This is a previous draft of the article "Explaining the Knobe effect" that is published in: Luetge, Hannes Rusch & Matthias Uhl (eds.), *Experimental Ethics*. Palgrave Macmillan 65-79 (2014).
Please, do not quote from the draft.
Contact: v.wagner@uni-konstanz.de

Verena Wagner: Explaining the Knobe effect

Abstract: In this paper I reject the view that the famous 'Knobe effect' reveals an asymmetry within people's judgments concerning actions with good or bad side effects. I agree with interpretations that see the ascriptions made by survey subjects as *moral* judgments rather than ascriptions of intentionality. On this basis, I provide an explanation as to why people are right in blaming and 'expraising' agents that acted on unacceptable motives, but praise and excuse agents who meet intersubjective expectations by acting on acceptable motives. The asymmetry only arises when blameworthiness and praiseworthiness are seen as instances of one and the same concept: moral responsibility. This analysis is backed by a study of Joshua Shepherd who extended and varied Knobe's original vignettes.

1 Introduction

Joshua Knobe famously conducted several case studies in which he confronted survey subjects with a chairman who decides to start a new program in order to increase profits and by doing so brings about certain foreseen side effects. Depending on what the side effect is in the respective case, either harming or helping the environment, people gave asymmetric answers to the question as to whether or not the chairman brought about the side effect *intentionally*. 82 per cent of those subjects confronted with the harm scenario judged the chairman to have *harmed* the environment intentionally, but only 23 per cent of the subjects confronted with the help scenario judged the chairman to have *helped* the environment intentionally (Knobe 2003a). This at first sight surprising asymmetry is called the "Knobe effect" and together with the explanation Knobe provided for his findings it gave rise to a great amount of responses in the literature. Many follow-up studies were conducted that were meant either to confirm or to reject the Knobe effect and many comments were written on how to interpret the data correctly. Most of these very different responses share the view that the asymmetry is surprising and has to be explained: the chairman went through the same reasoning and decision process, both side effects are equally foreseen by the chairman, his motivation (which is making profit) is both times the very same and there is no external influence that could explain why people judge his *harming* the environment to be brought about intentionally, but *helping* the environment not to be. The only difference seems to be that harming the environment is considered to be *bad* and helping the environment is considered to be *good*. Indeed, this asymmetry is in need of explanation.

In this paper I aim at providing an explanation of the Knobe effect that is based on the claim that people are in fact judging an agent morally when they ascribe intentionality to an agent's behaviour. This kind of interpretation involves no new insight and variations of it are given in several responses to the Knobe effect, e.g. in Adams and Steadman 2004ab. But unlike many other defenders of this claim, I do not think that this is sufficient for an explanation. Such an approach can be a first step, but it provides no explanation of the asymmetry itself, because the same asymmetry arises when the chairman is judged to be *responsible* for harming, but *not* responsible for helping the environment. A sufficient explanation additionally requires either an explanation of why the *concept* of moral responsibility, in contrast to the concept of intentional action, contains an asymmetry that justifies people's diverging judgments; or it has to explain why people wrongly apply the concept of moral responsibility in an asymmetric way. There is no explanation given by merely stating that people judge the chairman to be responsible for the bad side effect but not to be responsible for the good one under the same circumstances. I defend the view that people – though being mistaken in their asymmetric ascription of *intentionality* – are doing the right thing in judging the harming chairman to be blameworthy, but the helping chairman to be not praiseworthy. However, I aim at showing that the ascriptions interpreted in this way do not involve an asymmetry after all. The asymmetry only arises when praise and blame are subsumed under one the same concept: the concept of *moral responsibility*.

Among all the follow-up studies that were made after Knobe's original findings, there is one study I will refer to in detail. Joshua Shepherd conducted a series of surveys in 2011 that are based on Knobe's chairman cases and an example of Phelan and Sarkissian (2008) of a city planner who produces good or bad side effects in the course of cleaning an area from toxic waste. Shepherd's study is more conclusive than Knobe's, because Shepherd varies the valence of the main goal (making profit, creating jobs, cleaning toxic area etc.), the valence of the side effect (poisoning the groundwater, in-/decreasing joblessness, in-/decreasing cancer levels etc.) and he also varies the agent's verbally expressed attitude ("I don't care about", "I feel terrible about", "That's great news about" etc.).

