Moral bookkeeping

Igor Douven, Frank Hindriks, Sylvia Wenmackers

Moral Bookkeeping

Igor Douven
IHPST / CNRS
igor.douven@univ-paris1.fr

Frank Hindriks
Faculty of Philosophy, University of Groningen
f.a.hindriks@rug.nl

Sylvia Wenmackers
Institute of Philosophy, University of Leuven
sylvia.wenmackers@kuleuven.be

Abstract

There is widespread agreement among philosophers about the Mens Rea Asymmetry (MRA), according to which praise requires intent, whereas blame does not. However, there is evidence showing that MRA is descriptively inadequate. We hypothesize that the violations of MRA found in the experimental literature are due to what we call "moral compositionality," by which we mean that people evaluate the component parts of a moral problem separately and then reach an overall verdict by aggregating the verdicts on the component parts. We have subjected this hypothesis to the test and here report the results of our experiment. We explore several explanations of the experimental findings and conclude that they present a puzzle to moral theory.

1 Moral compositionality

When people praise or blame someone, they evaluate what this person did and what she thought about this. More precisely, they evaluate whether what she did was right or wrong and which attitude she had toward her action. The blame attributed to a murderer will be sensitive not only to the act—that the agent killed someone—but also to the agent's attitude—that the killing was premeditated. This distinction between action and attitude plays an important role in criminal law where, in addition to a guilty act (*actus reus*), a guilty mind (*mens rea*) is required for a conviction. What role these components play exactly in how people should attribute moral responsibility is a matter of ongoing controversy. Meanwhile, most philosophers do agree about what one could call "the Mens Rea Asymmetry" (MRA), according to which praise requires intent, whereas blame does not.

However, there is evidence that the folk attribute praise and blame in a way that conflicts with MRA (for overviews, see Anderson et al. [2020] and Guglielmo and Malle [2019]). For instance, Bertram Malle [2006] reports that people attribute a substantial amount of praise to agents who accidentally bring about a good outcome that they had no reason to expect. In fact, they ascribe more praise when the outcome is good than blame when the outcome is

bad.[1] These folk attributions are not in line with MRA, which requires foresight for praise and blame. The question arises as to why they are inconsistent.

We propose a new hypothesis according to which violations of MRA are due to the fact that, in attributing moral responsibility, people first evaluate how a person acted and what her attitude toward the act was separately and then combine those evaluations to reach their overall moral verdict. We use the term "moral compositionality" to refer to this hypothesis.

Moral compositionality is an open-ended idea that leaves room for a wide range of precisifications. In this paper, we will be mostly concerned with the arguably simplest way of operationalizing it. According to what we call "the Moral Bookkeeping Hypothesis" (MBH), the total responsibility (TR) that the folk attribute to an agent for performing a particular action is simply the *sum* of the attributions to its components. As we consider only unintended side effects, from this point onward we will use the term "outcome" instead of "action." Thus, the two components we distinguish are outcome and attitude. Correspondingly, we distinguish between two kinds of component responsibilities, namely, outcome responsibility (OR) and attitude responsibility (AR). The Moral Bookkeeping Hypothesis then amounts to the following equation:

$$\text{TR} = \text{OR} + \text{AR}. \tag{MBH}$$

Among other things, this hypothesis entails that the total amount of blame people ascribe to an indifferent chairman who harms the environment is the sum of the blame people attribute to him for harming the environment and the blame they ascribe to him for his indifference. (We are referring here, of course, to Joshua Knobe's [2003] celebrated chairman scenario, which will also prominently figure in the following.)

So, our hypothesis explains the previously observed violations of MRA and the operationalization is specific enough to generate numerically precise predictions, which can be tested empirically. Investigating these matters is worthwhile because an explanation of said violations can reasonably be expected to enhance our understanding of both how moral responsibility is actually attributed and how it should be ascribed. In Section 4, we report the findings of an experiment designed to put MBH to the test. Whereas existing research provides for qualitative predictions such as that intentionality increases blame (Malle, Guglielmo, and Monroe [2014]), our hypothesis generates quantitative predictions. Thereby we hope to shed new light on the structure of folk attributions. In Section 5, we ask why our findings are inconsistent with MRA. As this asymmetry is based on moral theories that are taken to be closely related to our folk practices, a discrepancy between the asymmetry and the practices is alarming. We explore a number of explanations and conclude that the findings conflict with moral theory.

2 The Mens Rea Asymmetry

According to MRA, the requirements for praise are more demanding than those for blame: praise requires intent, whereas blame does not. More specifically, an agent is praiseworthy for performing a good action only if she intended that action, whereas an agent need not intend a bad action in order to be blameworthy for performing it. We take Michael Stocker [1973:60] to subscribe to MRA when he claims that "an act is not praiseworthy nor, therefore,

[1] In the scenario that was used, an employee accidentally calls someone at home. Depending on whether this person likes or dislikes receiving phone calls at home, people attribute praise or blame to the employee (M = 2.5 and M = 1.5 respectively on a 7-point Likert scale with anchors "a little" and "a lot"; Malle [2006:94]).

is it supererogatory unless done with a good intention. . . . [A] blameworthy act need not be done with a bad intention."[2]

In this paper, we are concerned with the side effects of people's actions. We take it to be uncontroversial that the side effects of an action can bear on the moral quality of that action. The underlying idea is that an agent should attend to the morally significant side effects of her action when deliberating about what to do.[3] For the purposes of this paper, then, we rely on a formulation of MRA in terms of outcomes:

Mens Rea Asymmetry (MRA): Praise requires an intention to bring about a good outcome, while blame does not require an intention to bring about a bad outcome.

Outcomes can be good or bad because they are beneficial or harmful. We also allow for an outcome to be good or bad in virtue of the maxims or reasons on which an agent acted.

MRA will play a central role in our interpretation of the empirical findings we report in this paper. Hence, it is important to be clear about what motivates MRA. The most straightforward defense of MRA appeals to the reasons for which someone acts. A praiseworthy person does the right thing for the right reasons (Wolf [1990:84], Sher [2009:142]). A blameworthy person flouts the reasons she has, either by ignoring them or by defying them and acting contrary to those reasons (Scanlon [1998:271]). These two claims directly bear on the agent's motivation. The first claim entails that an agent has to be motivated to bring about a good outcome in order to be praiseworthy. Blame, on the other hand, does not require being motivated to bring about the bad outcome, as a lack of motivation for avoiding it will do (Hindriks [2008:632]). These two claims imply that praise requires intent, whereas blame does not.

In much the same vein, Nomy Arpaly [2002] argues that, in order to be praiseworthy, an agent needs to exhibit appropriate moral concern or a good will. Blameworthiness is consistent not only with ill will, but also with a deficiency in good will: malice forms a good basis for blame, but so does indifference (see Strawson [1962:4 ff] for a similar proposal). Arpaly [2002:231] explicates good will in terms of responsiveness to moral reasons, and ill will in terms of responsiveness to sinister reasons. Someone's will is deficient when she is insufficiently responsive to moral reasons. Acting for the right reasons entails intent; not acting out of malice does not.

Outcomes that are due to recklessness or negligence provide further support for MRA. We blame people when the outcomes are harmful, but we do not praise them when they are beneficial. This is consistent with the claim that praise requires intent. At the same time, it suggests that inattentiveness suffices for blame and that foresight is not required. Note that, as it reveals a lack of moral concern, inattentiveness as such is blameworthy.

