DOI: 10.1111/mila.12197

WILEY

ORIGINAL ARTICLE

# What is the cognitive basis of the side-effect effect? An experimental test of competing theories

Marina Proft | Alexander Dieball | Hannes Rakoczy

Department of Developmental Psychology, Georg-August-Universität Göttingen, Göttingen, Germany

**Correspondence**
Marina Proft, Georg-August-Universität Göttingen, Waldweg 26, 37073 Göttingen, Germany.
Email: marina.proft@psych.uni-goettingen.de

Recent work on the *side-effect effect* has shown that subjects' intentionality judgments are influenced by moral evaluations. In six experiments, we tested four different candidates for the cognitive foundation derived from prominent explanatory accounts (prescriptiveness, [un-] expectedness, blame and a shift in default attitudes) against each other in three steps. First, Study 1 showed that the effect even extends to certain descriptive norms. Second, Studies 2–5 investigated the candidates more directly. Results reveal that intentionality judgments could best be explained by underlying shifts in default attitudes. Third, Study 6 experimentally manipulated this default attitude, leading to the predicted change in intentionality judgments.

**KEYWORDS**

intentionality, moral judgment, norms, side-effect effect, social cognition, theory of mind

## 1 | INTRODUCTION

Common sense and empirical research have long known that our moral judgments are deeply influenced by our intentionality ascriptions: unintentionally harming someone, for example, is generally seen as less bad and less blameworthy than intentionally doing so (see, e.g., Cushman, 2008; Guglielmo, Monroe & Malle, 2009). A new line of research, however, has shown that there is also influence the other way round. Surprising, and on first glance counter-intuitive, the so called *side-effect effect* (SEE) demonstrates that moral judgments themselves can influence our non-moral cognition, such as intentionality ratings (e.g., Knobe, 2003a, 2003b, 2005). When told a story about a person who intentionally performs a neutral action which has a foreseen side-effect that is either morally good or bad, subjects tend to say that the person intentionally brought about the negative but not the positive side-effect. In the original vignette (Knobe, 2003a), a chairman plans to increase profits (the intended main effect) by starting a new program. When informed by his vice president that the program would also harm/help the environment (the foreseen side-effect), the chairman replies that he does not care about what happens to the environment but only aims at increasing profits and thus decides to implement the new program. Consequently, the environment is harmed/helped. When asked to judge "Did the chairman harm/help the environment intentionally?," people tend to give higher intentionality ratings in the harm than in the help situation. This effect has been widely replicated with various kinds of material (see Knobe, 2010, for an overview), even across different cultures (Knobe & Burra, 2006) and in young children (Leslie, Knobe & Cohen, 2006; Michelin, Pellizzoni, Tallandini & Siegal, 2009; Pellizzoni, Siegal & Surian, 2009).

What remains largely unclear, however, is what cognitive foundation the effect builds on. Which cognitive processes underlie the different representations of the help and the harm scenarios that manifest themselves in the different intentionality judgments? A closely related question is: how content-specific is the SEE? Is it specifically related to morally relevant side-effects and thus an effect to do with moral cognition more narrowly? Or might the effect extend to side-effects that relate to other types of prescriptive or perhaps even descriptive norms and thus turn out to be an effect of a more general form of cognition? The crucial question thus is the following: What is the cognitive difference between the help and harm scenarios tested so far that underlies the diverging representations of the two cases? Existing data are compatible with a number of different options and leave open a number of different accounts concerning the cognitive basis and content-specificity of the SEE.

Rationality accounts (e.g., Holton, 2010; Uttich & Lombrozo, 2010) assume that, despite appearances to the contrary, upon closer inspection the effect itself is not surprising or counter-intuitive at all. On the contrary, it originates from rational strategies of ascribing mental states. According to one variant of rationality accounts, the effect arises from an underlying asymmetry concerning the role of knowledge when ascribing mental states (Holton, 2010). In our mental state attribution, the rational strategy is to consider whether a given action intentionally violates or intentionally conforms to a norm, and this is mediated by the agent's knowledge: to intentionally violate a norm, it is sufficient to knowingly violate it. Knowingly violating a norm involves an intention to disregard the given norm and is thus seen as an intentional norm violation. In contrast, to intentionally conform to a norm, knowingly following the norm is not sufficient. For a norm-conforming behavior to be perceived as intentional, the agent has to be counterfactually guided by the norm—that is, she has to actively adapt her behavior so that it corresponds to the requirements of the norm.

Relating to the original vignette, the criteria for the intentional norm-violation are fulfilled in the harm scenario, while the criteria for intentional norm-conformity are not fulfilled in the help scenario (the chairman does not care about the side-effect). Concerning the extension of the SEE, Holton's analysis of the critical difference between help and harm scenarios (knowingly acting in accordance vs. in violation of a norm) applies to all kinds of norms that are potentially action-guiding; that is, to all kinds of prescriptive norms (that prescribe what ought to be done).

