The Role of the Primary Effect in the Assessment of Intentionality and Morality

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Alex Wiegmann (alex.wiegmann@psych.uni-goettingen.de)
Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

In moral dilemmas performing an action often leads to both a good primary and a bad secondary effect. In such cases, how do people judge whether the bad secondary effect was brought about intentionally, and how do they assess the moral value of the act leading to the secondary effect? Various theories have been proposed that either focus on the causal role or on the moral valence of the secondary effect as the primary determinants of intentionality and morality assessments. We present experiments which show that these theories have neglected a further important factor, the primary effect. A new theory is proposed that is based on the key assumption that people's judgments of intentionality and morality depend on the strength of assumed reasons the agent has for the primary and secondary effects.

Keywords: intentional action; doctrine of double effect; moral dilemmas; causal structure; trade-off.

Introduction and Overview

How do people judge whether an effect was brought about intentionally? Since Joshua Knobe's seminal paper (2003) this question has gained increasing attention in recent cognitive science (see Waldmann, Nagel, & Wiegmann, 2012, for an overview). In the present paper we focus on the following question: If an agent in a moral dilemma performs an action that leads to two effects[1], a good one, which is the primary goal of the agent, and a bad secondary one, how do subjects judge whether the bad effect was brought about intentionally?

Currently two very different theories are competing to answer this question. Knobe (2003) has proposed that intentionality attributions regarding the secondary effect depend on its moral value. If the secondary effect is morally bad, his theory predicts high intentionality ratings, whereas the ratings are lowered when the secondary act is good. The second class of theories focuses on the causal structure linking primary and secondary effects. If the secondary act is a means for achieving the primary one, then high intentionality ratings and low morality judgments are to be expected. When the secondary act is just a causal side effect, intentionality ratings are predicted to be lower, and morality judgments higher (see Mikhail, 2011).

We are going to propose a third account. Our main claim is that intentionality attributions are a function of the strength of the assumed reasons that can be attributed to the agent for causing the primary and the secondary effect. These reasons are inferred on the basis of observable cues, with the inferences being influenced by both the causal structure of the scenario, and the trade-off between good and bad effects. Morality judgments are also influenced by the trade-off, although, based on previous research, we expect that moral judgments are influenced less by causal role than intentionality judgments (see Waldmann et al., 2012; Waldmann & Dieterich, 2007). We will outline these three theories in greater detail below, and then present two experiments testing them.

Knobe's (2003) Side-Effect Effect

Theories of intentional action have gained a lot attention since Knobe (2003) had discovered that subjects rate the assumed intentionality regarding a secondary act higher when this effect is morally bad than when it is morally good. Consider Knobe's (2003) famous vignette:

> "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." (p. 191)

In a second version of this scenario, the word "harm" was replaced by "help". When subjects were asked whether they think the chairman intentionally harmed the environment, 82% answered in the affirmative. In contrast, in the help condition 23% said that the agent did bring about the good side effect intentionally. Knobe concluded that in judging whether the side effect (harming and helping the environment, respectively) was brought about intentionally, the (moral) value of the side effect is crucial. People seem considerably more willing to say that a side effect was brought about intentionally when they regard it as bad, than when they regard it as good.

In the present research we focus on moral dilemmas in which the secondary effect is kept invariant and constantly bad (killing one person). For such scenarios, Knobe's (2003) version of his theory entails two theoretical predictions: (1) As long as the secondary effect is invariant, equal degrees of intentionality ratings should be observed. Since in our cases, the secondary effect is bad, generally high ratings are to be expected. (2) The moral value of the secondary effect is an intrinsic property of the act leading to the bad outcome. For example, harming nature is bad, helping nature is good. Of course, Knobe's (2003) theory

[1] Following the literature we here use the term primary and secondary effect as referring to acts leading to a good or bad outcome (e.g., harming nature).

does not explicitly rule out that other factors, such as the primary effect, may play a role in the assessment of intentionality and moral value. However, the present version of the theory neglects the potential role of the primary effect, and solely focuses on the role of the secondary effect. This seems like a crucial oversight given that the primary effect is constitutive for assigning the other effect the role of being secondary. Moreover, it is far from clear whether this additional factor can be simply added to the present theory without changing key theoretical assumptions (see also General Discussion).

