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.

j.walmsley@ucc.ie

[2] Département d'Études Cognitives, École Normale Supérieure, Paris.

cathalcom@gmail.com

Abstract

In this paper, we describe a hitherto undocumented fallacy—in the sense of 'a mistake in reasoning'—that occurs when people assume that an agent's *worst* motive for an action is thereby their *main* motive. We call this the 'Worst Motive Fallacy.' We outline the results of an experimental study in which we demonstrate that the commission of this fallacy is also a hitherto undocumented cognitive bias to which people are systematically prone. 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; 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 demonstrate that people *are* subject to such a bias, which we call the 'Worst Motive Fallacy'. We outline the results of an experimental study showing that people systematically assume that an agent's *worst* motive for an action is thereby their *main* motive, and 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.

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 could be 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. But if an observer were to systematically suppose, all else being equal, that the worst of these motives was the main motive, this would be a mistake. The study we present here investigates whether we are biased toward making this mistake.

3 The Worst Motive Fallacy: An Experiment

To explore whether people have a cognitive bias towards committing the Worst Motive Fallacy, we ran an experiment. 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 which option the character in the story will pick, and asked to rate the goodness or badness of each motive. 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 motive.

3.1 Materials and Methods

Each participant was told 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 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. Setting the study up in this way has several advantages. It gives us a continuous scale of 'badness' of

[1] All four vignettes used in the study are included as Appendix 1.

motives to use as a predictor of whether a participant will expect a motive to be followed. A continuous scale is generally more sensitive statistically than a categorical measure, so taking this approach means we are more likely to identify an effect if it exists. Further, if the outcome were to prove significant, it allows us to say not simply that participants generally expect agents to do the worse of two options, but that there is a linear effect of 'badness'; the worse a motive is rated, the more likely it is to be expected of the agent. Finally, it allows us to be confident that participants are picking what they themselves regard as the worse motive, rather than seeing which action they pick and simply assuming they agree with us that this is the worst action.

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 as an alternative 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.

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 them-

selves if they were in the position of the protagonist. We included this question to allow us to rule out that participants were simply ascribing to the protagonists the course of action that the participants would prefer themselves. 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; that participants would generally express a preference for doing the opposite of the course of action they predict that the protagonist will take.

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 the appendix). 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 aimed to collect answers from at least 320 participants, since we had four vignettes and eight counterbalanced versions of each vignette, such that 320 participants would give us ten participants in each of the smallest cells of the study. We ran three 'atten-

tion' 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. 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.

3.2 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 as the dependent variable and the difference between the participant's ratings of the motives behind the two options as the main predictor, expecting that the worse a participant rated one motive relative to the other, the more likely they would be to expect the protagonist to pursue that option that satisfied that motive. We also wanted to include a control to measure whether participants were choosing the option they ranked simply as the most 'extreme' of the two. That is, if they rated one motive as -10 (toward 'bad'), and the other as +2 (toward 'good'), then we might expect they would pick the option satisfied by the motive rated at -10; but if they rated a motive as plus 10 toward good, and another at minus 2 toward bad, they would be more likely to pick the good motive. This would mean the participant was making a prediction not on the basis of which motive was worse, but on the basis of which motive which motive was more extreme. To rule this out, we included a measure of the difference in absolute ratings of the two options in the full model along with the interaction between this and the main predictor. We identified significance levels using likelihood ratio tests to compare the fit of different models.

3.3 Results

What we wanted to narrow in on was the effect of a participant's rating of the motives described on their expectation that the protagonist would act according

to that motive. We expected that the worse a participant rated an agent's motives, the more likely they would be to expect the agent to act according to that motive. To test for the main effect of 'badness', we therefore compared a model that included the participant's rating of the agent's motives against a 'null model' that did not include that rating. We found that including the rating in the analysis significantly improved the fit of the model (n=323, χ^2 = -6.3425, p=0.0415). This outcome confirms our hypothesis: the worse a participant rated the motive of the agent in the story, the more likely the participant was to expect the agent to choose the course of action that satisfied that motive. These results are plotted in Figure 1.

