to a Spanish language school in Spain, because the classes that he will take there will help him to do better in his studies, and also because he thinks he will get to cheat on his girlfriend with the girls he meets there.
>
> In the end, the course in Spain is cancelled, so Patrick cannot go. He has two other options.
>
> He could go to stay with his cousins in Argentina, where he will learn Spanish that will help with his studies, but where he does not expect that he will meet girls. Or, he could go to stay for the summer in France, where he expects to meet girls that he can cheat with, but where he will not learn Spanish to help with school.
>
> Which option do you think Patrick will pick?
>
> - Argentina, to help with his Spanish.
> - France, to cheat on his girlfriend.

"Trip to Town"

> Simon wants to take the car into town early in the morning. He has two reasons. He wants to go to the shops in town, in order to pick up a going away present for his neighbor who will be moving out that day. He also wants to get to the welfare office where he knows pensioners are cashing their monthly pension cheques, so that he can assault one and take their money.

However, it turns out that his car has broken down, so he can't take it.

He has two other options. He could take the bus which passes the shops in town, but does not go to the welfare office; this will allow him to get a present for his neighbor, but will not allow him to beat and rob a pensioner. Or, he could take the train which goes to the welfare office but does not pass the shops in town; this would mean that he could beat and rob a pensioner, but that he wouldn't be able to get his neighbor a present.

Which option do you think Simon will pick?

- The bus, so he can get his neighbor a present.
- The train, so he can beat and rob a pensioner.

"Party"

Sally is going to Kate's birthday party, and has to decide what to wear. She decides to wear her red dress for two reasons. First, because Sally's mother gave her the dress and she wants to make her mother happy, and second, because she knows it's a nicer dress than Kate's and she wants to embarrass Kate.

However, when the date of the party comes around, Sally cannot find her red dress.

Instead, she has two other options. She could wear a blue dress that she bought last year, which will very likely embarrass Kate, but will not make her mother particularly happy. Or, she could wear a pair of jeans her mother bought her, which will make her mother happy but will not be likely to embarrass Kate.

Which do you think she'll pick?

- The New Jeans, to please her mother
- The Blue Dress, which will embarrass Kate

"Election"

A politician has some funding left over from her campaign, and she decides to use it to hire a computer engineer that she knows. She does this for two reasons. First, the engineer has recently lost his job and is in need of new work, and the politician wants to help him out. Second, the politician wants the engineer to send misleading messages to her opponent's supporters to send them to vote on the wrong day.

When she describes the work to the engineer, however, the engineer says he will not do it.

The politician has two further options. She could hire the unemployed engineer anyhow, to do ordinary computer maintenance work. This will help the engineer, who needs income, but won't help the politician to mislead voters. Or, she could hire a computer hacker who has no problem sending misleading messages. This will help the politician to mislead voters, but will not help out the unemployed engineer.

Which option do you think the politician will pick?

- Hire the engineer, to give him work
- Hire the hacker, to mislead the voters

"NGO"

Bob is looking for a new job, and an NGO offers him a job vaccinating and taking care of children in a poor country. Bob is very tempted to accept it, both because it will allow him to help those in need as a part of his work, and also because taking the job will make his elder brother jealous, who has always wanted to travel.

However, in the end the NGO retracts their offer before Bob can accept it.

Bob soon receives two more offers of work. First, another NGO offers Bob a job taking care of poor children, in his home town. This won't make his brother jealous, but it will help those in need. Second, he receives an offer from a travel agency, travelling to review possible locations for holidays. This will make his brother jealous, but it won't help those in need.

What do you think Bob will do?

- Take the NGO position, which will help those in need
- Take the travel-agency job, which will make his brother jealous

"Cheers"

Mike is the owner of a bar where Gary is a regular customer. One evening, when the bar is almost empty, Gary runs out of money, and so cannot afford to buy any more beer. Mike wants to give Gary some free whiskey instead, for two reasons. First, because he wants to do a good deed for a regular customer. Second, because it will get Gary drunk quickly, so Mike can go home.

In the end, Mike discovers that he has run out of whiskey, and so he cannot give any to Gary.

There are two options instead. First, Mike could give Gary some free orange juice. This would be doing a good deed for a regular customer, but won't allow him to throw Gary out early. Second, Mike could tell Gary they've run out of beer. This means he could go home early, but it wouldn't be doing a good deed for a regular customer.

What do you think he will do?