The Doctrine of Double Effect

The Doctrine of Double Effect (DDE) is one of the oldest and best known moral principles (cf. Mikhail, 2011). There are several versions of the DDE, here is one by Timmons (2002): Whenever an action would produce at least one good effect and one bad or evil effect, then one is permitted to perform the act if and only if all of the following conditions are met:

1. The action in question, apart from its effects, must not be wrong.
2. The bad effect must not be intended by the agent. There are two principal ways in which an effect might be intended:
    a) Any effect that is a chosen end of action is intended.
    b) Any effect that is a means for bringing about some intended end is also intended.
3. The bad effect must not be "out of proportion" to the good effect.

To illustrate the DDE, let us apply it to two popular trolley cases (see Waldmann et al., 2012, for an overview of trolley research). In case one ("bystander"), a runaway trolley is threatening five people and the only possibility to rescue the five people is to re-direct the trolley onto another track where only one person would die. In case two ("push"), the initial situation is the same but this time the only possibility to rescue the five is to throw a heavy man from a bridge into the path of the trolley. The trolley would be stopped due to the weight of the heavy man and, not surprisingly, the heavy man would die in the collision. According to the DDE, redirecting the trolley might be permissible but throwing the heavy man from the bridge is not, since throwing the heavy man is a means for bringing about the intended end (2b).

Although the primary goal of the DDE is to offer a guide for the moral evaluation of moral dilemmas, the DDE also provides a criterion that tells us when an act or an effect is intended: Any effect that is a means for bringing about some intended end, is itself intended, whereas secondary effects that are only causal side effects are merely foreseen, but not intended. Cushman and Young (2010) have presented a series of studies in which they showed that means elicited higher intentionality ratings than side effects. This pattern holds for both moral dilemmas and isomorphic non-moral scenarios. Again, the primary effect does not figure in the predictions of intentionality. It does play a limited role in permissibility judgments, however, but the DDE only states that the secondary bad effect must not be out of proportion to the good effect.

Trade-Off of Lives in Moral Dilemmas

Whereas Knobe (2003) focuses on the moral value of the secondary effect in his predictions about intentionality ratings, the DDE focuses on its causal role (means vs. side effect). However, there is an additional neglected factor that is actually constitutive for the labeling of one of the effects as secondary: the primary effect. A typical feature of moral dilemmas is that they contain a trade-off between acts saving and killing people. For instance, in the standard version of the bystander dilemma, the act causes the death of one person and saves five persons. If you consider instead a version of this case in which acting kills one person but nobody is saved, it seems clear that causing the death of the one person was brought about intentionally because otherwise it is hard to explain why the agent should have intervened. However, if one person is killed and one is saved one might judge that saving the one person was the primary intended goal. If more lives are saved than killed it seems even more likely that subjects will view the good outcome as a mitigating reason for generating the bad side effect, and will therefore be less inclined to regard the bad effect as strongly intended. Thus, based on the assumption that intentionality ratings are influenced by the inferred reasons an agent might have for causing a bad secondary effect, the prediction can be derived that intentionality ratings should decrease the more reasons the agent has for causing the primary effect. In trolley dilemmas, this means that lower intentionality ratings with respect to the bad secondary effect should be expected, the more people are saved by the act. The number of saved people provides excellent reasons for acting, and allows the agent to dismiss the bad secondary effect as intentionally pursued.

The trade-off between primary and secondary effect may not only affect intentionality assessments but also moral evaluations. Again here our trade-off hypothesis makes predictions different from Knobe (2003) and the DDE. Whereas Knobe (2003) treats the badness of the secondary effect as an intrinsic feature of this effect, the DDE predicts that secondary effects should be judged generally worse when they constitute a means compared to a side effect. By contrast, our trade-off hypothesis claims that the judged badness of the secondary effect is not only based on intrinsic harmful features of the corresponding act, but also on the strength of reasons for accepting a bad secondary effect in light of a good primary goal. Thus, again we predict that moral evaluations will be sensitive to the relationship between the primary and secondary effects. Empirically this prediction is supported by research presenting trolley dilemmas with catastrophe conditions in which one person needs to be killed to save, for example, 1,000.000 (Nichols & Mallon, 2006; see also Bartels, 2008). Typically permissibility judgments rise with increasing

numbers of saved people. Intentionality judgments have not been studied in these experiments, however.

The Role of Causal Structure

So far we have elaborated the role of the primary effect as a mitigating reason for accepting a bad side effect. However, reasons for acting are not only influenced by the size of the primary effect, but also by the causal relationship between primary and secondary effect. Following the DDE, we argue that secondary effects that play the role of means need to be intended more strongly than when they are in the position of side effects because agents should be aware of the fact that the secondary effect constitutes a necessary step on the path to the primary goal. Thus, causal structure also should determine the strength of reasons an agent has for causing a secondary effect.