The only philosopher we can think of who might reject MRA is John Mackie. He defends what he calls "the straight rule of responsibility," according to which "an agent is responsible for all and only his intentional actions" (Mackie [1977:208]). In addition to intended actions, he takes intentional actions to encompass unintended but foreseen or accepted events,

[2] Note that MRA does not claim that an intended good outcome is sufficient for praise. For all we know, praise is in order only when the action performed exceeded legitimate moral demands (Sher [2009:13], Smith [2008:281 n15]). Furthermore, MRA leaves open that an intention adds to the degree to which the agent is blameworthy.

[3] Hindriks [2008] refers to this as the Side Effect Deliberation Norm.

actions that are "obliquely intended."[4] As all the examples Mackie discusses concern harmful events, however, it remains unclear whether he in fact accepts that obliquely intended beneficial events might be praiseworthy.

3 How to test the Moral Bookkeeping Hypothesis

As mentioned in the introduction, Malle [2006] finds that people attribute praise and blame for a good or bad outcome in the absence of any attitude toward the outcome. The scenario that Malle discusses is one in which the agent intends to call a particular person. In one condition, the agent calls the person she intends to call. The other condition involves a central switchboard error due to which she ends up calling someone she did not want to call. This scenario is of limited use for our purposes. The reason for this is that in both conditions the agent intends to call someone. This means that there is no guarantee that people who attribute responsibility in the condition involving the switchboard error respond to the outcome only. The intention to call someone else could influence the attributions as well. In light of this, we use different scenarios that present a stronger test of the Moral Bookkeeping Hypothesis.

In order for the participants of our experiment to respond to the outcome of an action in isolation of any attitudes the agent might have, we resort to scenarios that involve a side effect. More specifically, we use the type of scenarios with which Joshua Knobe discovered an asymmetry concerning our attributions of intentionality. Knobe presented data showing that people tend to evaluate in an asymmetric way a person's bringing about a negative side effect and that person's bringing about a positive side effect when that person expresses indifference with respect to that effect. While people are strongly inclined to characterize the former as an intentional action, they are much less strongly inclined to regard the latter as an intentional action. This asymmetry has become known as "the Knobe Effect." This finding has been replicated in numerous studies (see Knobe [2010] for references). The scenario with which it was first elicited—and that since has been frequently used in replications of Knobe's results—is this:

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment." The chairman of the board answered, "I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was harmed. (Knobe [2003:191])

Knobe also presented his participants with a help version of this story, the only difference being that, in it, the environment is said to be helped as a side effect (and the chairman expresses indifference toward helping the environment), rather than to be harmed.

In addition to the Knobe Effect, Knobe [2003] reports a parallel asymmetry concerning the responsibility attributions people make: in the harm version, the modal response is that the chairman is very blameworthy, while in the help version, the modal response is that he is at most moderately praiseworthy. Whereas the blame is consistent with MRA, the praise that is attributed is not. Indifference does not, after all, satisfy the volitional requirement of

[4] In criminal law, the court is entitled to infer intention from foresight of harm. In effect, this means that the court employs a similarly broad conception of intention; see Ashworth [2006:176 ff].

praise. The discrepancy in the case of praise is in line with Malle's finding that was discussed in the introduction.[5]

Knobe's chairman scenario is useful for our purposes because it can easily be modified in such a way as to exclude any attitudes on the agent's part because they pertain to a side effect. The vignette Knobe used can be modified by, first, eliminating the reference to the vice-president informing the chairman about the effect that the strategy might have on the environment, and second, removing the subsequent expression of indifference on the part of the chairman.

MBH provides an explanation of the aforementioned inconsistency with MRA that starts from the observation that the scenario in either version contains components that can be evaluated separately from each other. On the one hand, there are the outcomes of the action. The intended outcome of the chairman's decision is the same in the harm and in the help version, and if it were just for this effect, the story would probably hardly elicit a moral response: every reasonable person with the chairman's responsibilities would have decided likewise. On the other hand, the side effect does elicit a moral response. While the chairman is not being paid to be concerned about the environment, most of us feel that he should nonetheless be concerned about it, and should not try to maximize the profits of his company at the expense of the environment. Naturally, if he can maximize the profits of the company and at the same time thereby help the environment, then so much the better.

On the other hand, there is the expression of indifference about the environment, which is the same in both versions of the story. This attitude can be evaluated as a distinct component of the story in isolation from the outcome of the action. This attitude of indifference appears morally reprehensible on its own (Adams and Steadman [2004], Wagner [2014]). Strikingly, indifference has not been studied as such. Instead, the attitude of indifferent agents such as Knobe's chairman has been considered merely as one of foresight of the side-effect (Guglielmo and Malle [2019]).[6]

That the two aspects—side effect and expression of indifference, outcome and attitude—*can* be evaluated separately does not mean that, when asked for a moral verdict on the chairman, people do evaluate them separately when both are present. That makes MBH, which asserts that this is how people reach their verdict indeed, a substantive hypothesis.

To see how the hypothesis makes the asymmetry in responsibility attributions that Knobe's experiment uncovered a matter of course, note that in the harm version of the chairman story, the components may plausibly be assumed to be both evaluated negatively, thereby yielding two wrongs, while in the help version one component (the side effect) is plausibly evaluated positively and the other component (the expression of indifference) is still evaluated negatively, yielding one right and one wrong. Thus, if MBH holds true, then even if no asymmetry is involved in evaluating the components—in the harm version the outcome is evaluated negatively to the same extent that, in the help version, the outcome is evaluated positively, and the expression of indifference is evaluated the same in the two versions—we

[5] Knobe [2003:193] also reports that the intentionality and responsibility attributions were relatively highly correlated ($r(120) = .53$, $p < .001$). This has led people to believe that the responsibility attributions explain the intentionality attributions. However, Hindriks, Douven, and Singmann [2016] report an experiment showing that the two attributions are correlated only in the harm condition, and not in the help condition.

[6] Guglielmo and Malle do observe that blame judgments are sensitive to a wider range of mental states than praise judgments. They take this to be reflected in the fact that "*mens rea* terms [used] in the law and everyday life . . . such . . . as *knowingly*, *negligent*, *reckless* . . . do not have positive counterparts (Guglielmo and Malle [2019:3]).

still end up with an asymmetry in our overall verdict. Putting the point schematically, and in its simplest form, it is as though we have $-a - b$ in the harm case, and $-a + b$ in the help case, and not $a + b$, as would be required to have symmetry in the overall verdicts in the harm and help versions.

That MBH provides an exceedingly simple explanation for the designated asymmetry is a point in its favor. But in the following, we want to subject the hypothesis to systematic testing. More exactly, we want to present the harm and help versions of the chairman story as well as of two similar stories and see whether the overall verdict in each case is (at least roughly) the sum of the verdicts on the morally evaluable component parts.

4 Experiment

To the best of our knowledge, so far no experimenters have tried to also elicit moral verdicts on the component parts of the two versions of the chairman story, that is, the chairman's decision to act and his expression of indifference.[7] The experiment we conducted had the purpose of filling that lacuna, with an eye toward testing MBH. The present section describes, and reports the results of, this experiment.