According to a second variant of rationality accounts, the "rational scientist view," the crucial difference between help and harm cases lies in their differential diagnostic validity (Uttich & Lombrozo, 2010). The general idea is that, when confronted with SEE scenarios, subjects attempt to make rational inferences on the basis of the agent's behavior about her motivational set and the intentions underlying her actions. Generally, behavior that is in accordance with a default is less informative than behavior that deviates from a default. One specific case of this general principle is the case of norm-related behavior. Norm-conforming behavior is uninformative in terms of the agent's internal motivation: she might have done what she was supposed to do or what everyone else did, without a specific motivation of her own for this kind of action. Norm-violating behavior, in contrast, is generally more diagnostic for intentionality attribution since it implies mental states on the part of the agent that are sufficiently strong to make her act against the defaults of the norm. Consequently, the chairman's behavior is judged as intentional in the norm-violating but not in the norm-conforming situation. Crucially, this analysis of the difference in diagnostic quality concerning norm-conforming versus norm-violating behavior for mental state ascription should not be confined to prescriptive norms, but also applies to descriptive norms—norms that describe what is generally the case: “A behavior that violates a statistical norm is not ‘expected’, and hence provides information about the agent's underlying mental states” (Uttich & Lombrozo, 2010, p. 99). For example, imagine several gardens full of yellow flowers. Seeing a gardener in one of the yards planting another yellow flower, no one would seriously wonder, “Why does he plant a yellow flower now?” or even infer “He really seems to like yellow flowers!” In contrast, seeing him planting a blue flower might well be surprising and unexpected (and the question “Why is he planting a blue flower now?” or the inference “This gardener really wanted to plant a blue flower!” might be perfectly appropriate). In sum, according to the rational scientist view, the SEE is a consequence of the differential diagnostic validity (often manifested in the differential expectedness) of norm-violating compared to norm-conforming behavior, and should thus emerge in the case of prescriptive and descriptive norms alike. Concerning the relation of the two rationality accounts (Holton's and the rational scientist view) to each other, what differs between them are their diverging predictions regarding the extendibility of the effect to prescriptive versus descriptive norms.

Morality accounts (Alicke, 2000; Knobe, 2010) emphasize the role of specifically *moral* considerations and their influence on intentionality judgments. According to one morality account, the “culpable control” model (Alicke, 2000), morally wrong actions elicit spontaneous negative reactions, leading to a blame attribution towards the agent. In order to validate this blame judgment, the intentional role of the agent is then emphasized post-hoc. Thus, morally wrong actions are more likely to be judged as intentional than morally right actions, since the latter usually do not elicit negative reactions in need of being justified. This account in terms of a general bias in blame attribution is compatible with existing findings concerning the original and related SEE vignettes (Alicke & Rose, 2010; Alicke, Rose & Bloom, 2011). Regarding content-specificity, the account would predict that the SEE is confined to contrast cases that differ in the elicitation of spontaneous blame and related negative reactions towards the agent. Note that this does not imply that the effect should be restricted to moral norms per se. It has been amply documented that also non-moral, for example, conventional or disgust norms, can elicit negative reactions such as blame in participants (e.g., Nichols, 2002; Smetana et al., 2012).

A second variant of morality accounts, the “default-shift account,” views the relations of moral and intentionality ascription as even more intimate (Knobe, 2010; see also Pettit & Knobe, 2009). According to this account, moral considerations themselves are integrated in our intentionality evaluations. That is, they do not only influence one another but the former are part of the latter. Intentionality ascriptions are seen as an attitude dimension, a continuum between the highly positive and highly negative value of the attitude intentionality. To ascribe an intentional attitude to a person one has to view this attitude on the continuum in relation to a default (i.e., what people normally think about the behavior). If the intentional attitude being judged is located on the positive side of the default attitude then the action is seen as intentional. Take, for instance, the case of growing yellow flowers in a garden. If people normally have a neutral attitude towards having yellow flowers, but the agent is really excited about it, people will judge his actions of growing yellow flowers as intentional (as “being excited” is more positive than “neutral”). If an action is located on the negative side of the default attitude, however, this action is judged as unintentional.

What differs concerning moral and immoral behavior is the location of the default attitude regarding the behavior. The default attitude of a person towards moral behavior should be slightly positive (one should be slightly pro-environment even if not a strong environmentalist) while the default attitude towards unmoral behavior should be slightly negative. Having a *neutral* attitude in either scenario, the chairman thus meets the criteria for intentionality (being on the pro-side in relation to the default attitude, as "neutral" is more positive than "slightly negative") in the harm-scenario but not in the help-scenario (here the default is slightly positive, hence a neutral attitude is more negative than the default). The critical difference between help and harm cases is thus the default attitude between moral and immoral behavior. Though this account has so far been primarily applied to prescriptive, in particular moral norms (Knobe, 2010), the core idea of the default shift does not need to be limited to these cases. In fact, there might be such default shifts for any kind of descriptive or prescriptive norm. For example, one might expect everyone to be slightly pro "greeting if you enter a room with other people," and slightly negative towards not doing so. A fundamental problem, however, is that a priori it is not clear exactly in which cases the model would predict such default shifts as in the help/harm contrast. However, the account does make very clear conditional predictions: *if* a differential default shift towards pro-attitudes for norm-conforming and con-attitudes for norm-violating behavior happens, *then* there should be a corresponding asymmetry in intentionality judgments. And these conditional predictions can straightforwardly be tested by independently exploring both intentionality judgments and default attitudes and then probing their relations.

