Title

Outcomes Speak Louder than Actions?
Testing a Challenge to the Two-Process Model of Moral Judgment

Permalink

https://escholarship.org/uc/item/8mq6491t

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 41(0)

Authors

Prochownik, Karolina
Cushman, Fiery A.

Publication Date

2019

Peer reviewed

Outcomes Speak Louder than Actions?
Testing a Challenge to the Two-Process Model of Moral Judgment

Karolina Prochownik (karolina.prochownik@rub.de)
Faculty of Law, Ruhr University Bochum, Universitätsstrasse 150, 44801 Bochum, Germany
Faculty of Law and Administration, Jagiellonian University, Krakow, Poland

Fiery A. Cushman (cushman@fas.harvard.edu)
Department of Psychology, Harvard University, William James Hall 1480, 33 Kirkland St.
Cambridge, MA 02138, USA

Abstract

Curiously, people assign less punishment to a person who attempts and fails to harm somebody if their intended victim happens to suffer the harm for coincidental reasons. This "blame blocking" effect provides an important evidence in support of the two-process model of moral judgment (Cushman, 2008). Yet, recent proposals suggest that it might be due to an unintended interpretation of the dependent measure in cases of coincidental harm (Prochownik, 2017; also Malle, Guglielmo, & Monroe, 2014). If so, this would deprive the two-process model of an important source of empirical support. We report and discuss results that speak against this alternative account.

Keywords: blame blocking; two-process model; punishment; outcomes; actions; pragmatics

Introduction

Imagine that two runners compete in a championship race. One of the runners is a frequent winner, and so another racer decides to kill him and exclude him from the competition. He mistakenly believes that his rival is fatally allergic to poppy seeds, and so he sprinkles some on his rival's food at the banquet. The champion is not allergic to poppy seeds at all, however, but instead to hazelnuts. What's more, completely by coincidence, the chef happens to have served a hazelnut salad, and the champion dies as a result of consuming it.

The "blame blocking" phenomenon, first reported in the set of studies by Cushman (2008), is that people will tend to reduce blame and punishment assigned to the attempted harmdoer because of the coincidental harm caused by the salad. In other words, if the salad has no hazelnuts and the intended victim survives, the attempted harmdoer is blamed and punished more. This effect is notably large. Specifically, in the study based on the above story Cushman (2008) found that about half as many subjects assigned *no punishment* to the runner where no harm occurred compared with the case in which the rival coincidentally died (p. 374). That is, the coincidental death of the rival made participants twice as likely to let the runner off the hook.

One explanation of this puzzling effect posits two processes of moral judgment that render moral judgments separately on the basis of (1) causal responsibility for harm, or (2) a culpable mental state, such as intent to harm (Cushman, 2008). According to the model, then, when there was no causal input in the story (i.e., no coincidental harm occurs), the "mental state process" dominates and punishment judgments are therefore based on the evaluation of the agent's mental states alone. Because the relevant mental state was severe intentional harm, this tends to result in non-zero levels of punishment. On the other hand, when causal inputs are present (i.e., a coincidental harm occurs) but the runner himself is non-causal, the process of moral judgment predicated on causal responsibility competitively dominates (or "blocks") the evaluation of his mental states. The causal responsibility process assigns no punishment to the runner (who, of course, has no causal responsibility for the harm). Stated more generally, a two-process model of moral judgment can accommodate the pattern of results because it posits competition between a causal process seeking full exculpation (no punishment) and a mental state process seeking full inculpation (punishment) in cases of failed attempts to harm with independently caused harm, while the relative influence of the causal process is minimized in cases of pure failed attempts.

The two-process model is compatible with theories of moral judgments that identify intentional and causal evaluations as primary contributors to blame and punishment (e.g., Alicke, 2000; Alicke & Rose, 2012; Carlsmith & Darley, 2008; Darley & Shultz, 1990; Fincham & Jaspers, 1979; Shultz, Schleifer, & Altman, 1981; Shultz, Wright, & Schleifer, 1986; Weiner, 1995; Guglielmo, Monroe, & Malle, 2009; Malle, Guglielmo, & Monroe, 2014; Piaget, 1932/1965). It departs from most of these theories, however, in the assumption that causal and mental state evaluations proceed separately and compete during moral judgments of blame and punishment, rather than being combined and integrated in a single process.