Combining Trade-off and Causal Structure

We have identified two sources of reasons an agent might have for acting in a moral dilemma in which a varying primary effect is pitted against an invariantly bad secondary effect: The causal role of the secondary effect and the trade-off between the primary and secondary effects. There are two possibilities how the two factors can be combined. One possibility would be to claim that the primary effect only matters when the secondary effect is a side effect, not when it is a necessary means for the primary effect. This pattern of interaction could be motivated by the assumption that means are necessary steps on the way to the more distant effects, and therefore need to be intended regardless of the quality of the distant (i.e., primary) effects. The other possibility might be that the strength of reasons for the primary goal affect intentionality ratings in both conditions, but to a stronger extent in the side-effect condition in which it is easier to imagine that people just foresee, but do not intend the secondary effect. In this case, we also expect an interaction but also a main effect that is driven by the strength of the primary goal.

The same alternatives also arise for moral judgments about the badness of bringing about the secondary effect (i.e., killing one person). Whereas Knobe (2003) predicts invariant ratings signifying the invariant badness of killing one person, the DDE predicts a main effect driven by the causal role of the act leading to death. By contrast, based on previous research we expect that mitigating factors arising from the evaluation of the goodness of the primary effect will also affect the moral assessment of the secondary effect. However, previous findings on moral judgments in trolley dilemmas cast doubt on the hypothesis that causal role plays the same role in moral judgments as in intentionality judgments. The empirical evidence rather points to other factors (i.e., aversiveness, attentional focus, directness) as the crucial factors affecting moral judgments whereas causal role vanishes as a factor once its confounds are controlled (see Waldmann et al., 2012). Thus, intentionality and moral judgments need not be driven by the same factors, as implied by the DDE.

Experiment 1a

Experiment 1a focuses on intentionality attributions. To test our predictions we used two variants of trolley cases in which we manipulated the causal structure (means versus side effect) between subjects, and the number of lives saved (0, 1, 5, 100) within subjects. While the number of lives saved was manipulated across conditions, the secondary effect always involved killing one person.

Figure 1: Illustration of the bystander scenario (Experiment 1).

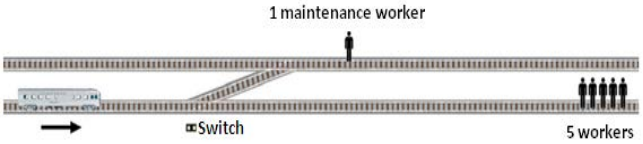

Figure 2: Illustration of the trap scenario (Experiment 1).

Design, Materials, and Procedure 140 subjects were recruited for a compensation of £ 0.50 via an online database located in the U.K.. Subjects were randomly assigned to the means or side-effect condition, and were then presented with the corresponding scenarios in which the number of lives saved was varied in a randomized order within subjects (0, 1, 5, 100 victims). It is important to note that the 0 condition only served as a control condition because in a condition in which one person was killed without anybody being saved, the means or side effect relations cannot be realized. This condition is expected to yield maximal intentionality and badness ratings because there are no mitigating reasons for the act.

The initial situation was identical in both conditions:

> "On the test ground of a modern railroad property an unmanned speed train (that normally can be remote-controlled) is out of control due to a technical defect. This speed train is heading towards five railroad workers that are checking the tracks. Since these workers are wearing a novel type of hearing protection, they would not notice the speed train on time and hence would be run over by it. Peter, an employee of the rail track control center, recognizes the upcoming accident. However, it is not possible to stop the train on time anymore."

Next the conditions, which were additionally illustrated (see Fig. 1, 2), varied (examples with 5 saved workers).

Side-Effect Condition ("bystander"):

> "There is only one possibility to avoid the death of the five workers: Peter could push a button and thereby re-direct the speed train from the lower track onto a parallel upper track before it reaches the five workers on the lower track. On the upper parallel track the speed train would run into a worker maintaining the tracks. The maintenance worker on the upper track would lose his life due to the collision."

Means Condition ("trap"):

> "There is only one possibility to avoid the death of the five workers on the tracks: Peter could push a button that would open a trap door and thereby causing a maintenance worker on top of the bridge to fall on the tracks. The speed-train would collide with the maintenance worker and be stopped before it reaches the five workers on the track. The maintenance worker would lose his life due to the collision."