So, which of these four different accounts best explains the SEE? Since existing data remain inconclusive, the rationale of the present study was to systematically test between the competing accounts. To this aim, we administered a novel approach, where the new methodology is a stepwise approximation of the following structure. In a first step, we investigated the generality and the limits of the SEE: for what kinds of norms do the differences in intentionality ratings (not) arise? In a second step, we explored in which respects the norms that elicit the effect differ from those that do not elicit the effect, using as anchors the proposed candidates from the four theories described earlier. The logic behind this is the following: if a candidate serves as a potential cognitive foundation of the difference in intentionality ratings, then the pattern for this candidate concerning the different types of norms should mirror that of intentionality ratings. In a third step, we directly tested the causal influence of the one candidate that stood out in the second step.

Adapted to the specific research aim, the approach was implemented in the following way. First, we contrasted different kinds of prescriptive and descriptive norms with regard to their elicitation of the SEE (Study 1). Results indicated that the effect is not restricted to prescriptive, but can be extended to certain kinds of descriptive norms. Second, in Studies 2–5 we investigated the four different candidates for the crucial cognitive difference between help/harm cases, suggested by the two rationality and the two morality accounts under study: prescriptiveness (Holton, 2010), (un-)expectedness (Uttich & Lombrozo, 2010), blame attribution (Alicke & Rose, 2010) and default shift (Knobe, 2010), and how they relate to the pattern of results found in Study 1. Results emphasized the default shift as the most suitable candidate (out of the four tested) to explain intentionality judgments. Third, in Study 6 we then directly tested whether an experimental manipulation of the default attitude affected intentionality judgments in the predicted ways.

## 2 | GENERAL METHODS FOR STUDIES 1–5

The first five experiments all followed the structure of a 2 (valence of side-effect: norm-conforming/norm-violating) × 4 (type of norm) between subjects design, presenting each subject with a vignette that was adapted from the typical SEE vignette to the requests of the current study. All participants were recruited on the campus of the University of Göttingen, were instructed to read the vignette carefully, and received a chocolate bar for their participation.

The vignettes all basically followed the same structure (for slight changes see the method sections of the respective studies): Snoj, who lives on planet Zeik, wants to bring about a main effect (growing his tree as high as possible). As a means to this goal he goes to a shop to buy a fertilizer for his tree. The shop owner tells Snoj about the side-effect that the fertilizer will have on the (kind of flowers) growing next to his tree. Snoj replies that he does not care about the side-effect but only about the main effect and thus buys the fertilizer and uses it for his tree. As a consequence, the side-effect occurs as well.

What varied across conditions was (a) the valence of the side-effect as norm-violating or norm-conforming and (b) the relevant norm. From a theoretical point of view, the main norm distinction was between prescriptive and descriptive norms. Prescriptive norms prescribe what to do, whereas descriptive or statistical norms describe what is "normal" or what people "typically" do (in the statistical sense). However, statistical norms of human social behavior often acquire prescriptive force over time (e.g., Lewis, 1969). Consequently, statistical norms, even if actually describing what people usually do, might be (mis)interpreted in a mixed descriptive-prescriptive way: as describing what people usually do, with the implication that members of the community in question ought to conform to this regularity. In order to address this potential ambiguity directly, we introduced two types of descriptive norms: statistical norms that might plausibly be understood in these mixed descriptive-prescriptive terms (norms describing what people normally do, in the following called statistical norms with social-conformity) and those for which such a prescriptive interpretation seems unlikely as they refer to regularities given by the environment (in the following, called statistical norms without social-conformity).

All in all, we thus tested four different types of norms: moral, social–conventional, statistical with social-conformity, and statistical without social-conformity. In the *moral* case, the fertilizer either destroyed someone else's flowers or led to their growth. In the *social–conventional* case, a rule on planet Zeik stated that people were only allowed to have flowers that are yellow in their gardens; the fertilizer either led to blue or yellow flowers. Regarding the *statistical norm with social-conformity,* it was described that all the other people on planet Zeik only grow yellow flowers in their gardens. Again, the fertilizer either led to blue or yellow flowers. In the *statistical without social-conformity* case, it was described that all flowers on planet Zeik are yellow. Again, the fertilizer either led to blue or yellow flowers. See Table 1 for a systematic structure of the vignettes.

## 3 | STEP 1: GENERALITY OF EFFECT

### 3.1 | Study 1: Intentionality

The aim of Experiment 1 was to investigate whether the typical side-effect effect, that is, higher intentionality ratings for norm-violating than for norm-conforming behavior, could be extended to descriptive norms. The distinction between prescriptive and descriptive norms reflects the key difference between the theories of Holton (2010) and the rational scientist view (Uttich & Lombrozo, 2010). While the latter predicts the typical divergence also for the case of statistical norms, the former predicts the effect only to occur in cases of prescriptive norms.