An ideal design for our purposes would be one in which participants were presented either with the harm or the help version of the chairman story and asked to evaluate the relevant component parts both separately and in tandem, where then the participants would be able to really consider the component parts as standing on their own and not be influenced by their joint occurrence in the story. Just as good would be a set-up in which participants were presented the full story (be it in the harm or help version), asked for their verdict, were made to forget their evaluation, then presented with a partial version of the story which mentions the effect of the act but not the expression of indifference, were asked for their verdict, were made to forget that evaluation as well, and finally were presented with another partial version of the story that contains the statement of the attitude but not that of the effect, and asked for their verdict about the attitude. In actuality, this is difficult to realize, and we were concerned that asking participants for their verdicts on the component parts as well as for their overall verdict would to lead to carryover effects in that their answer to one question influenced their answer to the next question.

Nevertheless, precisely because it is so straightforward to ask participants both for separate evaluations of the outcome and attitude components and for an overall evaluation, we started with what we thought of as a pre-test that did exactly this: presenting every participant with the harm or help version of the chairman story as well as with the harm or help versions of two other stories structurally almost identical to the chairman story (see below), and then asking for verdicts on the morally evaluable components of the three stories as well as for an overall verdict. The pre-test is reported in the Supplementary Materials to this paper, which can be downloaded from the following repository: `https://osf.io/7xcae/?view_only=9a8e747e295f46a58bbd1f4e64f02e66`. As explained there, although the results prima facie support the idea of moral compositionality and even the more specific MBH, they also raise a number of alarms. Notably, the earlier-mentioned praise/blame asymmetry that was observed in *all* relevant previous experimental work was absent from

[7] Pellizzoni, Girotto, and Surian [2010] use a scenario in which the chairman is not informed about the side effect (even though the vice-president knows about it). However, they are only concerned with intentionality and do not ask questions about responsibility.

our data. More problematical still was the seemingly paradoxical result that, for all three stories, participants in the harm condition tended to judge more harshly the expression of indifference toward the negative side effect of a decision (e.g., the damage being done to the environment, in the chairman story) when considered on its own than the combination of expressing that attitude *and* taking the decision leading to the negative side effect. One would think that, regardless of whether there is anything to moral compositionality or MBH, the amount of blame apportioned to someone for committing two wrongs should in general exceed the amount of blame apportioned for each of the wrongs individually.

As remarked in the Supplementary Materials, the results from the pre-test suggest some interesting avenues for future research which, however, are largely orthogonal to our present focus on moral compositionality and MBH. Thus, instead of delving deeper into what may have caused the somewhat puzzling results from the pre-test, we move on to describe another experiment whose design was meant to prevent carryover effects from the start by introducing additional versions of the three stories used in he pre-test in which either the outcome occurs without the attitude, or the attitude without the outcome.

4.1 Method

Participants

There were 292 participants in the study, which was run online using the Qualtrics platform. In return for their cooperation, the participants were paid a small amount of money. All participants were from Australia, Canada, the United Kingdom, or the United States. Repeat participation was prevented.

Data from participants who indicated that they were nonnative speakers of English were excluded from the analysis, as were data from participants who indicated that they had not responded seriously (a question concerning this appeared at the end of the survey, as recommended by Aust et al. [2013]) and data from participants who had returned incomplete response sets. This left us with 264 participants. These participants spent on average 338 seconds on the survey (SD = 3036 s). From this group, we excluded from the analysis the fastest as well as the slowest 5 percent responders, leaving us with 251 participants. These remaining participants spent on average 144 seconds on the survey (SD = 54 s). 175 of them were female; 173 had a university education, 77 had a high school or secondary school education, one participant only had a primary school education. The mean age of these participants was 39 years (SD = 12). Participants included in the analysis did not differ significantly from excluded participants in age, sex, or level of education.

Materials and design

To forestall potential carryover effects, two extra stories were invented that were meant to closely parallel Knobe's chairman story. Thereby, each participant could be asked for his or her overall verdict on a person in one story, his or her verdict on the act with the (either positive or negative) side effect in the context of a different story, and his or her verdict on the expression of indifference about a similar side effect in the context of a third story.

The stories that were used in the experiment, along with the chairman story, stick closely to the main structure of Knobe's story in the sense that they are about someone's taking a profitable decision that has either a negative or a positive side effect, about which the person expresses lack of concern in either case. The main differences between the three stories concern the personae and the settings.

Specifically, the following two stories were invented for the experiment:

> The mayor of a small city has to decide whether or not to build a bridge over a nearby river. One of his advisors tells him that doing so will improve the flow of traffic in the city but will also negatively affect the wildlife around the city. The mayor responds: "I don't care about what will happen to the wildlife. I want to improve the flow of traffic. So, let's build the bridge." The bridge was built. As a result, the wildlife around the city was negatively affected.

> The minister of infrastructure of an African country advises the president of the country to construct a big dam to improve the country's irrigation systems. The minister tells the president that the dam will adversely impact the irrigation systems in the neighboring countries. The president responds that he doesn't care about the other countries, and he orders that the dam be built. The dam is built, and the irrigation systems in the neighboring countries suffer as a result.

These are the harm versions of the stories, which had corresponding help versions, with again the only difference being that, in those versions, the wildlife was positively affected and the irrigation systems in the neighboring countries benefited, respectively. From here on, a (harm or help) version of a story is referred to as "a scenario"; thus, all in all there were six scenarios.

Furthermore, there were three *modes of presentation* in which each of the scenarios appeared in the experiment. The above (harm) scenarios all mention both the side effect or outcome (O) and the person's attitude of indifference (A) toward that side effect. This is one mode of presentation, which here will be called "the OA mode." Next to this mode of presentation, there was a mode that mentioned only the side effect or outcome as well as a mode that mentioned only the attitude. Call these "the O mode" and "the A mode," respectively.

For example, the O mode of presentation of the original Knobe (harm) scenario reads as follows:

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits." The chairman of the board was all for it, saying: "I want to make as much profit as I can. Let's start the new program." They started the new program. As an effect of the program, the environment was harmed.

Here, there is only mention of the chairman's attitude to the intended effect of the decision, not of his attitude to the side effect.

The corresponding A mode of presentation reads as follows:

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits." The vice-president mentioned to the chairman that the effects of the program on the environment are unknown. The chairman answered, "I don't care at all about the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program.

Note that in this mode there is no mention of the side effect of the decision—it is explicitly stated to be unknown what the side effects on the environment will be, if there will be any—but now the chairman does express his indifference toward whatever those side effects may be.

Also note that the A mode of presentation cannot have both a harm and a help version, simply because it is left unspecified whether the side effect will be positive or negative (or nonexistent).

As a result, there were in total 15 different vignettes used in the experiment: $\#\big(\{\text{chairman, mayor, president}\} \times \{\text{harm, help}\} \times \{\text{OA, O}\} + \{\text{chairman, mayor, president}\} \times \{\text{A}\}\big) = 12 + 3$.

Every participant in the study was presented with three vignettes, each instantiating a different story (chairman/mayor/president), and each being in a different mode of presentation. Also, the OA and O modes were either both of the harm variety or both of the help variety. Figure 1 gives an example of the three consecutive screens a participant might have been shown.[8] There are 12 possible combinations of three different modes of presentation of three different scenarios, if the scenarios are required to be all of either the harm variety or the help variety. To make the design fully orthogonal, all these combinations were used, and participants were randomly assigned to one of them.

Each of the three scenarios that were presented to the participant appeared on a separate screen. On the same screen, the participant was asked to indicate how blameworthy or praiseworthy the chairman/mayor/president was for acting as he did. The participant was asked to indicate his or her verdict by means of a slider, which could be moved on a scale ranging from −10 to 10, where it was explicitly stated that −10 indicates maximal blameworthiness and 10 indicates maximal praiseworthiness. To forestall response bias, the initial position of the slider was always at the midpoint of the scale. The scenarios appeared in an individually randomized order.