2 Interpreting the data

Knobe's original study contains two case descriptions of a chairman who decides to start a new program in order to make profit. The only difference between the two descriptions concerned the respective side effect that resulted from starting this program: either the environment was helped or it was harmed.

> 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 [and] it will also harm [help] the

> environment.' The chairman of the board answered, 'I don't care at all about harming [helping] 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 [helped]. (Knobe 2003a)

Knobe presented the following results after questioning survey subjects as to whether the chairman brought about the respective side effect intentionally: 82 per cent of the subjects confronted with the harm scenario judged the chairman to have *harmed* the environment intentionally, but only 23 per cent of the subjects confronted with the help scenario judged the chairman to have *helped* the environment intentionally. Knobe himself interprets his findings by stating that "people's intuitions as to whether or not a behavior was performed intentionally can sometimes be influenced by moral considerations" and that "when people are wondering whether or not a given behaviour was performed intentionally, they are sometimes influenced by their beliefs about whether the behaviour itself was good or bad". (Knobe et al. 2006, p. 205) In a later article, Knobe claims "there is a psychological process that makes people more willing to apply the concept [e.g. of intentional action] in cases of morally bad side-effects and less willing to apply the concept in cases of morally good side-effects." (Pettit and Knobe 2009, p. 590) For sure, moral considerations *do* play a role here, but it is doubtful whether Knobe is right that the application of concepts is generally *affected* by moral considerations about the goodness or badness of the side effect such that people are influenced in a way they should not be.

Adams and Steadman (2004a, 2004b) interpret the data by reference to a pragmatic usage of intentional language: while "the folk do not normally possess a clearly articulated theory of the mental mechanisms of intentional action[, they] do possess a very clear notion of the pragmatic features of intentional action and talk of intentional action due to the role of talk of intention in social praise and blame." (2004a, p. 177) Further, Adams and Steadman contend:

> We suspect that what is going on in the minds of the folk is that they disapprove of the chairman's indifference to the harm of the environment. They want to blame that indifference and they know that their blame is stronger and more effective at discouraging such acts, if the chairman is said to have done the action *intentionally*. (Adams and Steadman 2004a, p. 178)

I agree with views like the one of Adams and Steadman that survey subjects actually ascribe blameworthiness to the harming chairman but withhold praiseworthiness from the helping chairman, and for this purpose they use the vocabulary of *intentionally* harming but *not* intentionally helping the environment. But subjects are mistaken here. It may be useful within a *pragmatic* context to ascribe intentionality towards an agent that is regarded as a proper target of blame ("It's your fault, you did it on purpose!") and to withhold the ascription of intentionality towards an agent who is seen as no proper target of praise ("It's not your credit, you didn't do it

intentionally!") in order to emphasize the underlying moral judgment; nevertheless, this is no good reason to conclude that the concept of intentional behaviour inherits a real asymmetry people refer to. Indeed, in both scenarios the chairman brought about the side effect intentionally: he intentionally harmed the environment and he intentionally helped the environment though he intended neither of it. Albeit both chairmen brought about their respective side effect intentionally, the one is blameworthy for the produced harm, but the other is not praiseworthy for the produced help. The asymmetry within the ascription of intentionality, leads back to this asymmetry between praise and blame. However, this explanation provides no answer to the question concerning the *source* of the asymmetry and merely locates the asymmetry somewhere else: in the concept of moral responsibility. In the following, I will refer to this asymmetry between the ascription of blame and praise rather than the asymmetry within subjects' ascription of intentional action. Further, I will argue that there is no such asymmetry when praise and blame are not subsumed under one concept.

3 Some Shepherd effects

In this section I will point out briefly in what way Shepherd extended Knobe's original setting and what effects resulted from these changes. Further, I will aim at explaining the Knobe effect and the additional effects found by Shepherd. In the surveys, Shepherd sticks to Knobe's original example of a chairman who starts a new program and produces environmental side effects. Additionally, he uses an example of Phelan and Sarkissian (2008, p. 296f) in which a city planner starts a program in order to clean up toxic waste, which also produces certain side effects. Shepherd compares the original not caring but profit driven chairman with a chairman who is still profit driven but has another *attitude* towards the side effect: he either feels terrible about harming the environment or he is happy about helping the environment. Additionally, Shepherd compares all these cases with a chairman who has a nobler *main goal*: creating jobs for the homeless and disadvantaged. Correspondingly, Shepherd makes certain variations in the city planner's example: he manipulates the valence of the main goal (making profit vs. cleaning up toxic waste), the badness of the side effects (poisoning the groundwater vs. raising joblessness) and the goodness of the side effects (decreasing joblessness vs. decreasing cancer levels). Shepherd predicts that his results will show effects on the agent's *attitude*, on the valence of the *main goal* and on the valence of the respective *side effects*. Surprisingly, only an insignificant effect for the agent's attitude emerged when the side effect was bad, but there was a significant effect on the agent's attitude when the side effect was good. Changes of the respective valence of the main goal and the side effect did produce an effect when the side effect was bad, but none when the side effect was good. Shepherd interprets these findings as follows:

> While the valence of an agent's main goal or side effect significantly impacts folk ascriptions in harming cases, the agent's attitude does so in helping cases. [...] the asymmetry might result from the spontaneous triggering of a schema linking norm violations with intentionality – a schema not triggering by instances of helping. (Shepherd 2011, p. 181)

In the following I will comment on Shepherd's assumptions and findings. I will start with my own explanation of the Knobe effect based on Shepherd's results. Then, I will discuss Shepherd's result that the agent's attitude seems to have an effect in helping cases though no significant effect in harming cases. I will not give further arguments for treating ascriptions and non-ascriptions of intentionality as moral judgments in the way Adams and Steadman propose. My central aim is to show that there is no asymmetry in blaming the harming chairman though not praising the helping chairman.

3.1 Subjective and intersubjective valence

Shepherd describes the valence of creating jobs for the homeless and disadvantaged as "better" than the chairman's original main goal of making profit. In the same sense the side effect of poisoned groundwater is set as "worse" than raising joblessness, and decreasing cancer levels is set to be "better" than decreasing joblessness. Shepherd seems to assume that people commonly share this structure of valence in general; and he seems to assume an intersubjective agreement among people's judgment according to which a main goal or a side effect is judged to be good or bad, and according to which some are "better" or "worse" than others. This is a very important assumption, because it shows on what basis survey subjects are invited to form their moral judgments. Further, it is this intersubjective agreement that underlies such valence talk that gives rise to certain moral expectations towards agents and their behaviour in certain circumstances. If the valence of the side effect's occurrence is found to be worse than the non-achievement of an agent's main goal, it is expected from the agent that she refrains from performing the action in order to avoid the side effect. The profit-driven chairman is morally expected *not* to start the program in order to avoid environmental harm, because the valence of harming the environment is seen as far worse than the valence of the chairman's not making profit. Correspondingly, avoidance of environmental harm is seen as far more important than the chairman's making profit. The chairman, of course, thinks otherwise. For him, not making profit is worse than harming the environment and, therefore, he starts the program disregarding the fact that by doing so the environment will be harmed. This does not necessarily mean that the chairman does not know about the intersubjective perspective and the corresponding expectations - as a matter of fact, the experimental setting make sure that he is aware of this: the vice-president mentions the harming side effect as a potential problem and the chairman affirms

that he knows about the resulting harm. Nonetheless, the chairman rather acts on his main goal disregarding the intersubjective expectation. That is why subjects judge him to be blameworthy for harming the environment. His *subjective* valence or preference structure "making profit is more important than the environment" does not match the *intersubjective* expectation "the environment is more important than making profit". Because of this mismatch, the chairman is blamed for having started the program disregarding the resulting environmental harm. This mismatch does not only explain why the harming chairman is judged to be blameworthy (by means of ascribing intentionality), but also explains why the helping chairman is not found to be praiseworthy for the good side effect.

Blaming an agent is not the only negative reactive attitude: one can have a negative attitude towards a person who does bad *and* towards a person who does good. But of course, one cannot *blame* a person for producing good side effects even if she did not care about producing these. Yet, what one can do is to *withhold* moral praise. Withholding praise from a person is a negative attitude, too. This kind of resentment is what I call 'expraising' an agent. An agent who is expraised is not judged as a proper target of praise as well as an agent who is excused is not a proper target of accusation or blame. The chairman who started the program in order to make profit and thereby, as a side effect, helped the environment has the very same *inacceptable* preference as the chairman who harmed the environment by starting the profitable program: neither of them cares about the environment and both strongly prefer to make profit. The chairman's subjective preference "making profit is more important than the environment" is not in accordance with intersubjective expectations. This is so independently of the respective goodness or badness of the outcome. The goodness or badness of the outcome does not change the *general negative reactive attitude*, but only determines whether an agent is to be blamed or expraised: since the action in the help scenario did not have a bad side effect, the chairman cannot be blamed for having done the action under his bad preference structure. However, he can be expraised from having performed the action that led to good side effects. Because both reactive attitudes – blaming and expraising an agent – are negative attitudes of resentment, there is no asymmetry in subjects' judgments concerning the chairman when they are read as moral judgments.