**TABLE 1** Systematic structure of the vignettes

| Moral | Social–conventional | Statistical with social-conformity | Statistical without social-conformity |
|---|---|---|---|
| Snoj lives on planet Zeik. | | | |
| Snoj has a tree. His tree grows on a lawn among many flowers that belong to Hork, another inhabitant of Zeik. Hork really loves his flowers and wants to have as many as possible. | Many people on planet Zeik have flowers. On planet Zeik there is the rule that inhabitants are only allowed to have flowers that are yellow in their gardens. Snoj has a tree. | Many people on planet Zeik have flowers. People on planet Zeik only grow yellow flowers in their gardens. Snoj has a tree. His tree grows in a garden surrounded by other gardens. | On planet Zeik there are many flowers. All flowers on planet Zeik are yellow. Snoj has a tree. His tree grows on a lawn surrounded by yellow flowers. |
| Snoj wants his tree to grow as high as possible. For his tree growing as high as possible, he needs to fertilize his tree. To fertilize his tree, Snoj needs fertilizer. He goes to a shop to buy fertilizer. He says to the salesman: "I want to have fertilizer to fertilize my tree." The salesman answers: "Here I have fertilizer which you can use to fertilize your tree. | | | |
| However, if you do so, Hork's flowers will die [more of Hork's flowers will grow] next to your tree." Snoj answers: "I don't care if Hork's flowers die [more of Hork's flowers are going to grow] next to my tree. I just want my tree to grow as high as possible." Snoj buys the fertilizer and goes to his tree. He fertilizes his tree with the fertilizer and in fact Hork's flowers die [more of Hork's flowers grow] next to his tree. | However, if you do so, blue [yellow] flowers will grow next to your tree." Snoj answers: "I don't care if blue [yellow] flowers will grow next to my tree. I just want my tree to grow as high as possible." Snoj buys the fertilizer and goes to his tree. He fertilizes his tree with the fertilizer and in fact blue [yellow] flowers grow next to the tree. | | |

### 3.1.1 | Methods

A total of 320 participants ($N$ = 80 per type of norm) each received one of the above described versions of the SEE vignette. They were then asked to rate how much they agreed with the statement "Snoj brought about the side-effect intentionally"[1] on a rating scale of 1–7 with 1 being "no agreement at all" to 7 "total agreement", with 4 being "unsure."

### 3.1.2 | Results and discussion

Participants' ratings of how much they agreed with the statement that Snoj brought about the side-effect intentionally are displayed in Figure 1. The 2 (valence of side-effect) × 4 (type of norm) ANOVA revealed main effects for both valence $F(1,312) = 11.63$, $p < .001$, $\eta_p^2 = .04$, and type of norm $F(3,312) = 2.68$, $p = .047$, $\eta_p^2 = .03$. The interaction between the two factors almost reached significance $F(3,312) = 2.58$, $p = .054$, $\eta_p^2 = .02$. Planned $t$-tests for the different kinds of norms showed that people rated a norm-violating side-effect as more intentional than a norm-conforming one for moral (norm-violating: $M = 4.88$, $SD = 1.95$, norm-conforming: $M = 3.45$, $SD = 2.29$; $t(78) = 3.00$, $p < .01$, $r = .32$), social–conventional (norm-violating: $M = 4.23$, $SD = 2.09$, norm-conforming: $M = 3.23$, $SD = 2.12$; $t(78) = 2.12$, $p = .04$, $r = .23$) and statistical norms with social-conformity (norm-violating: $M = 4.6$, $SD = 2.07$, norm-conforming: $M = 3.5$, $SD = 2.04$;

[1] Note that the notions "side-effect" as well as "norm-conforming/norm-violating behavior" serve as generic placeholders here. The test questions always included the respective actual wording (e.g. "Snoj intentionally grew yellow flowers next to his tree.")

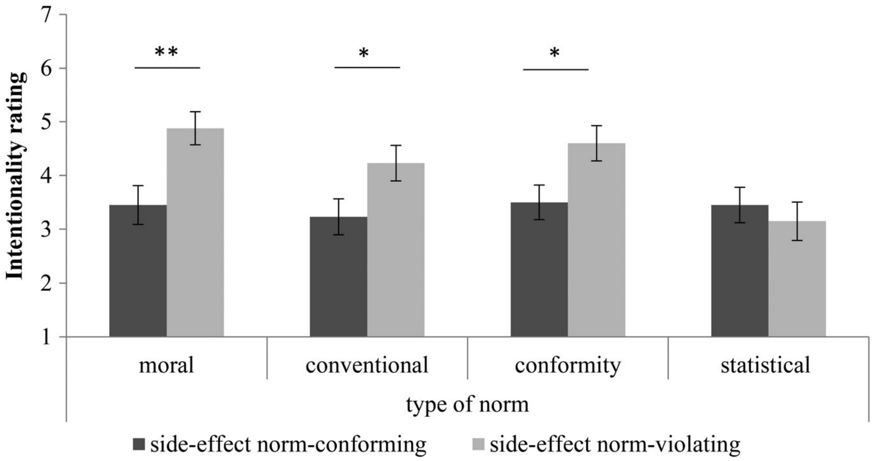

**FIGURE 1** Ratings of how much the person agrees with the following sentence: "Snoj brought about the side-effect intentionally" on a scale from 1 (not at all) to 7 (totally). *Note.* Error bars display standard errors, $^*p < .05$, $^{**}p < .01$