The aim was to see whether the participant's rating of the blameworthiness/praiseworthiness of the person in whichever scenario had been presented to the participant in the OA mode equaled the sum of the ratings of the blameworthiness/praiseworthiness of the persons in the scenarios that had been presented to the participant in one of the other modes. To that end, it was planned to conduct a t-test on the mean of the verdicts in the OA cases and the mean of the sums of the verdicts in the O and A cases. MBH predicts that these means will not differ significantly.

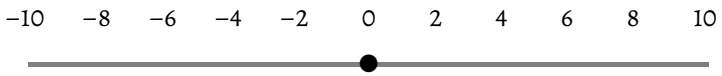
To make sure that the test has the conventionally required 80 percent chance of detecting an effect of adding up moral verdicts on the component parts of the stories, we did a power calculation in advance. According to Cohen [1988], an effect size of $d = .2$ is considered to be small, which in our case would mean that eliciting moral verdicts on the component parts and adding those up makes little difference as compared to eliciting the overall verdict. Reckoning with the possibility that the three stories are not completely parallel—a chairman's not caring about the environment may not be judged by everyone to be equally morally reprehensible (if at all) as a mayor's not caring about the wildlife around his city or a president's not caring about the irrigation systems in neighboring countries—we should certainly allow for the possibility of a small effect. Allowing for an effect of $d = .2$, and requiring a significance level of .05, a power calculation showed that 200 participants were needed for a two-sided t-test to reach a power of 80 percent. The power calculation was done using the package `pwr` (Champely [2020]) for the statistical programming language R (R Core Team [2021]).

Most importantly, it was planned to perform a series of regression analyses, with a

[8] These are the screens relevant to the experiment proper; the introductory screen and screens asking for demographic information etc. are not shown.

The mayor of a small city has to decide whether or not to build a bridge over a nearby river. One of his advisors tells him that doing so will improve the flow of traffic in the city. The mayor responds: "I want to improve the flow of traffic. So, let's build the bridge." The bridge was built. This had a negative effect on the wildlife around the city.

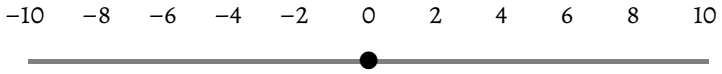
In your opinion, how blameworthy or praiseworthy is the mayor for acting as he did? Indicate your verdict by means of the slider below, where −10 indicates maximal blameworthiness and 10 indicates maximal praiseworthiness.



(next screen)

The minister of infrastructure of an African country advises the president of the country to construct a big dam to improve the country's irrigation systems. The minister tells the president that the dam will adversely impact the irrigation systems in the neighboring countries. The president responds that he doesn't care about the other countries, and orders that the dam be built. The dam is built, and the irrigation systems in the neighboring countries suffer as a result.

In your opinion, how blameworthy or praiseworthy is the president for acting as he did? Indicate your verdict by means of the slider below, where −10 indicates maximal blameworthiness and 10 indicates maximal praiseworthiness.



(next screen)

The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits." The vice-president mentioned to the chairman that the effects of the program on the environment are unknown. The chairman answered, "I don't care at all about the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program.

In your opinion, how blameworthy or praiseworthy is the chairman for acting as he did? Indicate your verdict by means of the slider below, where −10 indicates maximal blameworthiness and 10 indicates maximal praiseworthiness.

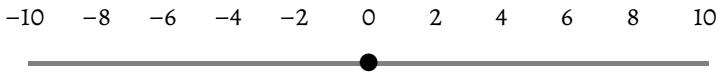


Figure 1: Example of the consecutive screens that a participant might have seen, the first screen presenting the O mode of the harm version of the mayor story, the second presenting the OA mode of the harm version of the president story, and the third presenting the A mode of the chairman story.

special interest in comparing models that included only one of the O and A variables as predictors, that included both, and that included the sum of O and A verdicts as a predictor. MBH predicts that the last model is as accurate as the model with both O and A as predictors,

and more accurate than the models that include only one of those variables as predictor.

4.2 Results

Figure 2 gives an overview of the data we obtained, separately for the different stories and, within those, for the different mode of presentation × harm/help conditions. The horizontal lines indicate the medians of the distributions, the orange dots, the means. Just eye-balling these plots, we expected both harm/help and mode of presentation to have made a significant impact on participants' responses. This impression was confirmed by conducting a 2 × 3 × 3 ANOVA, with harm/help as between-participants variable and story (chairman/mayor/president) and mode of presentation (OA/O/A) as within-participants variables. For the purposes of the ANOVA, we grouped the responses to the vignettes in the A mode of presentation—which, as said, were neutral with regard to the harm/help distinction—with the harm condition if the vignette had been presented with two other vignettes that belonged to the harm condition and with the help condition if the vignette had been presented together with vignettes in the help condition.

In line with the visual impression from Figure 2, and also as expected on the basis of previous experimental work by Knobe and others, there was a main effect of harm/help, $F(1, 747) = 67.23$, $p < .0001$, $\eta_p^2 = 0.08$, the mean for harm being −2.55 (SD = 6.72) and that for help being 1.06 (SD = 6.07). There was also a main effect of mode of presentation, $F(2, 747) = 45.56$, $p < .0001$, $\eta_p^2 = 0.11$, the means for OA, O, and A being −0.89 (SD = 6.77), 1.87 (SD = 6.21), and −3.25 (SD = 5.96). Note that the mean for the A mode is lowest. That is easily understandable, given that the other two means average over the harm and help conditions, whereas in the A mode there is always only an expression of a negative attitude. A main effect of story was significant at the $\alpha = .05$ level, $F(2, 747) = 3.02$, $p = .049$, $\eta_p^2 = 0.01$, indicating that the stories were not exactly parallel. However, the effect was minor, as

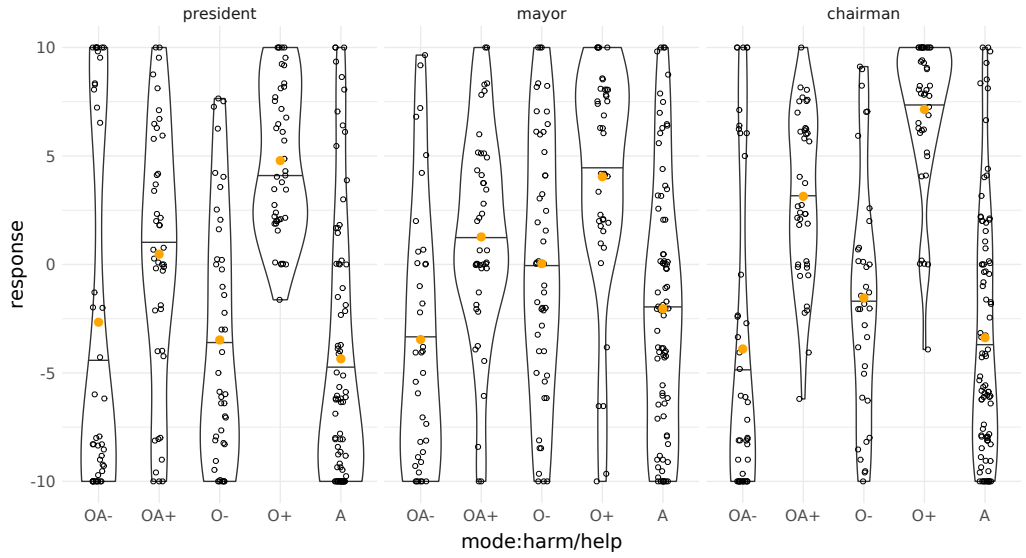


Figure 2: Violin plots of the data, per story, and for each story further split out per mode of presentation × harm/help condition (– indicates the harm condition, + indicates the help condition). The horizontal line within each distribution shows the median, the orange dot shows the mean.