In addition to the blame blocking phenomenon, some independent evidence provides support for the two-process model. Young, Cushman, Hauser, and Saxe (2007) found neurological signature of conflict for adult judgments of accidental harms in which intentional and causal evaluations point in different directions. Several studies show that punishment judgments are especially strongly influenced by the causal process in ordinary cases of harm (Martin & Cushman, 2015, 2016), and developmental evidence suggests that this pattern is a vestige of an early-emerging "causal" process of moral judgment augmented by a later-emerging "mental state" process (Cushman, Sheketoff, Wharton, & Carey, 2013).

Here, we consider another explanation of the blame blocking effect—one that depends on assumptions about how people interpret the pragmatics of the dependent measure used to trigger this effect. Specifically, the question "How much prison time does [agent] deserve?" used by Cushman (2008) might be interpreted by participants differently across conditions: as implicitly referring to punishment *for behavior* (how much should the runner be punished for trying to kill his rival with poppy seeds) in the "no harm" condition, but as implicitly referring to punishment *for a harmful outcome* (how much should the runner be punished for the victim's death by the hazelnuts) in the coincidental harm condition (see also Prochownik, 2017). If the agent in two scenarios were evaluated against these very different standards in each case it would explain the blame blocking effect without appeal to two processes of moral judgment. We call this alternative "pragmatics account" because it relies on an assumption that people take a broad context into account when deciding what for to punish others (cf. Prochownik, 2017).[1]

In this paper we examined this alternative hypothesis by conducting two experiments. In Experiment 1, we manipulated the question about punishment to ensure that it is interpreted with a wide scope, encompassing not only what the agent caused (or did not), but also what he intended. Next, in Experiment 2, we used the original dependent measure that was previously used to elicit the blame blocking effect, and then asked participants a series of questions designed to clarify how they understood it.

Collectively, the results of these experiments suggest that unintended interpretations of the dependent measure are not sufficient to explain the full blame blocking effect.

Experiment 1

The goal of Experiment 1 was to test whether a more precise phrasing of the dependent measure would eliminate the previously observed blame blocking effect. In the baseline condition ("unspecified") we left the question identical to previous experiments by Cushman (2008): "In your opinion, how much prison time does *X* deserve?" In the novel condition ("specified") we modified the question so that it more clearly pointed at the agent's total set of behaviors as the target of punishment, thus diminishing the chance that it would be interpreted in terms of outcome alone (following in this respect Prochownik, 2017): "Suppose that X were apprehended by the police and put on trial. Given the complete set of behaviors and facts, in your opinion how much prison time does he deserve?"

The language that we used in the "specified" condition was borrowed from earlier research. In particular, Prochownik (2017) found that people with legal education tended to manifest the blame blocking effect only when the punishment question was unspecified, but the effect disappeared when it was specified, suggesting a key role for pragmatics in this group of respondents. However, Prochownik & Unterhuber (2018) did not replicate this finding in their comparative study including both lay people and legal experts. In Experiment 1 we use the same version of the "specified punishment question" as these researchers, but we focus exclusively on lay people in a well-powered study, and also examine it more systematically (across sixteen scenario contexts instead of just two or three as in these previous studies).

Methods

We tested 20 participants in each of 64 cells of a 2 (harm vs. no harm) x 2 (specified vs. unspecified) x 16 (scenario context) design, for a total sample of 1280. Participants were recruited on MTurk in the US. After consenting to participate in a short study for small compensation ($0.30), they filled an online Qualtrics survey comprised of one scenario, a punishment probe, and demographic questions (age, gender, nationality, exposure to moral philosophy, religiosity, etc.). Participants marked their answers on a scale with 11 anchored options: "None", "1 week", "1 month", "3 months", "6 months", "1 year", "2 years", "4 years", "8 years", "16 years", "32 years".[2]

The total set of 16 scenarios varied along several dimensions. Most notably, half of them involved physical harm (burning, cutting, stabbing, etc.) while the remaining half involved property harm (arson, defacement, etc.). The full text of all study scenarios is available online as Supplementary Materials:
https://osf.io/9w4ke/.