$t(78) = 2.39$, $p = .02$, $r = .26$) but not for statistical norms without social-conformity (norm-violating: $M = 3.15$, $SD = 2.25$, norm-conforming: $M = 3.45$, $SD = 2.09$; $t(78) = .62$, $p = .54$, $r = .07$, $BF = 4.91$ [Bayes factor for null over alternative]).[2] From this data, the SEE seems to be extendable to descriptive norms if they are confounded with social-conformity in the form that everyone shows a certain behavior, in this case only planting yellow flowers in their gardens.

While these results give some information about the relevant factors influencing the intentionality bias due to the differences in the kind of norms, this information is indirect since we do not know whether the manipulation of the kind of the norm was in fact the crucial difference for our participants to give their respective intentionality ratings. Hence, as a second step, we directly asked participants to give judgments regarding the potential critical difference between norm-conforming/norm-violating situations according to the theories described above. The logic behind this method is the following: if a proposed factor (prescriptiveness, (un-)expectedness, blame attribution and default shift) can account for the difference between the norm-conforming and norm-violating cases, subjects in their direct ratings of the factors should display an equivalent pattern of results as in their intentionality judgments from Study 1. In Studies 2 to 5, we therefore tested for the four candidates suggested by the four theories: prescriptiveness (Studies 2a and 2b), (un-)expectedness (Study 3), blame attribution (Study 4), and default shift (Study 5).

## 4 | STEP 2: TEST OF CANDIDATES FOR COGNITIVE FOUNDATION

### 4.1 | Study 2a: Prescriptiveness

The aim of Study 2a was to investigate whether prescriptiveness ratings can account for the differences in intentionality ratings from Study 1. The idea was that if participants see a norm as prescriptive, they should say that the agent ought to buy the fertilizer that comes along with a norm-conforming side-effect but ought not to buy the one that comes along with a norm-violating

[2] Bayes factors were computed with version 0.9.8 of the BayesFactor package (R version 3.3.2 [October 31, 2016]) on i386-redhat-linux-gnu. All Bayes factors are based on the JZS prior, that is, Cauchy prior on effect size and Jeffreys prior on variance (see Rouder, Speckman, Sun, Morey & Iverson, 2009).

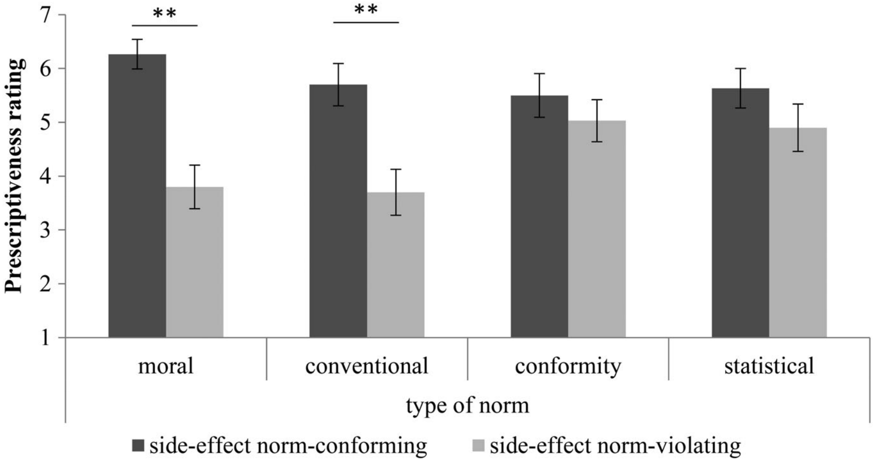

**FIGURE 2** Ratings of how much the person agrees with the following sentence: "Snoj ought to buy the fertilizer to fertilize his tree" on a scale from 1 (not at all) to 7 (totally). *Note:* Error bars display standard errors, $*p < .05$, $**p < .01$

side-effect. By contrast, a descriptive norm should not lead to that difference. As a consequence, if Holton's (2010) view is right and prescriptiveness is a necessary requisite for the SEE, then the prescriptiveness ratings between norm-conforming and norm-violating side-effects should only differ for the moral, social–conventional and statistical with social-conformity vignettes but not for statistical without social-conformity, since only the former three elicited differences in intentionality ratings in Study 1.

### 4.1.1 | Methods

A total of 240 participants ($N = 60$ per type of norm) each received a slightly modified version of the vignettes described in the general methods section: the vignette ended after the shop owner told Snoj about the side-effect of the fertilizer. Participants were asked to rate how much they agreed with the statement "Snoj ought to buy the fertilizer to fertilize his tree"[3] on a rating scale of 1–7 with 1 being "no agreement at all" and 7 "total agreement", with 4 being "unsure."