Table 1: Descriptive statistics for the three modes of presentation of the three stories, specified separately for the harm and help conditions (the A mode is neutral as regards the harm/help distinction).

	harm						help		
	n	M	SD	*n*	M	SD	*n*	M	SD
OA	126	−3.34	7.34				125	1.58	5.08
O	126	−1.58	5.87				125	5.34	4.32
A				251	−3.25	5.96			

indicated by the η_p^2-value, and more importantly, in Tukey's HSD follow-up tests none of the pairwise comparisons reached significance.

A quick inspection of the averages for the modes of presentation as considered per condition gives a first indication that MBH is on the right track. MBH predicts that the average values for OA in both conditions should be close to the corresponding sums of O and A. Table 1 gives these averages together with standard deviations and number of ratings the statistics are based on. Two t-tests from these summary statistics showed that the mean of OA in the harm condition is not significantly different from the sum of the means of O in the harm condition and A, and similarly for the help condition: the mean of OA in the help condition is not significantly different from the sum of the means of O in that condition and A.

However, the findings mentioned thus far are still not very revealing with respect to MBH, as the averages are insufficiently informative about the individual verdicts: they are compatible with a general disparity between the participants' judgments for OA and the sum of their judgments for O and A; averaging can have the effect of canceling out such disparities. MBH is a claim about how individual people come to their moral judgments, so our analysis must address the individual level.

To that end, we added each participant's response to whichever vignette in the O mode he or she had been presented with to the participant's response to whichever vignette in the A mode he or she had received, and then compared those sums to participants' responses to whichever vignette in the OA mode they had received.

We first carried out a paired t-test on the two groups of values. Supposing the hypothesis at issue, we should have little hope of being able to reject the null hypothesis of there being no difference between the means of these groups, and indeed we were not able to do so, $t(250) = -1.10$, $p = .274$. Given the number of participants, the power of the test to reject the null hypothesis, on the supposition that it is false, and supposing that a small effect (Cohen's $d = .2$) would still be compatible with MBH, equals .88. The effect we found was actually close to 0: $d = .058$. Especially in view of the fact that the stories were not entirely symmetrical, MBH is perfectly compatible with such a small effect.[9]

[9]One might worry that the stories are still similar enough to each other to have given rise to carryover effects in our data. To address this concern, we used the randomization data from Qualtrics and compared ratings for the vignettes as they occurred at different places in the order of presentation. There were 15 vignettes in the study, for each of which we grouped ratings as they occurred at place i in the order, for $i = 1, 2, 3$, and then ran for these groups three t-tests (comparing first occurrences of a vignette with second occurrences, first occurrences with third occurrences, and second occurrences with third occurrences). So, we ran $3 \times 15 = 45$ t-tests in total. Of these, two came out significant at a level of $\alpha = .05$ and one of these was also significant at a level of $\alpha = .01$. These results are not significantly different from what one would expect to find by chance

It is to be emphasized, however, that the result from the t-test is not to be interpreted as *evidence* for the null hypothesis. In fact, it is commonly held that t-tests never allow one to state evidence for the null hypothesis. The notion to be used in this context is rather that of corroboration. It is therefore interesting to note that, in the increasingly popular Bayesian approach to statistics, researchers have developed an alternative to the (frequentist) t-test that does allow one to state evidence for the null hypothesis, in terms of a Bayes factor (see, e.g., Rouder et al. [2009]). Where E is one's current evidence, H_0 is the null hypothesis, and H_1 the alternative hypothesis (i.e., the hypothesis that the null is false), then $\Pr(\mathcal{E} \mid H_1)/ \Pr(\mathcal{E} \mid H_0)$, often abbreviated as BF_{10}, is the Bayes factor for the alternative, which expresses how much the evidence favors that hypothesis over the null; conversely, BF_{01} expresses how much the evidence favors the null over the alternative. Using the package BayesFactor for R (Morey and Rouder [2018]), we performed a Bayesian t-test on the OA responses and the sums of the O and A responses and obtained a Bayes factor of $\mathrm{BF}_{01} = 7.8$, thus indicating that the data are 7.8 times more likely assuming the null hypothesis than assuming the alternative hypothesis that postulates the presence of a difference. Following Jeffreys' [1961:432] classification scheme, this means that there is *substantial* support for the null hypothesis that there is no difference between the OA responses and the sums of the O and A responses.

Because it would still be more informative to know how well participants' OA responses were predicted by their O and A responses we also fitted a number of linear mixed-effects models to the data, using the lme4 package for R (see Bates et al. [2015]).[10] The models we compared all had OA response as response variable. One model had the sum of O and A responses as fixed effect (the O + A model), one had O response and A response as separate fixed effects (the O & A model), one had only O as fixed effect (the O model), and one, finally, had only A as fixed effect (the A model). All models had a full random effects structure for participants (as recommended in Barr et al. [2013]), meaning that they included random intercepts per participant and also a random slope per participant for each fixed effect they contained. All fixed effects were scaled to facilitate interpretation of the regression outcomes. Table 2 gives these outcomes for the various models and Table 3 presents several model comparison statistics.[11]

In Table 2, we see that, in the O + A model, as the sum of O and A goes up by one standard deviation, the OA response (i.e., the full-story judgment) goes up by 0.63 standard deviation. Similarly, in the O & A model, for every standard deviation that O goes up, keeping A fixed at its mean, the OA response goes up by 0.4 standard deviation, while for every standard deviation that A goes up, keeping O fixed at its mean, the OA response goes up by 0.37 standard deviation. All these relationships hold reliably, at least at the level of $\alpha = .01$.

In Table 3, k is the number of parameters and LL the log-likelihood. AIC is the Akaike

if there were no carryover effects and so order of presentation did not matter, as confirmed by one-sample proportion tests.

[10] Mixed-effects models allow us to focus more clearly on the variability in the data that is due to predictors of interest by filtering out the variability that is due to the fact that the experimenters happened to recruit this rather than that group of participants. See, for instance, Singmann and Kellen [2020] on the advantages of using mixed-effects models over standard regression models.

[11] An anonymous referee remarked that the models with only one fixed effect (so the O and A models) do not correspond to any position actually maintained in the debate about moral responsibility. While that is true, these models are still of interest, given that if one of them had come out on top, that would have cast doubt on MBH. Moreover, it would also have been reason to start thinking about a philosophical account that could have explained the superiority of that model with a single predictor.

Table 2: Results from our four linear mixed-effects models.