- Give Gary some orange juice on the house, to help out a regular customer.
- Tell Gary they've run out of beer, so that he can go home early.

"Ice Cream"

Mario owns an ice-cream shop, and is planning on opening a second branch. He hopes to open on the corner of the village for two reasons. First, because the owner of the building is a pensioner who could really use the rent money, and Mario would like to help him out. Second, because his biggest competitor, Gino's Sorbets, has a shop immediately next door, and if Mario opens here he will take some of Gino's business.

However, in the end Mario can't get planning permission to open in this location.

He has two other options. First, he could open in another building on the opposite side of the village, owned by the same pensioner. This will help out the old man, but won't hurt Gino's business. Second, he could open on the other side of Gino's shop, in a building owned by a young entrepreneur. This won't help the pensioner, but will damage Gino's business.

Where do you think he will open his shop?

- In the building owned by the pensioner, to help him out.
- In the building next to Gino's Sorbets, to beat the competition

.

"Lazy Prof"

A professor is on his way to class for which is he unprepared, when he sees a pedestrian get struck by a tram. He rushes over to the site of

the accident, both because he wants to offer to take the pedestrian to hospital, and also because he wants to have an excuse to cancel his class.

Once he gets to the pedestrian, however, two nurses arrive at the same time who are happy to take the pedestrian to hospital.

At this point the professor could offer to drive the pedestrian home from hospital later on. This would be very helpful to the pedestrian, but would provide the professor with no excuse to cancel his class. Or, he could accompany the pedestrian to the hospital along with the nurses. This will not be particularly helpful to the pedestrian, but it will give the professor an excuse to cancel class.

What do you think he will do?

- Offer to come back later, to drive the pedestrian home.
- Accompany the group to the hospital now, to avoid class.

"New Car"

Mike wants to buy a used car. He's interested in buying his friend's jeep, because his friend needs to make some money and Mike wants to help, and also because the jeep will make him look cool when he drives around town.

However, it turns out that in fact the jeep has a damaged chassis, and won't pass a safety test so that Mike can't buy it.

There are two more options, that will cost Mike the same. First, he could buy a three door hatchback owned by his friend. This won't make him look cool, but it will help his friend. Second, he could buy a used Mustang he has seen on eBay. This will make him look really cool, but it won't help his friend out.

What do you think he will do?

- Buy the hatchback, to help his friend
- Buy the used Mustang, to look cool

"Seamus"

Seamus is an inventor living on a small island where most of his neighbors are potato farmers. He has thought of a new machine for planting potatoes. He wants to patent the machine and to get it built, both because he thinks this will make life easier for the potato-farmers that live on his island, and also because he expects that selling the machine will allow him to make a lot of money.

However, Seamus discovers that a similar machine has already been built, and he cannot get the patent.

Seamus has two more ideas that he is considering working on. One is a strategy for improving potato harvests, that allows farmers to tell which are the best times to plant over the year. The strategy cannot be sold, so Seamus won't make money from working on it, but he would help the farmers on his island if he perfects it. The other is a computer app that allows users to efficiently predict market trends, and increase their luck in betting on stocks and trades. Perfecting and selling this app would make Seamus a lot of money, but of course it will not help the potato-planting farmers that live on his island.

What do you think Seamus will spend his time on?

- The crop-planting strategy, to benefit the islanders
- The investment app, to make big money

"Surveillance"

The president of a small country has requested that the local telephone company allow his government to surveil citizens' phone calls. He wants to do this to increase the chances that the police will catch criminals and make the country safer, but also because he wants to find out how people are voting in different areas by listening to their conversations.

The communications company does not want to allow the government access to the phones, and succeeds in protecting their customers privacy.

The president has two further plans. First, he could donate money to a charity that helps youths to stay out of gangs. This will make the country safer, but won't help with his re-election. Second, he could hire a company that predicts how people vote by monitoring their online behavior. This will help his re-election campaign, but won't make the country safer.

He does not have enough money to do both. Which do you think he will do?

- Donate money to help youths stay out of gangs
- Hire the online monitoring company to help his re-election

"Teacher"

Mervyn is a teacher from Missouri. He wants to move to New York City for two reasons. First, because there's a great teacher training program

there, and he wants to learn new skills so that he can later return home to teach young people. Second, because he wants to experiment with a wide variety of hard drugs, and they are more plentiful, cheaper, and easier to get in New York City.

In the end, Mervyn cannot afford to move to New York City.