Both scenarios ended as follows:

> "Peter understands the situation and knows the consequences of the action just described. Peter decides to throw the switch and the maintenance worker dies."

In the control condition (zero people saved) we used the same instructions mentioning re-directing of the train or the trap door but these acts were not motivated by saving anybody. After reading the scenario description, subjects were asked to judge whether Peter "caused the death of the maintenance worker intentionally" on a six-point Likert scale ranging from "certainly no" to "certainly yes." On the last page, subjects were asked some demographic questions, and were given a simple logical question unrelated to the experiment to identify the subjects who did not pay attention to the task.

Results and Discussion Eight subjects were removed from the analyses because they failed to solve the logical question or completed the whole survey in less than a minute. The results for all scenarios are depicted in Figure 3.

We generally excluded the control condition in the ANOVAs of all experiments because here the difference between means and side effects could not be realized. Looking at the remaining three conditions, the intentionality ratings proved generally higher in the means than in the side-effect condition ($F_{1,\ 130}$=22.158, p<.0001). Moreover, the analysis yielded a significant interaction, $F_{2,260}$=3.0911, p<.05.

More detailed planned comparisons showed that in both conditions intentionality ratings decreased with increasing numbers of lives saved. However, in the means condition the ratings were statistically equivalent in the three conditions in which the means relation could be realized (1, 5, 100). In contrast, in the three side-effect conditions (1, 5, 100) ratings dropped from 3.97 when one life was saved to 3.41 when one hundred lives were saved (p<0.001). In both causal conditions, the control scenario in which nobody was saved by killing one received significantly higher ratings than the averaged ratings for the three other scenarios (means condition: p<0.01; side-effect condition: p<0.0001). Thus, in general the presence of a primary good effect lowers intentionality assessments regardless of causal status, but the influence of the quantitative size of the primary goal affects the side-effect condition more than the means condition.

Interestingly, subjects' intentionality ratings for the cases in which 5 and 100 lives are saved did not differ. Possibly subjects are only sensitive to qualitative differences, that is, it only matters if fewer lives (0), just as much (1), or more lives (5 and 100) are saved but not how many more are saved.

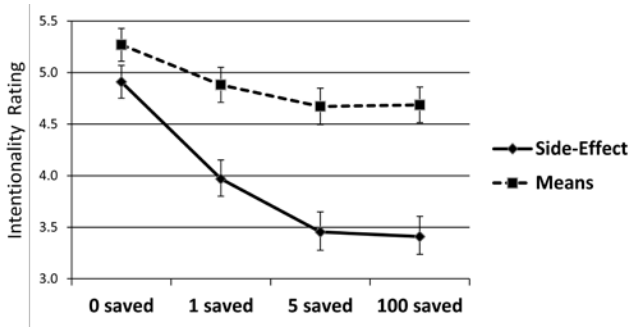

Figure 3: Results of Experiment 1a. Error bars indicate standard error of means.

Experiment 1b

Experiment 1b uses the same conditions and manipulations as Experiment 1a. The only difference is that we asked a different test question with which we assessed the moral evaluation of the secondary effect: "How bad is Peter's causing the one person's death?" Subjects responded using a 6-point rating scale ranging from "not bad at all" to "very bad."

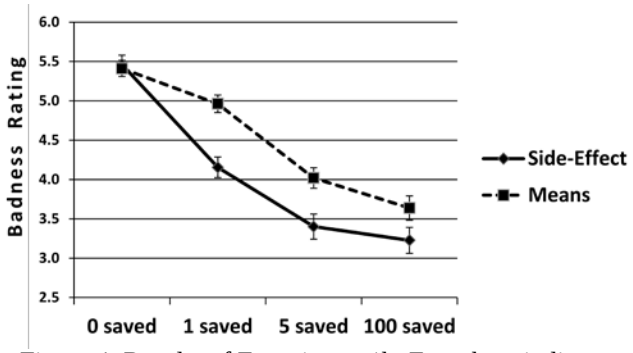

Figure 4: Results of Experiment 1b. Error bars indicate standard error of means.

Results and Discussion The results are based on 147 participants from the same online site used in the previous experiment (29 from the larger sample needed to be removed). Figure 4 shows the results. Excluding the control condition (0 people saved) which yielded uniformly high badness ratings, we found a main effect between the side-effect and means condition, $F(_{1,145})$=7.43, p<.01, along with a main effect of the size of the primary effect, $F(_{2,290})$=56.004, p<0.0001. Thus, in contrast to the predictions of the DDE, the primary effect affects moral evaluations. The main effect between side-effect and means conditions is consistent with the predictions of the DDE, but could also indicate differences in aversiveness of death (Waldmann & Dieterich, 2007). Experiment 2 will address this question.