Model	Predictor	Estimate	SE	df	*t* value	Pr
O + A	(Intercept)	0.02	0.11	4.98	0.19	.86
	O + A	0.63	0.09	4.80	7.26	<. 001
O & A	(Intercept)	−0.02	0.11	4.84	0.17	.87
	O	0.40	0.07	4.60	5.68	<.005
	A	0.37	0.09	4.93	3.93	.01
O	(Intercept)	0.04	0.09	5.00	0.51	.63
	O	0.60	0.10	4.71	5.96	<.005
A	(Intercept)	−0.03	0.19	5.00	0.14	.90
	A	0.47	0.12	4.92	3.87	.01

Information Criterion and BIC the Bayesian Information Criterion. Both AIC and BIC weigh model fit and model complexity against each other (see Burnham and Anderson [2002]). ΔAIC and ΔBIC are the values for each model minus the smallest AIC or BIC value. It is to be noted that AIC and BIC values are not interpretable per se but only comparatively, in that models with smaller values are better descriptions of the data. R^2 is the squared correlation between the fitted and observed values.[12] We see that, across all model comparison criteria, both the O + A and the O & A model do much better than both the O and the A model: according to Burnham and Anderson [2002:70 f], for relatively large data sets (like ours), differences in AIC value greater than 10 indicate that the models with the higher values enjoy basically no empirical support. On the other hand, small differences in AIC or BIC value, like those between the O + A and O & A models, have little to no empirical meaning. In addition to this, the O + A and O & A models are exceedingly close in terms of R^2 values. Figures 3 and 4 plot the best models.

That the O + A model and the O & A model do about equally well strongly suggest that the way the O and A responses predict the response variable is in accordance with MBH, namely, by adding up. Indeed, we see that, in the O & A model, the two fixed effects contribute about equally to the predictions. Realistically speaking, we should in fact expect that people do not give *exactly* the same weight to the two summands in MBH, which is to be regarded as an idealization. That the O & A model allows the two components to be weighted differently explains why it is able to do slightly better than the O + A model in terms of R^2 value. (That the former model does slightly worse than the latter in terms of AIC and BIC values is because

Table 3: Comparison of the four regression models.

	k	LL	AIC	ΔAIC	BIC	ΔBIC	R^2
O + A model	4	−747.39	1502.78	0.00	1516.88	0.00	.52
O & A model	5	−747.38	1504.77	1.99	1522.40	5.52	.53
O model	4	−777.69	1563.38	60.60	1577.49	60.61	.42
A model	4	−771.82	1551.64	48.86	1565.74	48.86	.49

[12] R^2 values were calculated using the MuMIn package for R, which relies on the method described in Nakagawa, Johnson, and Schielzeth [2017].

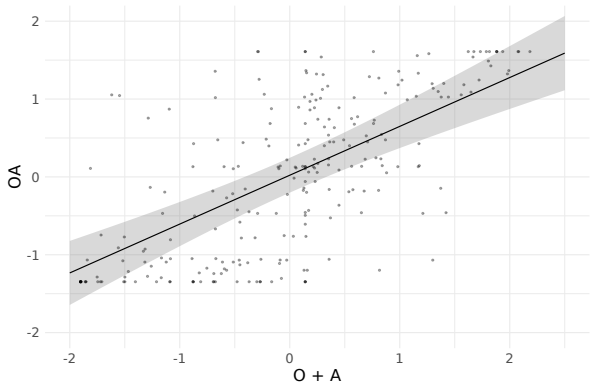


Figure 3: Plot of the O + A model; the shaded band indicates a 95 percent confidence region.

these also take into account model complexity, next to model fit.)

Finally, we would like to draw attention to a finding concerning outcome responsibility—the responsibility attributed in the O mode—that is directly relevant to MRA. Table 1 shows that, averaged over the three scenarios, the mean for the O mode in the help condition equals 5.34 (SD = 4.32) while in the harm condition it equals −1.58 (SD = 5.87). These means are not only significantly different from one another (two-sample $t(249) = 10.62$, $p < .0001$, Cohen's $d = 1.34$); each of them is also significantly different from 0. For the help condition, the mean is significantly greater than 0 (one-sample $t(124) = 13.81$, $p < .0001$, one-sided, Cohen's $d = 1.24$), and for the harm condition the mean is significantly smaller than 0 (one-sample $t(125) = -3.01$, $p = .0016$, one-sided, Cohen's $d = 0.27$). This is striking, inasmuch as the personae in the O mode of the scenarios express the same positive attitude toward the same intended effect and express no attitude toward the side-effect, about which they are not informed, and happens to be positive in the help condition and negative in the harm condition. In spite of this, participants apparently attribute what may be called "a praise bonus" in the help condition and "a blame penalty" in the harm condition. We come back to this finding.

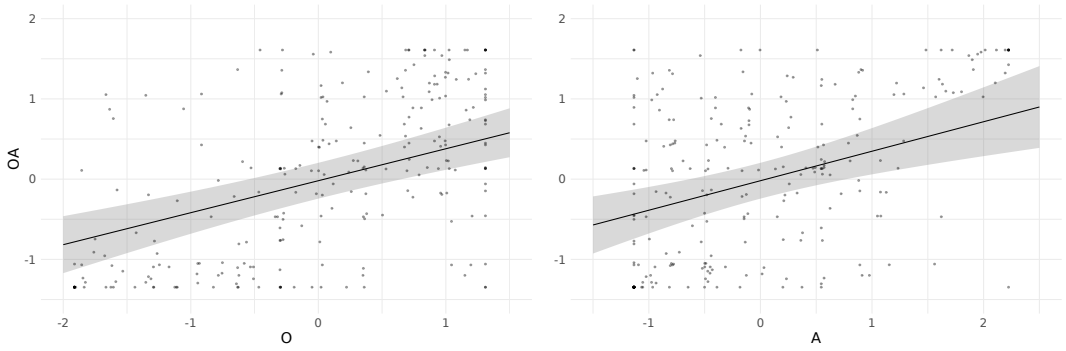


Figure 4: Marginal effects plots of the O & A model; the shaded bands indicate 95 percent confidence regions.

4.3 Discussion

MBH predicts that if people are asked to evaluate the component parts of an action, then by totaling those separate evaluations one will approximately arrive at what one would have obtained had one asked the same people for their overall evaluation of the problem case. The test we reported showed that, indeed, there was no significant difference between people's overall moral verdict on a case and the sum of their moral verdicts on the separately evaluable components of that case. Often, in empirical research one aims to demonstrate that there *is* a significant difference between two values. In such cases, it is of the utmost importance that one's test is "strict" enough and does not proclaim to have detected a significant difference when in reality there is none; the probability of committing a Type I error should be low. In contrast, in the kind of situation at hand, where the claim is that there is no significant difference between two values (the overall verdict and the sum of the verdicts on the parts), one's test must be "strong" enough so as not fail to detect a significant difference when in reality there is one; the probability of committing a Type II error should be low. Hence our concern with power in the above: we showed that the test we conducted had sufficient power to detect a significant difference between total moral verdict and sum of the verdicts on the parts. Therefore, that no such difference was found to obtain corroborates MBH. Also, we supplemented the frequentist t-test we conducted with a Bayesian one, for which power considerations are of no (Rouder [2014]), or in any case of much lesser (Kruschke [2015, Ch. 13]), concern. Moreover, the outcome of the Bayesian t-test allowed us to directly assert that the data from our experiment provide evidence in favor of the null hypothesis implied by MBH, even substantive evidence.

Furthermore, a series of regression analyses showed that by adding up verdicts on the component parts of the stories—thus turning them into a single variable—one can predict the overall verdict as accurately as when one bases predictions on the component parts. This strongly suggests that the verdicts on the component parts allow one to predict the overall verdict precisely by adding up.