Results

As summarized in Figure 1, we observed the basic blame blocking effect in both the "specified" and "unspecified" conditions. Indeed, if anything, the blame blocking effect was slightly larger in the new "specified" condition. In order to analyze the data more fully we conducted a linear mixed effect analysis. First, we constructed a null model without fixed effects for harm or punishment question type, but including a random effect for scenario. We then found that this model was significantly improved by modelling the harm factor, $\chi^2(3) = 61.49$, $p < .001$. Next, we found that this "harm only" model was not significantly improved by modelling the punishment question type factor $\chi^2(4) = 1.17$, $p = .8826$, or

[1] The importance of pragmatic considerations for participants´ (re)interpretations of research stimuli has been also raised by some recent studies (e.g., Guglielmo & Malle, 2010; Samland & Waldmann, 2016; Wiegmann, Samland, & Waldmann, 2016; Hagan & Rozyman, 2017).

[2] The "unspecified" punishment question was taken from Cushman (2008): Experiment 4. However, the scale differed from the 9-points scale used by Cushman in his study as for the majority of scenario contexts we did not use attempted murders but attempts of less severe crimes (including bodily injuries and damages to property) for which we needed a greater range of less severe sentences. As a result, we could also examine if the previous findings replicate when a different scale of punishment ratings is used.

by modelling both this factor and its interaction with harm χ^2 (9) = 4.1, p = .9047. In summary, then, the best-fitting model included only harm as a factor. In other words, we observe a significant effect for the harm vs. no harm factor, but no significant effect for the specified vs. unspecified factor, or for its interaction with harm.

We next assessed whether there are significant differences between scenarios in the magnitude of the blame blocking effect that they induce by testing whether random intercepts (i.e., an interaction between scenario context and the effect of the "harm" variable) contribute significantly to the model. They do, χ^2 (2) = 20.9, p < .001, indicating meaningful variability between vignettes. We next tested whether the model was improved by adding a fixed effect for "physical" versus "property" harms, but it was not χ^2 (9) = 7.94, p = .54. The precise nature of the relevant differences between scenarios therefore remains an important topic for further research.

Discussion

Experiment 1 shows that the blame blocking effect is not diminished by an alternative phrasing of the dependent measure designed to clarify that punishment could apply to any aspect of an attempted harmdoer's conduct—including, most importantly, the attempted harm.

These results speak against the alternative interpretation of that effect in terms of the pragmatic constraint on the way ordinary people assign punishment, and instead support the two-process model of moral judgment.

However, one limitation to this experiment is that by asking participants to consider the entire event when making their punishment judgments, we cannot completely exclude the possibility that some participants interpreted the question as referring to the outcome alone. If so, it is still possible that people who interpreted the question as referring to the outcome were driving the blame blocking effect. To address this problem we conducted an additional experiment which faithfully replicated the original "runners study" by Cushman (2008) but differed in one important element: participants in the "Harm" condition were presented with an additional question about how they understood the question about punishment (i.e., what they thought the punishment was meant to be for).

Experiment 2

In this experiment we replicated Cushman`s Experiment 4 (2008) but we presented participants in the "Harm" condition with an additional question about how they understood the punishment question after they have responded to it. Specifically, we asked explicitly whether they understood the question "how much punishment does *X* deserve?" to refer to "punishment *for the actual harm*" (e.g., death of a runner) in the coincidental harm condition. Such an interpretation, which is consistent with the pragmatics account, would explain away the purported "blame blocking effect."