Experiment 2a

One might object that in Experiment 1a the act in the means condition is more aversive than the one in the side-effect condition (falling through a trap-door and getting hit by a train vs. just getting hit by a train). In fact, the moral judgment data revealed that the trap dilemma was generally assessed as more aversive than the bystander dilemma. Hence, the difference in intentionality ratings could be caused by this factor and might not be due to different causal structures. To counter such an objection we designed a new experiment with more similar scenarios. In Experiment 2a we focused on intentionality assessment, in Experiment 2b on morality judgment.

Design, Materials, and Procedure 142 subjects were recruited and compensated as in Experiment 1a. The same design as in Experiment 1a was used. Moreover, we used the same side-effect scenarios except that we placed the single worker inside a train to make his death appear less cruel. Additionally, we made the means scenario less violent by using a different, more technical mechanism. The key differences between the old and new means scenarios lie in the point of intervention and in the equipment the workers are wearing. In the new means scenario it is not the runaway train that is redirected but the train containing the one worker who is going to be killed. Furthermore, the workers in the means condition are wearing a security system that causes all running trains to stop when a worker's heart stops beating. The crucial paragraph that distinguishes the means and side-effect conditions is the following:

> "There is only one possibility to avoid the death of the five workers: Every worker is wearing a security system that causes all running trains to stop when a worker's heart stops beating. Peter could throw the switch (by pushing a button), and thereby redirect the yellow train carrying one worker from the parallel upper track onto the main track. The speed-train would collide with this yellow train so that the one worker in this train would instantly lose his life in this accident. However, due to the security system the one worker in the yellow train is wearing, the speed train would stop before it reaches the five workers."

The security vest was introduced to highlight the role of the single victim as a necessary means. Without the vest one might argue that the five are saved by the train, and the single worker's death is only a side effect. The procedure and test questions were otherwise identical with the ones used in Experiment 1a.

Results and Discussion 22 subjects were removed for the same reasons as in the previous experiments. The results for all scenarios are depicted in Figure 5 and based on 120 subjects. Again the intentionality ratings were generally higher in the means than in the side- effect condition, $F_{1,118} = 4.51$; $p < .05$ (the control condition was again excluded from the analyses). Furthermore, this main effect was moderated by a significant interaction, $F_{2,236} = 5.23$; $p<0.01$. Intentionality ratings only decreased with increasing numbers of saved lives in the side-effect but not in the means condition. In the means condition, ratings decreased from 4.42 when no lives were saved to 4.31 when one hundred lives were saved ($p=0.55$). In the side-effect condition, ratings dropped from 4.20 when no lives were saved to 3.64 when one hundred lives were saved ($p<0.01$). Again, in both conditions ratings for the control scenario were significantly higher than the weighted ratings for the remaining three scenarios (means condition: $p<0.01$; side-effect condition: $p<0.0001$)

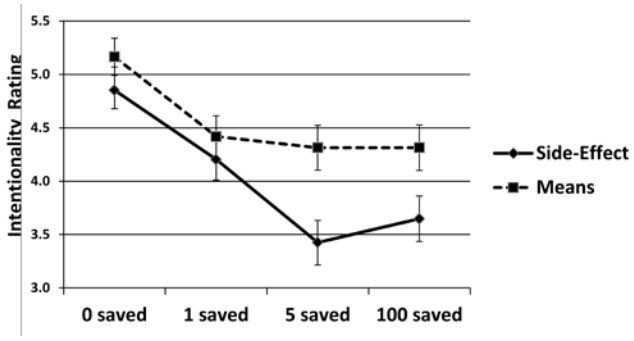

Figure 5: Results of Experiment 2a. Error bars indicate standard error of means.

As in Experiment 1a there was again no difference between the scenarios in which 5 and 100 lives are saved indicating that qualitative differences rather than quantitative differences might be crucial for ratings of intentionality.

Experiment 2b

Experiment 2b uses the same conditions and manipulations as Experiment 2a along with the moral test question from Experiment 1b.