As regards the goodness of fit of the various models that we looked at in our regression analyses, it is to be noted that, by conventional criteria, the fit of even the best models is only moderately good. To a large extent, this may be due to the fact that the stories we used were not fully symmetric, which was already clear from comparing the relevant violin plots in Figure 2 and was confirmed by our analysis. This finding was not completely surprising. It is well imaginable, for instance, that people regard it as morally unproblematic for a president of a country to promote the interests of his country at the expense of those of others. Alternatively, people might assume that the responsibilities that public officials have regarding externalities differ from those of businessmen: whereas the former should take unintended effects into account when they make decisions irrespective of whether they are positive or negative, businessmen only have reason to attend to negative side effects. It is well worth trying to come up with stories that match each other still more closely than the ones that were employed here and repeat the experiment with those.

We also found in our data an asymmetry in moral verdicts in the absence of intentionality, apparently resulting only from differences in outcome. People attribute more praise than blame in such situations. We call this "the outcome-responsibility asymmetry" and return to it below.

5 General discussion

The key result of this paper is that MBH is supported by empirical findings concerning vignettes that have the structure of Knobe's chairman scenario. In the help condition, the total praise attributed for the protagonist's action is not significantly different from the sum of the praise attributed for the beneficial outcome (outcome praise) and the blame attributed for the protagonist's indifference with respect to this outcome (attitude blame). In the harm condition, a roughly similar amount of attitude blame added to the attributed outcome blame is not significantly different from the total amount of blame that participants attribute.

These findings are radically different from what was to be expected, given the way in which moral philosophers believe that outcome responsibility should be attributed. According to MRA, responsibility is to be attributed for beneficial outcomes only when they are intended. No praise is to be ascribed to outcomes in the absence of this kind of attitude. We find, however, that people attribute praise on the basis of beneficial outcomes only (a finding we referred to as "the praise bonus"). In light of this, we propose that the praise bonus along with the blame penalty and the outcome-responsibility asymmetry are systematic biases.[13] And we proceed to explain why the attributions that support MBH are not in line with MRA.[14]

5.1 The praise bonus and the blame penalty

Our findings concerning attitude responsibility are as was to be expected. Participants attribute blame to an agent who expresses indifference. Such blame is readily intelligible given the lack of moral concern that the agent displays.[15] In contrast, the findings concerning outcome responsibility present us with a deep puzzle: Why do people attribute praise and blame when both intent and foresight are absent? The outcome-responsibility asymmetry gives rise to the question why the blame bonus is substantially larger than the praise bonus (in absolute terms). We argue that people seem to attribute praise and blame in a way that is, at times, entirely pointless.

Existing hypotheses in psychology are of little help in this respect. Interestingly, earlier findings suggest that people attribute more blame than praise. And in situations that feature both an attitude and an outcome we find the same (Table 1). In light of such findings, Guglielmo and Malle [2019] have formulated the Amplified Blame Hypothesis, according to which people blame others more than they praise them in situations that differ only in terms of valence. However, this hypothesis has not yet been considered in situations where the outcome is the only relevant variable (O). And in such situations, we find a praise bonus that is significantly larger than the (absolute value of the) blame penalty (Table 1). Thus, the outcome-responsibility asymmetry provides evidence against the Amplified Blame Hypothesis.

Anderson, Crockett, and Pizzaro discuss other asymmetries between praise and blame,

[13] See Kahnemann [2011] for a discussion of a wide range of cognitive and affective biases.

[14] An explanation that implies that a finding reflects a bias to which people are susceptible, at least in part, is a *bias explanation*. In contrast, a *competence explanation* accounts for the findings such that they reflect competent or reliable judgments (Nado [2008], Knobe [2010], Hindriks [2014]).

[15] A related ground for blaming the agent for his indifference is the fact that he violates the Side-Effect Deliberation Effect mentioned in note 3.

in particular that causality, intentionality and (intended) outcomes have a substantial effect on blame attributions and hardly any on praise attributions. They argue that, when attributing blame, people are primarily concerned with "how the action was performed," whereas praise attributions are based first and foremost on "what kind of person performed the action" (Anderson, Crockett, and Pizzaro [2020:701]). In particular, those who attribute praise are concerned with how trustworthy the other is and whether she "can be counted on to be a cooperative partner in the future" (Anderson, Crockett, and Pizzaro [2020:700]). Subsequently, praise acts to strengthen the relationship between the attributor and the recipient. However, in the vignettes without indifference, attributors have little if any information about the agent's character. So, it is unlikely that Anderson, Crockett, and Pizzaro's hypothesis explains our finding.

The same authors also mention that praise might serve "to reinforce that moral behavior in the recipient" (Anderson, Crockett, and Pizzaro [2020:700]). Inspired by this, one could consider a consequentialist explanation, which complements this claim about praise with one about blame: that it serves to discourage bad behavior. However, the praise bonus and the blame penalty are observed in scenarios in which the agent is left in the dark about any side effects his decision might have. This means that, as side effects played no role in the agent's deliberations, praise and blame cannot serve to reinforce or discourage motivations that informed his decision. The upshot is that, if people try to encourage agents to bring about beneficial outcomes and discourage them to bring about harmful consequences, they do so in a woefully inadequate manner.[16]

As the findings are inconsistent with MRA, they seem to reflect biases. The discussion thus far reveals that it is difficult to make sense of them. In fact, the outcome-responsibility attributions appear to be pointless. This does not mean that they defy explanation. We propose to explain them in terms of affect. The idea is that beneficial and harmful outcomes trigger positive and negative affect respectively. Furthermore, people are inclined to attribute praise when they are in a good mood, and blame when they are in a bad mood (praise bonus and blame penalty).[17] Finally, this effect is bigger for beneficial outcomes than for harmful outcomes (outcome-responsibility asymmetry). On this affect-based explanation, these attributions are undeserved, which means that both the praise bonus and the blame penalty are indeed systematic biases.

5.2 Concluding remarks

Moral compositionality is the idea that people split up the problem of attributing responsibility by ascribing responsibility to component parts and subsequently combining these component attributions to form a judgment about total responsibility. According to the Moral Bookkeeping Hypothesis (MBH), total responsibility simply is the sum of the component attributions. We have tested this by distinguishing two components of actions, to wit outcomes and attitudes. Our empirical findings support MBH. Of course, our conclusion remains somewhat tentative. For instance, the present study cannot rule out that there are

[16] It could be that the attribution of praise as well as blame is in part motivated by relation-based concerns. And it may well be that people can make friends by attributing undeserved praise. But that certainly does not hold for attributing undeserved blame. Furthermore, people attribute more praise in the absence of foresight than in its presence (Table 1). And this seems like poor relationship-management.

[17] Similar mood effects have been found in relation to helping behavior, which can be triggered by relatively arbitrary factors such as finding a dime or smelling freshly baked bread (Isen and Levin [1972], Gueguen [2012]).

weight factors involved in moral bookkeeping. Still, compared to the *status quaestionis*, our hypothesis and the presented evidence that is consistent with the simplest operationalization offer a promising new avenue for thinking about responsibility.

Moral compositionality conflicts with moral theory and common sense about responsibility in that neither allows for ascriptions of praise or blame on the basis of an outcome only. Yet, we find exactly this: people attribute praise and blame to agents who bring about a beneficial or harmful outcome without foreseeing that outcome and without even having been informed about the possibility that the envisaged plan might have a side effect. We have not been able to formulate a plausible competence explanation of these puzzling findings. Although a lot is to be said in favor of a bias explanation, more research is needed before this conclusion can be drawn, in particular to determine how robust these findings are and what mechanism or mechanisms give rise to them.