### 4.1.2 | Results and discussion

Participants' ratings of how much they agreed with the statement that Snoj ought to buy the fertilizer to fertilize his tree are displayed in Figure 2. The 2 (valence of side-effect) × 4 (type of norm) ANOVA revealed a main effect for valence $F(1,232) = 26.27$, $p < .001$, $\eta_p^2 = .10$, no main effect for type of norm $F(3,232) = .94$, $p = .422$, $\eta_p^2 = .01$, but an interaction between the two factors $F(3,232) = 3.07$, $p = .029$, $\eta_p^2 = .04$. Planned *t*-tests for the different kinds of norms show that people rated that Snoj ought to buy the fertilizer that comes along with a norm-conforming, but not the one that comes along with a norm-violating, side-effect for the moral (norm-violating: $M = 3.8$, $SD = 2.22$, norm-conforming: $M = 6.27$, $SD = 1.51$; $t(58) = 5.04$, $p < .01$, $r = .55$) and social–conventional norm (norm-violating: $M = 3.7$, $SD = 2.34$, norm-conforming: $M = 5.7$, $SD = 2.15$; $t(58) = 3.45$, $p < .01$, $r = .41$). However, they did not make this difference for the two statistical norms (with social-conformity: norm-violating: $M = 5.03$, $SD = 2.14$, norm-conforming: $M = 5.5$, $SD = 2.22$; $t(58) = 0.83$, $p = .41$, $r = .11$, $BF = 3.77$ [Bayes factor for null over

[3] Note, that the original wording was the German word "sollte" which was unambiguously used in a prescriptive sense.

alternative]; without social-conformity: norm-violating: $M = 4.9$, $SD = 2.41$, norm-conforming: $M = 5.63$, $SD = 2.01$; $t(58) = 1.28$, $p = .21$, $r = .17$, $BF = 2.46$ [Bayes factor for null over alternative]). These results suggest that participants understood the moral and the social–conventional norm as a prescriptive rule, while this was not the case for the two statistical norms.

However, it might be that the frames of reference used in Study 1 and Study 2a were different. So it might be that, while participants themselves did not see the conformity norm as prescriptive and the statistical one as descriptive (as would be in line with Holton's explanation of the data), they assume that such a pattern mirrors the normative structure on planet Zeik. To rule out this possibility, in Study 2b we explicitly asked participants to judge whether other people on planet Zeik would say that Snoj ought to buy the fertilizer.

## 4.2 | Study 2b: Prescriptiveness other

### 4.2.1 | Methods

A total of 240 participants ($N = 60$ per type of norm) each received one of the vignettes used in Study 2a. Only the test question slightly differed from Study 2a. Participants were asked to rate how much they agreed with the statement "The other people on planet Zeik would say that Snoj ought to buy the fertilizer to fertilize his tree" on a rating scale of 1–7 with 1 being "no agreement at all" to 7 "total agreement", with 4 being "unsure."

### 4.2.2 | Results and discussion

Participants' prescriptiveness ratings are displayed in Figure 3. The 2 (valence of side-effect) × 4 (type of norm) ANOVA revealed a main effect for valence $F(1,232) = 78.11$, $p < .001$, $\eta_p^2 = .25$, a main effect for type of norm $F(3,232) = 4.70$, $p = .003$, $\eta_p^2 = .06$, but no interaction between the two factors $F(3,232) = 2.12$, $p = .098$, $\eta_p^2 = .03$. Planned $t$-tests revealed that for all the four types of norms participants said that they agreed more with the statement that other people on planet Zeik would say that he ought to buy the fertilizer that comes along with the norm-conforming than with the norm-violating side-effect (moral: norm-violating: $M = 3.67$, $SD = 1.84$, norm-conforming:

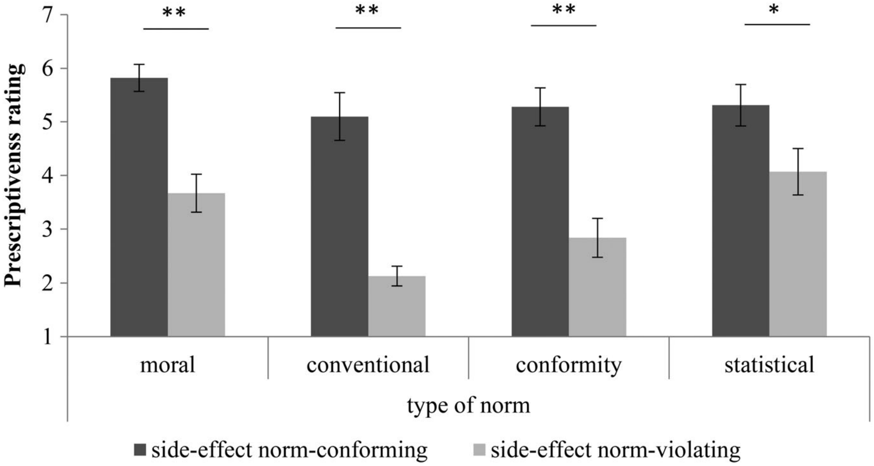

**FIGURE 3** Ratings of how much the person agrees with the following sentence: "The other people on planet Zeik would say that Snoj ought to buy the fertilizer to fertilize his tree" on a scale of 1 (not at all) to 7 (totally). *Note:* Error bars display standard errors, *$p < .05$, **$p < .01$