In the same way as we can have negative reactive attitudes towards agents who perform actions with bad side effects *and* towards agents who perform actions with good outcomes, it is possible to have positive reactive attitudes towards agents who perform actions with good outcomes *and* towards those who perform actions with bad outcomes. It is certainly debateable whether moral praise is reserved only for those who do better than what is morally expected or is also available for those who simply *meet* moral expectations in their actions. This is indeed an interesting and important question but it shall not be a central question of this paper. What seems to be clear is

that an agent is a proper target of moral praise only if she (at least) meets moral expectations. Knobe's original case descriptions do not provide for an agent who acts in accordance with intersubjective expectation. As I pointed out before, the chairman of the help scenario as well as the chairman of the harm scenario have unacceptable preferences: both do not care about the environment and strongly prefer making profit to caring about environmental issues. In both cases, survey subjects showed negative reactive attitudes of blame and expraise. Shepherd's extension makes us see positive attitudes towards an agent. In a variation of the help scenario, Shepherd makes the chairman say: "That's great news about helping the environment! Ultimately, though, I want to make as much profit as I can." (Shepherd 2011, p. 183) This caring chairman was judged by 62.5 per cent of the subjects to have helped intentionally, while only 33.75 per cent judged the not caring chairman to have helped the environment intentionally in a reproduction of Knobe's original study. Even if the results are not as high as in Knobe's own surveys, the difference between judgments about the caring and the not caring chairman is significant. Similarly but less significant, 54.67 per cent judged the city planner to have brought about the good side effect intentionally when he expressed a positive attitude towards it, but only 26.67 per cent agreed (and 60 per cent disagreed) that the city planner brought about the good side effect intentionally when he explained not to care about it. These results show that it is not only the *badness* of the side effect that triggers ascriptions of intentionality as Knobe claimed; when the agent shows sensibility to the goodness of the side effect, this agent is judged as having brought about the good side effect intentionally. When this ascription of intentionality is interpreted as an ascription of moral responsibility, the surveys' results indicate that the caring chairman is judged to be a proper target of praise and the caring city planner is at least considered as such. By stating that they consider it to be great news that their respective program will not only increase profit (or create jobs or clean an area from toxic waste) but additionally has good side effects, both agents at least meet the intersubjective expectation that the environment is something that has to be cared about.

After having discussed examples of positive reactions to good side effects, the question is whether there also is experimental support for my claim that people can have *positive* reactive attitudes towards an agent who performed an action that led to *bad* side effects. Again, Shepherd's study provides such a case: in the case description of the city planner who has as his main goal to clean an area from toxic waste, Shepherd changes the valence of the bad side effect from 'poisoning the groundwater' to the milder effect of 'increasing joblessness'. While 68.4 per cent of the asked subjects agreed that the bad side effect of poisoned groundwater was brought about intentionally, only 47.5 per cent agreed that increasing joblessness was brought about intentionally. Even if insignificant, there was also a difference between the city planner caring and the city planner not caring about increasing joblessness as a side effect of cleaning up toxic

waste; nevertheless, Shepherd points out that this difference "did approach significance." (Shepherd 2011, p. 176) This may not be perfectly analogous to cases of positive reactive attitudes in the good side effects cases, but it seems to speak for the interpretation that survey subjects were inclined to accept the caring city planner's preference to clean an area from toxic waste over the comparatively mild side effect of raising joblessness. At least in parts, this city planner seems to be excused from having performed an action that led to bad side effects. It can be speculated that this effect would be more significant if the case description were modified in the following way: if the city planner decides against cleaning up the polluted area, many citizens would become ill due to the toxic waste in their neighbourhood. Given this additional information, subjects would accept (if not expect from) any city planner to start the program even at the cost of increasing joblessness, because the consequences of not cleaning the area would be far worse. In this new scenario one can speculate that the city planner would be excused for having performed an action with bad side effects for a greater and intersubjectively accepted good. Survey subjects would refrain from ascribing intentionality to this city planner as a means to excuse him from having increased joblessness. Excusing an agent is, as is praising, a positive reactive attitude towards an agent's performing an action. Again, there is no asymmetry in help scenarios when we read subjects' ascription of intentionality as a moral judgment of an agent's praiseworthiness and the non-ascription of intentionality as a means to excuse an agent. In both scenarios, subjects express their *positive* attitude to the respective agent.