We offered participants two alternatives to this interpretation of the question: First, that "punishment" referred only to the attempted harm (e.g., the sprinkling of poppy seeds on a salad with intent to cause an allergic reaction) and, second, that it referred to *both* the attempted

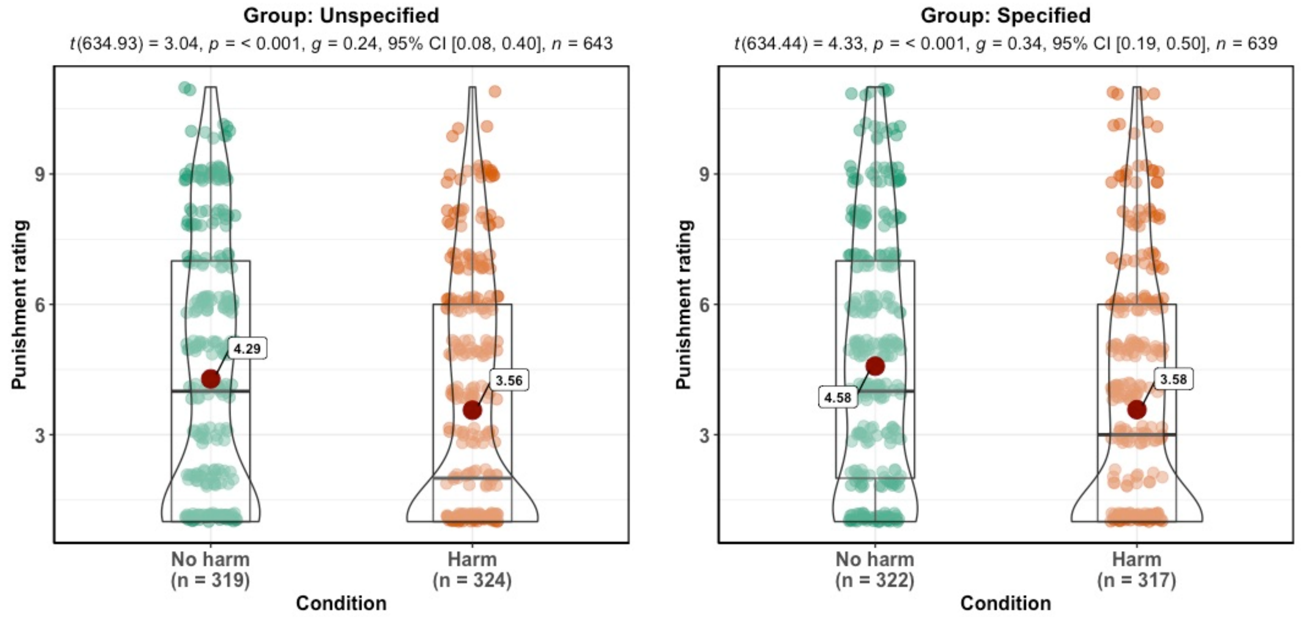

Figure 1: The blame blocking effect was replicated both in the "unspecified" condition, which directly matches the original demonstration (Cushman, 2008) and in the "specified" condition, which was designed to eliminate the alternative explanation of the effect in terms of pragmatics.

and actual harm. The blame blocking model is agnostic with respect to these alternatives—crucially, both of them entail sensitivity to the attempted harm, and thus the null prediction would be equal punishment across the no harm and coincidental harm conditions, both of which involve this attempted harm. The two-process model attempts to explain why participants who interpret the punishment question to *include* this key shared element—the attempted harm—would nevertheless be more likely to fully exonerate the attempted harmdoer in the coincidental harm case.

Study hypotheses, methods of analyses, sample size calculation and exclusion criteria were preregistered (the OSF preregistration document can be viewed at https://osf.io/pf574).

Methods

1007 complete responses were collected via TurkPrime in the US using Qualtrics anonymous link (we intended to recruit 500 participants per each of the study conditions). Participants were payed $0.50 for taking part in the survey.

Participants were asked to imagine that they are in a jury in a case of a defendant named Brown. In following, they were presented with a story of two runners named Brown and Smith competing in a championship race. One group of participants saw the variant of the story where Brown tries to kill Smith by sprinkling the poppy seeds on his food, but no harm results ("No Harm" condition). Another group of participants was presented with the story in which Smith dies because of the hazelnuts in the salad that he is served, completely independently of Brown´s actions ("Harm" condition). After reading the story all participants were asked: "How much prison time does Brown deserve?", and chose between the nine following options: "None", "6 months", "1 year", "2 years", "4 years", "8 years", "16 years", "32 years", "Life" (Cushman, 2008).