It is to be emphasized that we have only offered one piece of evidence in favor of MBH, crucially building on Knobe's seminal work. However, MBH is a broad thesis about moral responsibility, which is supposed to apply regardless of considerations having to do with intentionality or side effects.[18] Thus, an obvious avenue for future research is to conduct additional experiments using a richer, more varied set of materials. For instance, one possible follow-up experiment would be to test scenarios concerning intended effects and contrast those with scenarios involving effects that are not due to an intention at all. In this way, carryover effects are avoided that might arise, for instance, when intended effects are contrasted with effects that are due to deviant causal chains (as in Malle [2006]). Another way to make progress is to use more fine-grained decompositions of actions, which may help to put MBH to more severe tests. All in all, the findings presented in this paper are to be considered as just a first step in uncovering the structure of folk responsibility attributions and their relations to moral competence concerning responsibility ascriptions.[19]

References

Adams, F. and Steadman, A. [2004] "Intentional Action in Ordinary Language: Core Concept or Pragmatic Understanding?" *Analysis* 64:173–181.

Anderson, R. A., Crockett, M. J., and Pizarro, D. A. [2020] "A Theory of Moral Praise," *Trends in Cognitive Sciences* 24:694–703.

Aron, A., Aron, E. N., and Coups, E. J. [2009] *Statistics for Psychologists* (5th ed.), Upper Sadle River NJ: Pearson Education.

Arpaly, A. [2002] "Moral Worth," *Journal of Philosophy* 99:223–245.

Ashworth, A. [2006] *Principles of Criminal Law* (5th ed.), Oxford: Oxford University Press.

Aust, F., Diedenhofen, B., Ullrich, S., and Musch, J. [2013] "Seriousness Checks are Useful to Improve Data Validity in Online Research," *Behavior Research Methods* 45:527–535.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. [2013] "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal," *Journal of Memory and Language* 68:255–278.

Bates, D., Mächler, M., Bolker, B., and Walker, S. [2015] "Fitting Linear Mixed-effects Models Using lme4," *Journal of Statistical Software* 67::1–48

[18] Thanks to an anonymous referee for pressing us on this.

[19] We are greatly indebted to two anonymous referees for valuable comments on a previous version.

Bratman, M. [1987] *Intention, Plans, and Practical Reason*, Cambridge MA: Harvard University Press.

Burnham, K. P. and Anderson, D. R. [2002] *Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach*, Berlin: Springer.

Champely, S. [2020] pwr*: Basic Functions for Power Analysis*, R package version 1.3-0, http://CRAN.R-project.org/package=pwr.

Cohen, J. [1988] *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale NJ: Lawrence Earlbaum Associates.

Graham, P. A. [2014] "A Sketch of a Theory of Moral Blameworthiness," *Philosophy and Phenomenological Research* 88:388–409.

Gueguen, N. [2012] "The Sweet Smell of . . . Implicit Helping: Effects of Pleasant Ambient Fragrance on Spontaneous Help in Shopping Malls," *Journal of Social Psychology* 152:397–400.

Guglielmo, S. and Malle, B. F. [2019] "Asymmetric Morality: Blame Is More Differentiated and More Extreme than Praise," *PLOS One* 14: e0213544.

Hindriks, F. [2008] "Intentional Action and the Praise–Blame Asymmetry," *Philosophical Quarterly* 58:630–641.

Hindriks, F. [2014] "Normativity in Action: How to Explain the Knobe Effect and its Relatives," *Mind and Language* 29:1–22.

Hindriks, F., Douven, I., and Singmann, H. [2016] "A New Angle on the Knobe Effect: Intentionality Correlates with Blame, not with Praise," *Mind and Language* 31:204–220.

Isen, A. M. and Levin, P. F. [1972] "Effect of Feeling Good on Helping: Cookies and Kindness," *Journal of Personality and Social Psychology* 21:384–388.

Jeffreys, H. [1961] *Theory of Probability* (3rd ed.), Oxford: Oxford University Press.

Kahneman, D. [2011] *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux.

Kenkel, B. and Signorino, C. S. [2014] "Estimating Extensive Form Games in R," *Journal of Statistical Software* 56:1–27, http://www.jstatsoft.org/v56/i08/.

Knobe, J. [2003] "Intentional Action and Side-effects in Ordinary Language," *Analysis* 64:81–87.

Knobe, J. [2006] "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology," *Philosophical Studies* 130:203–231.

Knobe, J. [2010] "Person as Scientist, Person as Moralist," *Behavioral and Brain Sciences* 33:315–329.

Kruschke, J. K. [2015] *Doing Bayesian Data Analysis* (2nd ed.), London: Academic Press.

Malle, B. F. [2006] "Intentionality, Morality, and their Relationship in Human Judgment," *Journal of Cognition and Culture* 6:87–112.

Malle, B. F., Guglielmo, S., and Monroe, A. E. [2014] "A Theory of Blame," *Psychological Inquiry* 45:147–186.

Morey, R. D. and Rouder, J. N. [2018] BayesFactor*: Computation of Bayes Factors for Common Designs*, R package version 0.9.12-4.2, http://CRAN.R-project.org/package=BayesFactor.

Nado, J. [2008] "Effects of Moral Cognition on Judgments of Intentionality," *British Journal for the Philosophy of Science* 59:709–731.

Nakagawa, S., Johnson, P. C. D., and Schielzeth, H. [2017] "The Coefficient of Determination R^2 and Intra-class Correlation Coefficient from Generalized Linear Mixed-effects Models Revisited and Expanded," *Journal of the Royal Society Interface* 14:20170213.

Pellizzoni, S., Girotto, V., and Surian, L. [2010] "Beliefs and Moral Valence Affect Intentionality Attributions: The Case of Side Effects," *Review of Philosophy and Psychology* 1:201–209.

R Core Team [2021] *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, `http://www.R-project.org/`.

Rouder, J. N. [2014] "Optional Stopping: No Problems," *Psychonomic Bulletin & Review* 21:301–308.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. [2009] "Bayesian *t* Tests for Accepting and Rejecting the Null Hypothesis," *Psychonomic Bulletin and Review* 16:225–237.

Scanlon, T. [1998] *What We Owe To Each Other*, Cambridge MA: Harvard University Press.

Sher, G. [2009] *Who Knew? Responsibility Without Awareness*, Oxford: Oxford University Press.

Singmann, H. and Kellen, D. [2020] "An Introduction to Mixed Models for Experimental Psychology," in D. Spiler and E. Schumacher (eds.) *New Methods in Cognitive Psychology*, New York NY: Routledge, pp. 4–31.

Smith, A. M. [2008] "Control, Responsibility, and Moral Assessment," *Philosophical Studies* 138:367–392.

Stocker, M. [1973] "Act and Agent Evaluations," *Review of Metaphysics* 27:42–61.

Strawson, P. F. [1962] "Freedom and Resentment," *Proceedings of the British Academy* 48:1–25.

Wagner, V. [2014] "Explaining the Knobe Effect," in C. Luetge, H. Rusch, and M. Uhl (eds.) *Experimental Ethics: Toward an Empirical Moral Philosophy*, Basingstoke UK: Palgrave Macmillan, pp. 65–79.

Wolf, S. [1990] *Freedom Within Reason*, Oxford: Oxford University Press.