There are two more options. First, he could move to Kansas. There is an excellent teaching college in Kansas, but he won't have easy access to drugs there. Second he could move to Chicago. This will give him ready access to drugs, but the teaching college there is not very good.

What do you think he will do?

- Move to Kansas, for the great teaching college.
- Move to Chicago, to have access to drugs

.

—◇◇—

Appendix 2 (SOM-R): Further Details on Methods in Experiment 2

Experiment 2 was, like Experiment 1, conducted using the Qualtrics online survey platform, with participants recruited using Amazon's Mechanical Turk. Participants were paid 25c for their participation, which took about 1-2 minutes. We tested twelve vignettes that were counterbalanced (with the order in which motives and options were mentioned presented in eight different ways). Since this was three times as many as the first study which had originally aimed to have 320 participants, we aimed to collect data from three times as many participants: 960 participants. We ran three 'attention' questions at the beginning of each experiment, and excluded participants who did not correctly answer all of those questions. We also excluded participants who did not complete the whole test, or if a participant attempted to take the survey more than once. In total we collected 1657 responses and excluded 690 (41%), resulting in 967 responses. These were excluded for not passing attention questions (25.7%), taking the test more than once (14.6%) or not finishing the test (1.7%). The high drop rate raises some suspicions about the data quality, and the results should be considered in that light.

To analyse the new data we used a Generalised Linear Mixed Model in R package 3.4.3, using package lme4 and the logit link function. Again the full model included a dependent variable of the protagonist's choice (the option the protagonist in the story is expected by the participant to choose), as a binary variable (good or bad). The predictors were the difference score of the ratings of the two options, the difference score of the absolute rating of the options, and the participant's own choice. The full model included these factors plus a random intercept of vignette, while the null model lacked all terms except for the random intercept of vignette.

The full-null comparison was highly significant (N=967, χ^2 =141.98, df=4 $p<$ 0.0001). We then tested for an effect of the interaction between the rating and the 'absolute rating', as in the first study. There was no effect of the interaction and so we removed the term from the model. We now compared this reduced model to a further reduced model that lacked the term for 'rating'. This comparison proved insignificant. Finally, we compared the reduced model to a model further reduced by the exclusion of the participant's own choice, revealing an extremely significant improvement of fit (χ^2 =97.141, df=1 $p<$0.0001).

To explore why the first effect we found in Experiment 1 did not replicate, we subset the data into two groups: the vignettes that we had used in Experiment 1, and the ones only used in the Experiment 2. We applied the same model and procedure of analysis to both groups separately. The first group included the 332 participants who had seen the 'original vignettes' – namely those that were used in Experiment 1, which also appeared in Experiment 2 (see Appendix 1: Vacation, Trip to Town, Party, and Election). Here a full-null comparison was significant (N=332;

χ^2 =33.226, df=4 $p<$ 0.0001). The interaction between rating and absolute rating was not significant, and so it was removed. A comparison of models with and without rating, but where both included the absolute rating and the participant's own preference, showed a significant improvement of fit (χ^2 =5.9072, df=1 $p<$ 0.015). The effect of rating therefore replicated when considering the vignettes used in Experiment 1 alone.

Applying the same procedure to the new vignettes alone, the full-null comparison was significant (N=665; χ^2 =127.59, df=4 $p<$ 0.0001), the interaction was not significant, the effect of rating was not significant (χ^2 =1.328, df=1 p= 0.249), but we found an extremely significant effect of absolute rating (χ^2 =20.034, df=1 $p<$ 0.0001). This suggests that participants' high evaluation of the extremeness of the good motives compared to the bad in the new vignettes may have covered any bias they may have to expect the characters in the stories to act on the bad motives first. To check whether there was a difference in the absolute ratings of the good versus bad motives we checked the mean value of each group for the original and then the new vignettes. Participants rated the bad motives as having less extreme values on average than the good motives in the original vignettes as these featured in the replication (mean bad: 8.2; mean good: 7.45), but treated the good motives as being more 'extreme' than the bad motives in the new vignettes (mean bad: 5.63; mean good: 7.38). Finally we checked the mean value of the bad motives versus good motives overall in the replication study, and again found that the good motives were rated more extreme than the bad motives (mean bad: 6.51; mean good: 7.4). Once again this difference could explain why there was no overall effect of rating considering all twelve vignettes together, given that the overall difference in extremeness favoring the good motives will have counteracted any bias to pick the worst motive overall.

—◇◇—