On the next page of the survey, participants in the "Harm" condition were presented with the following "Harm Understanding Question":

"On the last screen you were asked to decide how much prison time Brown deserved. Which of the following did you think was meant by that:

1. How much prison time for sprinkling poppy seeds on Smith's food?
2. How much prison time for the death of Smith?
3. How much prison time for both sprinkling poppy seeds on Smith´s food and the death of Smith?"[3]

Participants who chose the third option ("for both") were additionally asked two questions about punishment for the action alone and for the outcome alone to enable the researchers better understand their previous answers: "How much prison time does Brown deserve only for the death of Smith?" and "How much prison time does Brown deserve only for sprinkling poppy seeds on Smith´s food?". Answers to both questions were marked on the same 9-points scale as above.

Finally, all participants were asked two comprehension questions about the story they read: "Why did Brown sprinkle poppy seeds on Smith´s food?" (multi-choice question) and "Did Smith die as a result of Brown sprinkling poppy seeds on his salad?" (two-choice question).[4]

Participants were excluded from the analysis if they answered incorrectly to any of the two comprehension questions (i.e., if to the first question they provided any answer other than "Because he wanted to kill Smith" and/or if they answered "Yes" to the second question). This resulted in 840 responses included in the analysis (N_{Harm} = 420, N_{NoHarm} = 420).

Results

Percentages of different responses to the "Harm Understanding Question" (n = 420, 100%) were as follows**:** 56.4% (n = 237) respondents understood the punishment question as being for sprinkling poppy seeds on Smith´s food, 19.3% (n = 81) as being for the death of Smith, and 24.3% (n = 102) as being for both sprinkling poppy seeds on Smith´s food and the death of Smith.

Consistent with our preregistered plan, and following the key analysis by Cushman (2008), we recoded the responses to the main punishment question to a binary variable with the following values: "No punishment" (all "None" responses) and "Any punishment" (all the responses assigning some punishment from "6 months" to "Life"). Subsequently, to test the main hypothesis we performed two chi-square tests comparing the frequencies of "No punishment" vs. "Any punishment" responses across two study conditions "Harm" and "No Harm": (1) a chi-square test with the overall sample (analysis repeating Cushman, 2008, Experiment 4), and (2) a chi-square test excluding people in the "Harm" condition who in the "Harm Understanding Question" replied that they understood the punishment question as referring to the outcome alone (i.e., the death of Smith).

Overall, 35% people assigned "No punishment" in the "Harm" condition, while in the "No Harm" condition barely half as many (18%) of people did so.[5] The difference was statistically significant, χ^2 (1, N = 840) 31.740, p < .001. Critically, this result held even after excluding participants who indicated that they thought the punishment question referred to punishment for the outcome only: among the remaining participants, 28% assigned "No punishment" in the "Harm" condition comparing to 18% in the "No Harm" condition. The difference was statistically significant, χ^2 (1, n = 759) 11.763, p = .001 (Figure 2).

[3] This and two following questions were omitted in the "No Harm" condition as no death of Smith resulted in this story.

[4] In the first multi-choice comprehension question participants could choose from the following responses: "Because he thought the poppy seeds would make Smith sick for a couple of days", "Because he thought Smith liked poppy seeds"; "Because he wanted to kill Smith", "Because he wanted Smith to go to the bathroom". In the second question they could choose between two options: "Yes" or "No".

[5] Note that in Cushman (2008) the numbers were very similar, with 34.5% participants in the "Harm" and 19.5% in the "No Harm" case deciding not to punish Brown at all (p. 374).