We also found a significant effect of the participant's *own* preference: participants significantly chose the opposite of what they expected the protagonist to choose (n=323, LRT = 6.866, p= 0.008), as can be seen in Figure 2. This is consistent with the literature showing that subjects are inclined to expect that they would act better than they expect others to act.

There was also an effect of the magnitude of a participant's absolute rating of the options (the distance from zero either positively or negatively): participants were more likely to expect the protagonist to take the action that they rated more extreme of the two options (n=323, χ^2 = -9.5331, p=0.008). However, there was no interaction between this absolute rating and the main predictor (n=323, χ^2=-0.54194, p = 0.4616). This means that although the 'extremeness' of a rating affected a participant's prediction about what a protagonist will do, this does not account for the tendency in the data for the participants to choose the worst motive.

There was no effect of vignette; all vignettes brought out the same bias in participants. There was no effect of duration of the experiment (how long participants spent).

Figure 1: The X axis shows the courses of action that participants predicted would be 'chosen' or 'not chosen' by the protagonist. The Y axis shows how participants rated the motives behind those actions. Disks represent individual participants' ratings of a given motive. 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.

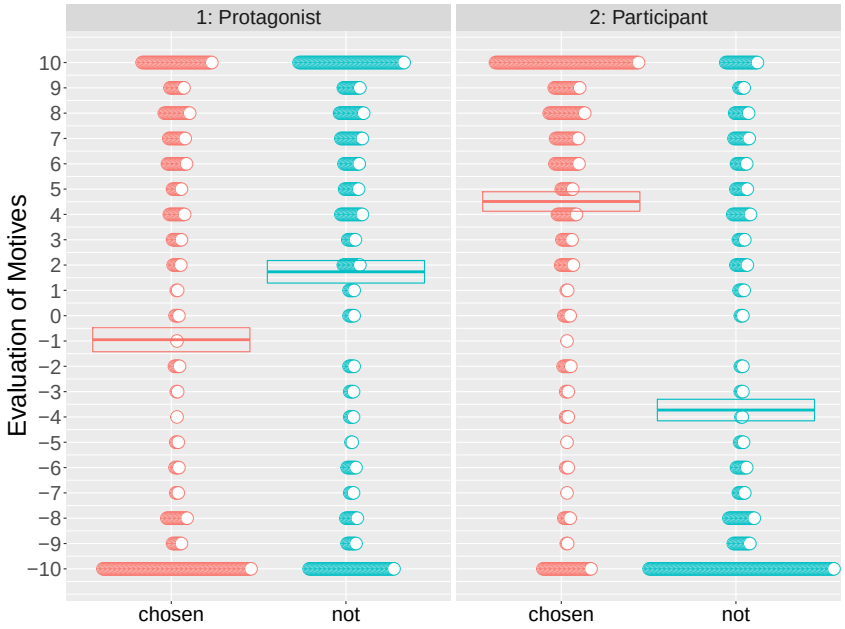

Figure 2: On the left are the data points from Figure 1 collapsed across all four vignettes, showing the tendency to commit the Worst Motive Fallacy. On the right are the options that participants stated they would prefer themselves. Here it is clear that although participants expected protagonists to make choices driven by the motives they rated as worst, the participants themselves tended to prefer the options driven by motives they rated as best.

4 General Discussion

The results of our experiment demonstrate 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, people are inclined to expect that agents are motivated primarily by the worst of the reasons that they have for a given action. The Worst Motive Fallacy fits naturally within the family of biases mentioned in Section 2 since it is, in effect, the discovery 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 tendency to evaluate others' characters more negatively than we do our own. Our finding that the participants in our experiment claimed that they themselves would have preferred to pursue the more positively rated course of action, in contrast to their interpretation of the protagonist in the stories, supports this interpretation.

We also 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:[2] '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.

What about those philosophical theories, considered at the outset, that ap-

[2]Usually attributed to Benjamin Disraeli.

peal 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 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. Whilst such 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 aims were simply to outline the contours of a hitherto un-noticed fallacy and to demonstrate that there is a statistically significant 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.

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.

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.

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.

Acknowledgements & Funding

—Redacted for anonymous review; to be added upon acceptance.

—◇◇—

Appendix 1: Four Vignettes Used

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).

"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