In summary, we have two pairs of reactive attitudes: the ascription of blame and expraise are negative reactive attitudes towards agents who act on unacceptable preferences that do not meet intersubjective expectations; and we have the ascription of praise and excuse as positive reactive attitudes towards agents who act on accepted or expected preferences. Maybe the ascription of moral praise has a special role here and requires from an agent not only to meet intersubjective expectations in her action but also to *exceed* these; yet, this is the topic of another article and cannot be discussed here.

Until now I based my explanation of the Knobe effect on Shepherd's results ignoring the fact that Shepherd explicitly separates effects that resulted from manipulating the agent's *attitude* and effects that resulted from manipulating the *valence* of the main goal and the side effect. One of his main results is that while there was a main effect on the verbally expressed attitude in help scenarios, there was none in harm scenarios; and that there was an effect in harm scenarios when he manipulated he valence of the main goal or the side effect, but none in help scenarios. In short, the agent's attitude seems to make a difference when the side effect is good, while it is the valence of the main goal and the side effect that matter when the side effect is bad. How to explain that? Is that not undermining my explanation, which generally makes use of the subject's

reference to the agent's attitude? In the next section I will discuss Shepherd's distinction of valence and attitude effects – a distinction I consider to be problematic.

3.2 Attitude as verbal expression

Shepherd separates the agent's attitude, on the one hand, and the valence of the main goal and the side effect on the other. In Shepherd's approach the valence of the main goal and the valence of the side effect are treated like agent-external factors within the case description that are said to have an impact on survey subjects' judgments, but which are considered to be distinct from the agent's attitude concerning her own action and the resulting side effect. Note that the agent's attitude refers to her own action and the resulting side effect and must not be confused with 'reactive attitudes' survey subjects have when they morally judge the agent in a case description. Shepherd identifies the agent's attitude with the *verbal expression* that the agent gives in the case description: "I don't care at all about [the side effect], I just want to [achieve the main goal]", "I feel terrible about [the side effect] but priority one is [to achieve the main goal]" or "That's great news about [the side effect], still priority one is [to achieve the main goal]". As I pointed out before, it is an interesting result of Shepherd's study that "the valence of an agent's main goal or side effect significantly impacts folk ascriptions in harming cases, [while] the agent's attitude does so in helping cases." (Shepherd 2011, p. 181)

As I will point out in the following, it is problematic to treat the change of the agent's attitude and the change of the valence of the main goal and the side effect as if they had a completely independent influence on what subjects conclude about the respective agent's motivation; the two must not be seen as distinct in the experimental setting. By changing the valence of the main goal or the side effect the agent's preference structure is affected, too, when the agent acts on the motive to achieve the main goal. An agent who performs an action disregarding known bad side effects, expresses by doing so her attitude that the achievement of the main goal is more important for her than the avoidance of the bad side effect. Merely adding a verbal expression of regret cannot overwrite the attitude that is expressed by performing the action though a verbal expression may sometimes provide some missing details about the agent's attitude. Therefore, it is not surprising that in cases of bad side effects, in which the agent decides *for* the main goal and – in a way – *against* the avoidance of the side effect, there is no significant effect visible in the results when the agent verbally expresses regret about producing the bad side effect. The reason for this is simple: had the regret been serious or strong enough, the agent would have refrained from performing the action in the first place. But since she did not, subjects are not convinced that a merely *verbal* expression truly mirrors the agent's attitude – and rightly so. Though having expressed regret about the side effect, the chairman is still judged to be

blameworthy for having harmed the environment. The reluctance to excuse the caring but still profit driven chairman explains the insignificance of an effect concerning the agent's verbally expressed attitude; this is simply not what the agent's attitude can be reduced to when the agent's action provides further and, in this case, more reliable information about the agent's motive.