In addition to the main analyses reported above, we also assessed whether the result is driven by remaining participants who understood the punishment question to refer to the attempted harm only, or by those who understood it to refer to both the attempted harm and the outcome. We found that the effect was maintained among those who understood the question to refer to both the attempted harm and the outcome: 43% people decided not to punish in this group comparing to 18% in the no harm group, χ^2 (1, n = 522) 29.801, p < .001. It was not significant, however, among those who understood the question to refer to the attempted harm only; among this group, 22% people assigned "No punishment" in the "Harm" condition, comparing to 18% in the "No Harm" condition, χ^2 (1, n = 657) 1.620, p = .203.[6]

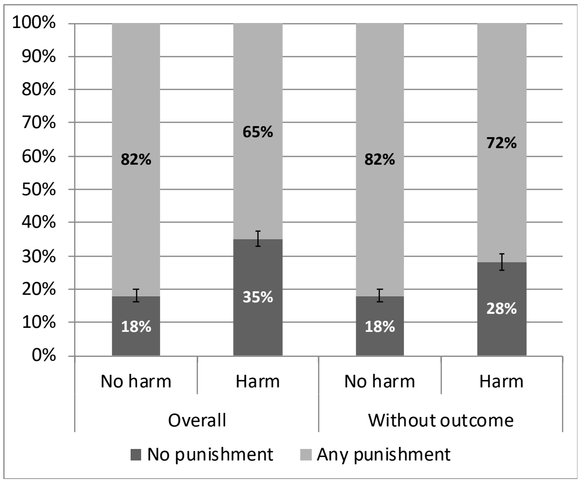

Figure 2: Percentage of punishment responses in different study conditions: overall sample and sample without people who referred to the outcome alone (error bars indicate standard error).

Discussion

Experiment 2 demonstrated that the blame blocking effect persists even after excluding participants who interpreted the punishment question as referring to the outcome alone. This suggests that the "pragmatics account" is insufficient to fully explain the effect.

Notably, however, the effect is driven by participants who say that they interpreted the punishment question to refer to "both" the attempted harm (sprinkling poppy seeds) and the coincidental harm (death by hazelnuts). It is weak, and perhaps entirely absent, among participants who say they interpreted the punishment question instead to refer exclusively to the attempted harm.

On the one hand, this data is consistent with a natural interpretation of the two-process model, according to which some attention to the harm (in the coincidental harm case) is necessary to produce the competitive interaction between the causal process and the mental state process. After all, according to the two-process model, it is precisely the attention paid to (the absence of) causal responsibility for the coincidental harm that competitively blocks assessment of the culpable mental state of the attempted harmdoer in the coincidental harm case.

On the other hand, this data is also consistent with an alternative explanation that we have not yet considered. A variant of this alternative is proposed by Malle, Guglielmo, and Monroe (2014)[7], who argue that blame in the coincidental harm case is the *average* of a high level of blame for the attempted harm and a low level of blame for the coincidental harm, whereas the blame in the no harm case is simply the high level for the attempted harm. In other words, when people interpret the punishment question as "both" about the attempted harm and the coincidental harm, they may therefore assign amount intermediate between these two values.[8]

We are in a good position to evaluate this alternative by analyzing an additional element of our data. Recall that, among people who said they interpreted punishment to refer to "both" the harm and the attempt, we then asked them to assign a specific amount of punishment to just the harm (presumably zero, as the harm was coincidental), and a specific amount of punishment to just the attempt. Thus, we can ask whether the *total* amount of punishment assigned was, on average, lower than the amount of punishment assigned to the *attempt alone*. This would be necessarily true on Malle and colleagues' hypothesis, since they assume that the total amount of punishment will be intermediate between the amount of punishment assigned to each of the two elements individually. Contrary to this prediction, however,

[6] We focused on the binarized results because these were the key analyses to report the blame blocking effect in the original study (Cushman, 2008). However, for the sake of transparency we also report the analyses with the full range of responses. Following the original study we ran Mann Whitney Ranked Sums tests for three groups of participants: (1) in the overall sample participants assigned more punishment in the "No Harm" case than in the "Harm" case (Mdn_{H} = 3, Mdn_{NH} = 4). The difference was statistically significant, Z(840) = 3.372, p = .001; (2) in the analysis without people who referred to the outcome alone, the difference was marginally statistically significant (Mdn_{H} = 4, Mdn_{NH} = 4), Z(759) = 1.897, p = .058, (3) in the analysis without people who referred to the outcome alone and both outcome and action, the difference was not statistically significant (Mdn_{H} = 4, Mdn_{NH} = 4), Z(657) = 0.565, p = .572. Note that for these supplementary analyses Cushman (2008) reported marginally significant results with p = .11 (cf. p. 374).