$M = 5.82$, $SD = 1.45$; $t(58) = 5.07$, $p < .01$, $r = .55$, social–conventional: norm-violating: $M = 2.13$, $SD = 1.02$, norm-conforming: $M = 5.1$, $SD = 2.40$; $t(58) = 6.33$, $p < .01$, $r = .64$, statistical with social-conformity norm-violating: $M = 2.84$, $SD = 2.00$, norm-conforming: $M = 5.28$, $SD = 1.91$; $t(58) = 4.82$, $p < .01$, $r = .53$ and statistical without social-conformity: norm-violating: $M = 4.07$, $SD = 2.29$, norm-conforming: $M = 5.31$, $SD = 2.19$; $t(58) = 2.14$, $p = .04$, $r = .27$).

Putting the results form Studies 2a and 2b together, the following pattern emerges. Participants in general did not interpret the two statistical norms (with and without social-conformity) as prescriptive. However, when asked to take as the frame of reference an inhabitant of planet Zeik, they indicated both statistical norms as prescriptive. As neither of these patterns matches the one of the intentionality rating from Study 1, prescriptiveness of the norm does not seem to be the (only) source for the differences in intentionality ratings. Additionally, as without any further prompts to switch perspectives participants rated the statistical norm with social-conformity as descriptive, the conclusion arises that the side-effect effect is in fact extendable to at least this particular case of descriptive norms.

## 4.3 | Study 3: (Un-)expectedness

Study 3 aimed to investigate the influences on (un-)expectedness of actions on intentionality judgments, as suggested by the rational scientist view (Uttich & Lombrozo, 2010). According to this theory, unexpected behavior, that is, behavior that deviates from a norm, is an important indicator for the revision of the baseline mental states, leading to an ascription of intentionality. Hence, the norm-violating actions that were judged as being intentional in Study 1 should also be judged as unexpected, while the norm-violating behavior in the statistical norm case without social conformity should not be judged as unexpected.

### 4.3.1 | Methods

A total of 280 participants ($N = 70$ per type of norm) each received one of the full vignettes as presented in the general methods section. They were then asked for their agreement with the test question: "That Snoj let the side-effect happen was surprising." On a scale of 1–7 with 1 being "no agreement at all" and 7 "total agreement", with 4 being "unsure."

### 4.3.2 | Results and discussion

Figure 4 shows participants' ratings for each of the norms. The 2 (valence of side-effect) × 4 (type of norm) ANOVA revealed a main effect for valence $F(1,272) = 20.65$, $p < .001$, $\eta_p^2 = .07$, but no main effect for type of norm $F(3,272) = 1.04$, $p = .373$, $\eta_p^2 = .01$, and no interaction between the two factors $F(3,272) = .33$, $p = .80$, $\eta_p^2 = .01$. Planned $t$-tests revealed that participants judged the norm-violating behavior as being more unexpected than the norm-conforming one for the social–conventional (norm-violating: $M = 3.09$, $SD = 2.01$, norm-conforming: $M = 2.06$, $SD = 1.64$; $t(68) = 2.35$, $p = .02$, $r = .27$) and the statistical norm without social-conformity (norm-violating: $M = 2.94$, $SD = 1.78$, norm-conforming: $M = 1.74$, $SD = 1.01$; $t(68) = 3.47$, $p < .01$, $r = .39$). A trend for this difference was observed for the other two norms (moral: norm-violating: $M = 3.11$, $SD = 2.00$, norm-conforming: $M = 2.29$, $SD = 1.76$; $t(68) = 1.84$, $p = .07$, $r = .22$, statistical with social-conformity: norm-violating: $M = 2.57$, $SD = 1.87$, norm-conforming: $M = 1.91$, $SD = 1.38$; $t(68) = 1.67$, $p = .10$, $r = .20$).

This pattern of results does not represent the pattern of intentionality ratings from Study 1, which suggests that at least for our scenarios, unexpectedness of behavior was not the (only) indicator for