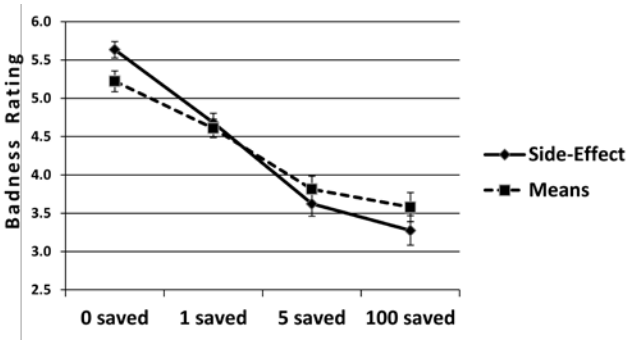

Figure 6: Results of Experiment 2b. Error bars indicate standard error of means.

Results and Discussion The results are based on 114 participants from the same online site used in the previous experiments (29 needed to be removed). Figure 6 shows the results, which are clear. We apparently managed to equate the moral aversiveness of harming the one worker. Thus,

our intentionality rating patterns are not moderated by different degrees of aversiveness across conditions. The only significant effect is the overall downward trend that is driven by the size of the primary effect. The more people are saved, the less bad the secondary effect was assessed, $F_{2,224}$=54, p<0.0001.These results refute the assumption of the DDE that there is a general moral difference between side-effect and means conditions beyond the typically confounded differences of aversiveness of death (see also Waldmann & Dieterich, 2007; Waldmann et al., 2012).

General Discussion

We have proposed a new theory of intentionality attributions for moral dilemmas, according to which judgments of intentionality regarding a secondary effect depend on the strength of reasons for accepting the secondary effect in the pursuit of the primary goal. We showed that this assessment is a joint function of the causal role of the secondary effect, and the size of the mitigating reasons provided by the primary goal. The more people are saved by an act that also has a bad secondary effect, the better the secondary act is evaluated on a badness scale, and the less intentionality is attributed to the agent with respect to this effect. Whereas moral evaluations were only sensitive to the size of the mitigating reasons provided by the primary cause and the general aversiveness differences between the scenarios, we found reliable interactions between the size of the primary effect and the causal role of the secondary effect in the intentionality attributions. Means generally were rated lower than ends, but beyond this effect the size of the mitigating reasons did not further affect intentionality assessments. In contrast, in the side-effect conditions intentionality assessments were lowered by the size of the primary effect.

The findings provide important constraints for theories of intentionality attributions. They show, for example, that Knobe's (2003) theory which focuses on the moral value of the secondary effect is at least incomplete. We doubt that the role of the primary effect can simply be added to the present theory, however. To explain our findings the assumption needs to be made that the moral evaluation of the secondary effect is a function of a trade-off with the primary effect. Moreover, in Knobe's (2010) theories causal structure is not viewed as a determinant of intentionality attributions. Causal intuitions are rather, similar to intentionality assessments, conceived as being triggered by moral evaluations of the outcomes. Other theories that have been proposed as competitors to Knobe's (2003) account have also difficulties with our findings since they also only focus on the secondary effect while typically holding the primary effect constant.

Theories based on the DDE are also only partially consistent with our results. In contrast to the predictions of the DDE we showed that the causal structure is not the sole determinant of intentionality attributions but needs to be augmented by our trade-off factor. Moreover, our moral evaluation data contradict the predictions of the DDE. Once the aversiveness of death differing between the means and side-effect conditions is equated, differences in moral judgments disappeared. This finding is consistent with previous results, and casts doubt on the adequacy of the DDE as a theory of moral evaluations (see Waldmann & Dieterich, 2007).

One general important result is therefore that intentionality and moral judgments, which are closely linked in the DDE, are not influenced by the same factors. Whereas both the primary effect and the causal role of the secondary effect influence intentionality assessments, only the former factor has an impact on moral evaluations, once aversiveness is controlled. This finding casts doubt on the often held assumption that intentionality and moral judgments are closely linked.

Acknowledgments

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG WA 621/21-1), and the Courant Research Centre "Evolution of Social Behaviour', University of Göttingen (funded by the German Initiative of Excellence).

References

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, *108*, 381-417.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from non-moral psychological representations. *Cognitive Science*, *35*, 1052–1075.

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

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*, 315–329.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language*, *23*, 165.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment.* New York: Cambridge University Press.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*, 530-542.

Timmons, M. (2001). *Moral theory: An introduction*. Rowman & Littlefield Publishers, Inc.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, *18*, 247–253.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 364-389). New York: Oxford University Press.