[7] Malle et al. (2014) apply this reasoning to judgments of blame, but they point out that there exists a similar pattern for judgments of criminal liability (p. 169). Since in the paper we focus on judgments of punishment, we consider their proposal in relation to this class of moral judgments.

[8] This proposal is similar to the account examined above as it assumes that people in different conditions may be judging the perpetrator for two different events. However, while the former would perceive the blame blocking effect as a result of people interpreting the dependent measure in terms of outcome alone, Malle and colleagues` account would explain it in terms of people judging the perpetrator for the conjunction of attempt and outcome.

the mean punishment for the "composite" event (*M* = 3.57, *SD* = 3) was not lower than the mean punishment for the attempt alone (*M* = 3.29, *SD* = 2.73).[9] Similarly, 43% (*n* = 44) of these participants assigned "no punishment" to the composite event, while 40% (*n* = 41) assigned no punishment to the attempt alone (consistent with the principle "no harm, no foul"). This suggests that people did not, for instance, feel that the attempt was punishable and yet assign no punishment for the *composite* event because one cannot be punished at all for something they did do *and something they did not*.

Collectively, these data further speak against pragmatic interpretations of the blame blocking effect. Even among people who say that they judged the coincidental harm case in part by assigning punishment to the attempted harm—and even when asked to make a punishment judgment strictly about that attempted harm—the blame blocking effect persists.

General Discussion

Recent proposals have advanced a potential alternative explanation of the blame blocking effect that does not invoke two independent processes of moral evaluation. According to the "pragmatics alternative" people could have interpreted the pragmatics of punishment question differently across versions of the story with and without harm that were used to trigger this effect in studies by Cushman (2008). In order to address this alternative we conducted two experiments. In Experiment 1 we used two different versions of the punishment question in order to test if the blame blocking would remain after we specify the question as more clearly referring to the total set of the agent´s behaviors as the target of punishment (developing previous research by Prochownik, 2017 and Prochownik & Unterhuber, 2018). The results indicated that the blame blocking effect occurs regardless of the phrasing of the dependent measure. However, a potential limitation of this study was that it did not completely exclude the possibility that some participants could have still interpreted the punishment question (even when specified) as referring to the outcome alone. To address this problem, we conducted another experiment. In Experiment 2 we replicated one of the original studies by Cushman (2008) with one modification: after assigning a specific amount of punishment to the defendant, participants in the "Harm" condition indicated what they thought the punishment was meant to be for (for the attempt, for the outcome or for both the attempt and the outcome). The blame blocking effect was present in the overall sample and also after excluding participants who indicated they thought the punishment was for the outcome alone. Taken together, these two experiments suggest that the blame blocking effect cannot be accounted for in terms of people´s presumed tendency to interpret the punishment question in terms of outcomes rather than actions.

In addition, Experiment 2 speaks against a slightly different proposal by Malle, Guglielmo, and Monroe (2014). According to these researchers, people´s judgements are for the agent´s attempt alone in the "No Harm" case, while they result from the average of the punishment for the attempt and the outcome in the "Harm" case. Yet, in contrast to this prediction, our results suggest that people judge the "composite" event of the attempt and outcome almost the same as they judge the attempt alone. Therefore, the blame blocking effect is not likely to occur due to averaging.