Shepherd admits that it "seems likely that in certain harming cases, an agent's attitude will have a significant impact." Further, he adds that "[i]t is worth noting, however, that it seems that the impact will be much less than that of the valence." (Shepherd 2011, p. 176) Against this interpretation, I think it is a mistake to conclude from the given data that in harm cases the agent's attitude has no impact on subjects' judgments and that only the valence of the main goal or the side effect has. When the side effect of the city planner's program is worsened from 'increasing joblessness' to 'increasing cancer levels', and in both scenarios the city planner says that he does not care about the respective harm, this is highly relevant for what kind of attitude is ascribed to the agent! The agent may give the very same verbal expression in both cases, but any judgment concerning the attitude of not caring about *x* and not caring about *y* depends on what *x* and *y* stand for and what their intersubjective valence is. An agent's attitude who does not care about increasing joblessness is bad enough, but somebody who does not care about increasing cancer levels is even worse. Here, it is the verbally expressed not caring attitude that – though syntactically identical in both cases – leads to different moral judgments concerning the agent. For judging an agent's attitude it is important *what* the agent does not care about.

After having explained why a change in the *verbal* expression of an agent's attitude does not lead to significant results, another question remains: why does a change in the agent's verbal expression have an impact on survey subjects' ascriptions of intentionality when the side effect is good? There is a difference between harm and help scenarios that concerns the agent's attitude and motivation. The agent's knowledge that a bad side effect will result from an action together with the performance of that action disregarding the consequences gives enough information about the agent's attitude: the achievement of the main goal is more important for her than the avoidance of the bad side effect. This conclusion cannot be made in help scenarios, because there is no conflict between the main goal and the side effect such that it is intersubjectively agreed upon that the action should be omitted for the sake of the side effect not to occur. The occurrence of a good side effect is as such never morally problematic as a bad side effect normally is. It is permissible to produce good side effects. However, there are intersubjective expectations also in help scenarios that concern the agent's attitude and motivation. A proper target of praise is required to (at least) meet these expectations. For example, the chairman who prefers making profit over environmental issues does not act on expected preferences even if the side effect turns out to be a good one. But in contrast to harm

scenarios, it is not the case in help scenarios that the motive can be deduced from the performance of the action together with the agent's knowledge about the side effect. In help scenarios further information is needed in order to judge whether the agent is a proper target of praise or should rather be expraised. The verbal expression of the agent's attitude provides this sort of missing information in the helping scenarios, while it is redundant or overwritten in harm scenarios. While the information that the agent knows about the bad side effect in harm scenarios is sufficient for judging her morally when she performs the action in question, the verbal expression of the agent's attitude is required in the help scenario for that judgment. In cases where more information is required the verbal expression determines whether subjects form a *positive* reactive attitude (praise) or a *negative* reactive attitude (expraise). That explains why manipulation of the verbal attitude has an effect in help scenarios but not a significant one in harm scenarios.

4 Why there *seems* to be an asymmetry

In the previous sections I argued against the common view that the effect Knobe and following experimenters found represents an asymmetry in survey subjects' judgments. In this section I will explain why the results can only be interpreted as an asymmetry when we try to explain the results within *one* concept, e.g. the concept of moral responsibility or the concept of intentional action. I will focus on the concept of moral responsibility here, because I think that the asymmetric ascriptions of intentionality is a pragmatic means by which subjects strengthen their moral judgments. There is no asymmetry when we refer to positive and negative reactive attitudes and distinguish between praise and excuse, on the one hand, and blame and expraise on the other. When the judgment 'A is blameworthy' is replaced by the allegedly synonymous judgment 'A is morally responsible', and 'A is not praiseworthy' is replaced by 'A is not morally responsible', then the asymmetry arises: though an agent who is blamed or expraised is judged negatively in both cases, she is seen as morally responsible in one case but *not* morally responsible in the other. The same is true for positive judgments: an agent who is praised is seen as morally responsible but not when she is excused for having done so. Further, an agent who is judged to be morally blameworthy is, according to that synonymous treatment, not distinguishable from an agent who is judged to be praiseworthy if we do not refer to the goodness or badness of the relevant side effect. In both cases, the agent is morally responsible for her action and the resulting side effects. Accordingly, an agent who is expraised is not distinguishable from an agent who is excused, because both are seen as not morally responsible. But praising and blaming are completely different reactive attitudes as well as excusing and expraising are: while blaming and expraising are negative judgments about an agent, praising

and excusing are positive judgments. That means that the goodness or the badness of an action or side effect does not determine whether the judgment is positive or negative, but only which kind of a positive or negative judgment is to be made.

It is the translation of expraise and excuse into 'not morally responsible' and blame and praise into 'morally responsible' that makes an asymmetry appear. Even if most interpreters do not explicitly use moral responsibility as a synonym for either praise or blame, they treat the notions 'blameworthy' and 'not praiseworthy' as if there is one and the same underlying concept – one time in the affirmative and one time in the negative. Implicitly it is referred to the concept of moral responsibility that is assumed to unite praise and blame. However, this is no reason to arrive at any conclusions about an asymmetry within the concept of moral responsibility; we rather should conclude that the concept of moral responsibility is not suitable for uniting judgments like praise and excuse on the positive side, and blame and expraise on the negative side. Indeed, we should stop looking for general conditions of moral responsibility as such, and rather provide separate accounts for positive and negative moral attitudes. With the replacement of '(not) blameworthy' by '(not) responsible' and the replacement of '(not) praiseworthy' by '(not) responsible' the respective judgment's context becomes lost.

5 Conclusion

Similar to other contributors of the debate, I have argued for interpreting ascriptions of intentionality as moral judgments about the agent and her action; but this was not my main point. Though people are misled by the experimental setting in what question to answer and wrongly try to emphasize their moral judgment by ascription or non-ascription of intentionality, they are doing the right thing when they judge the not caring harming chairman to be blameworthy (via an ascription of intentionality) and the not caring helping chairman to be not praiseworthy (via a false ascription of non-intentionality). Knobe claims that it is the badness of the side effect as such that makes subjects generally more inclined to ascribe intentionality, while the goodness of the side effect makes subjects refrain from doing so. This could be transferred to moral judgment: while the badness of the side effect makes subjects generally more inclined to blame an agent, the goodness of the side effect makes subjects refrain from doing so. But Knobe's interpretation is flawed. As it can be seen in Shepherd's extended study, there are also reversed ascriptions of intentionality towards agents who perform actions with *good* side effects and non-ascriptions of intentionality towards agents who perform actions with *bad* side effects. Knobe's original scenarios of a helping and a harming chairman are restricted to *negative* judgments only: the chairman of the harm and the one of the help scenario act on unacceptable preferences that do not match intersubjective expectations, because both do not

care about the environment and both prefer to make profit. Whether a moral judgment is negative or positive is not determined by the goodness or badness of the side effect; it is the agent's *preference structure* that is decisive for this. However, the goodness or badness of the side effect determines whether the judgment is, if negative, a judgment of blame or expraise and, correspondingly, whether the judgment is, if positive, a judgment of praise or excuse. That means that there is no asymmetry between the negative attitudes of blame and expraise, on the one hand, and none between the positive attitudes of praise and excuse on the other. An asymmetry only arises between negative and positive attitudes, e.g. between praise as a positive attitude and expraise as a negative; and the same between blame as a negative attitude and excuse as a positive. An asymmetry between ascriptions of blame and expraise only arises if the two are read as 'morally responsible' and '*not* morally responsible'. When blame and expraise are read in this way, it cannot be seen that both of them are negative judgments and perfectly symmetric.

In this paper, I did not analyse the concept of moral judgment sufficiently but only tried to make sense of the experimental data provided so far. I am convinced that the interpretation of such data can be useful for philosophical inquiry concerning the concepts of praise and excuse as well as blame and expraise if we are interested in what people are *doing* when forming moral judgments. Lastly, the data clearly supports my claim that moral responsibility is not a concept under which praise and blame can be equally subsumed.

References

F. Adams and A. Steadman (2004a) 'Intentional action in ordinary language: Core concept or pragmatic understanding?', *Analysis*, 64, 173–181.

F. Adams and A. Steadman (2004b) 'Intentional action and moral considerations: Still pragmatic', *Analysis*, 64, 268–276.

M. Phelan and H. Sarkissian (2008) 'The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it', *Philosophical Studies*, 138, 291–298.

J. Knobe (2003) 'Intentional action and side effects in ordinary language', *Analysis*, 63, 190–193.

J. Knobe (2006) 'The concept of intentional action: a case study in the uses of folk psychology', *Philosophical Studies*, 130, 203–231.

D. Pettit and J. Knobe (2009) 'The Pervasive Impact of Moral Judgment', *Mind and Language*, 24, 586–604.