Experiment 1 also recommends some further developments of the two-process model itself. In its original formulation, the model remained open regarding what type of consequences trigger the causal process of moral evaluation, and can eventually lead to the blame blocking effect. Scenarios used by Cushman (2008) featured harms to humans and presented coincidental harms that were roughly the same as the harms intended and attempted by the perpetrators (e.g., the same victim dies, and by similar means to those originally intended). The presence of the blame blocking effect across different scenario contexts in our first experiment suggests that this effect is robust across different types of harms including both severe bodily injuries and gross harms to property, as well as coincidental harms that are somewhat different than originally intended. This suggests a modification to the two-process model such that the blame blocking effect can be triggered by a wide variety of harmful events. However, more research in this direction would help to delineate the scope of the blame blocking phenomenon and the specific conditions under which it occurs.

Finally, although, our experiments suggest that the blame blocking effect cannot be accounted for simply in terms of people interpreting the dependent measure as referring to the outcomes and not the actions (thus outcomes do not speak louder than actions!), future research must test additional possible alternative explanations of blame blocking. Two stand out. First, it might be that people diminish the punishment in the "Harm" case comparing to the "No Harm" case because they think the harmful outcome would have occurred regardless of the agent's attempt to harm (e.g., because Smith would have been killed by the chef anyway people may perceive Brown´s attempted homicide as redundant and release him from responsibility). Second, the "Harm" case is more complex and contains more information than the "No Harm" case that may distract participants (e.g., that Smith ends up being killed by the chef may be an extra element drawing people´s attention away from Brown´s attempted homicide). Finally, in addition to testing these alternatives, the two-process model would benefit from more thorough research on how exactly the two processes of moral analyses operate and interact in everyday moral decision making.

[9] Because the mean might not be well suited to the ordinal scale like the one we used, we also calculated medians and modes for the two punishment questions. The results did not differ, as the medians (2="6 months") and modes (1="No punishment") were the same for both the main punishment rating and the punishment for the attempt alone (*n* = 102).

Acknowledgments

This work was supported by the research project No. 2014/13/N/HS5/01137 funded by the National Science Centre, Poland.

We would like to thank Alex Wiegmann and four anonymous reviewers for very helpful comments on the previous version of the manuscript.

References

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556-574.

Alicke, M. D., & Rose, D. (2012). Culpable control and causal deviance. *Social and Personality Psychology Compass, 6*(10), 723-735.

Carlsmith, K., & Darley, J. M. (2008). Psychological aspects of retributive justice. *Advances in Experimental Social Psychology*, 40, 193-235.

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353-380.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6-21.

Darley, J. M., & Shultz, T. R. (1990). Moral rules - their content and acquisition. *Annual Review of Psychology, 41*, 525-556.

Fincham, F. D., & Jaspers, J. (1979). Attribution of responsibility to the self and other in children and adults. *Journal of Personality and Social Psychology, 37*(9), 1589–1602.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 52*(5), 449-466.

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and social psychology bulletin*, *36*(12), 1635-1647.

Hagan, J. P., & Royzman, E. (2017). The shadow and the tree: inference and transformation of cognitive content in psychology of moral judgment. In J. F. Bonnefon & B. Trémolière (Eds.), *Moral inferences. Current issues in thinking and reasoning*. Oxford: Routledge.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

Martin, J. W., & Cushman, F. (2015). To punish or to leave: Distinct cognitive processes underlie partner control and partner choice behaviors. *PloS ONE*, *10*(4).

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition, 147*, 133-143.

Piaget, J. (1965). *The moral judgment of the child*. New York, NY: Free Press. (Original work published 1932)

Prochownik, K. (2017). Do people with a legal background dually process? The role of causation, intentionality and pragmatic linguistic considerations in judgments of criminal responsibility. In J. Stelmach, B. Brożek & Ł. Kurek (Eds.), *The province of jurisprudence naturalized*. Warsaw: Wolters Kluwer.

Prochownik, K., & Unterhuber, M. (2018). Does the blame blocking effect for assignments of punishment generalize to legal experts? In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of cognitive science society* (pp. 2285-2290). Austin, TX: Cognitive Science Society.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164-176.

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science, 13*(3), 238.

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development*, *57*(1), 177-184.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York, NY: Guilford Press.

Wiegmann, A., Samland, J., & Waldmann, M. R. (2016). Lying despite telling the truth. *Cognition*, *150*, 37-42.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*(20), 8235.