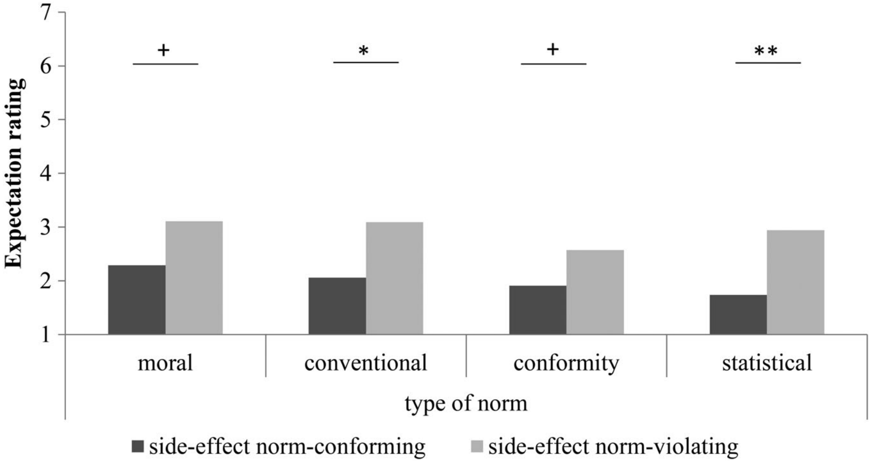

**FIGURE 4** Ratings of how much the person agrees with the following sentence: “That Snoj let the side effect happen was surprising.” on a scale from 1 (not at all) to 7 (totally). *Note.* Error bars display standard errors, $^{+}p < .10$, $*p < .05$, $**p < .01$

intentionality. Note also, that none of the means was above 4, indicating that though participants judged the norm-violating behavior as less expected, they still didn't judge them as unexpected.

## 4.4 | Study 4: Blame attribution

According to the culpable control model (Alicke, 2000; Alicke & Rose, 2010) the differences between intentionality judgments for norm-conforming versus norm-violating side-effects derive from the fact that in the norm-violating situation people tend to blame the agent for his action. To justify this blame judgment, the intentional role of the agent is enhanced. In the norm-conforming situation the agent is not blamed and therefore there is no need to ascribe him intentionality for the side-effect. In Study 4, we tested whether the pattern of subjects' blame judgments mirrors the pattern of intentionality ratings found in Study 1, such that participants give higher blame ratings for the norm-violating than the norm-conforming action for the moral, social–conventional and statistical norms with social-conformity but not for the statistical norms without social-conformity.

### 4.4.1 | Methods

A total of 240 participants ($N = 60$ per type of norm) each received one of the full vignettes as presented in the general methods section. They were asked to rate how much they agreed with the test question “The side-effect occurred. One can blame Snoj for that.” On a scale of 1–7 from 1 being “no agreement at all” to 7 “total agreement”, with 4 being “unsure.”

### 4.4.2 | Results and discussion

The 2 (valence of side-effect) × 4 (type of norm) ANOVA revealed a main effect for valence $F(1,232) = 12.68$, $p < .001$, $\eta_p^2 = .05$, a main effect for type of norm $F(3,232) = 7.48$, $p < .001$, $\eta_p^2 = .09$, as well as an interaction between the two factors $F(3,232) = 4.81$, $p = .003$, $\eta_p^2 = .06$. Planned $t$-tests show that participants' blame ratings differ in the predicted manner, that is, higher ratings for norm-violating than for norm-conforming behavior, for the moral (norm-violating: $M = 5.93$, $SD = 0.91$, norm-conforming: $M = 3.53$, $SD = 2.27$; $t(58) = 5.38$, $p < .01$, $r = .58$) and the social–conventional norm (norm-violating: $M = 5.87$, $SD = 1.43$, norm-conforming: $M = 4.4$, $SD = 2.27$; $t(58) = 3.00$, $p < .01$, $r = .37$; see Figure 5). No such difference could be found for the

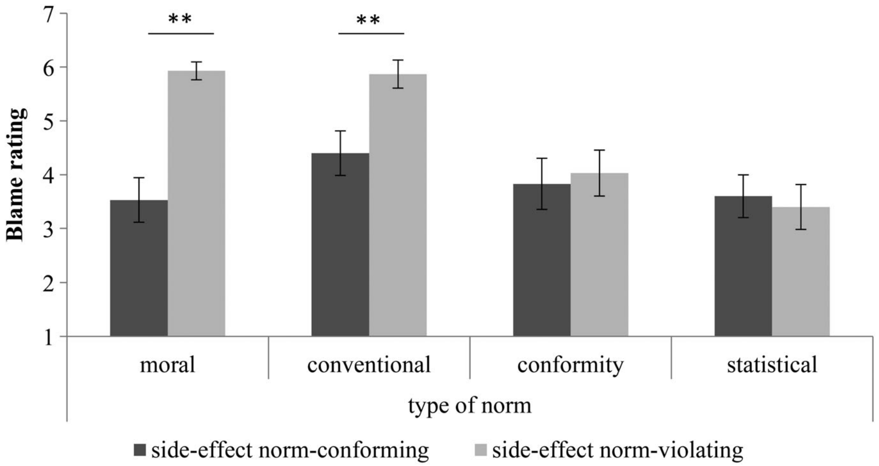

**FIGURE 5** Ratings of how much the person agrees with the following sentence: "The side-effect occurred. One can blame Snoj for that." on a scale from 1 (not at all) to 7 (totally). *Note.* Error bars display standard errors, $^{*}p < .05$, $^{**}p < .01$

two statistical norms (with social conformity: norm-violating: $M = 4.03$, $SD = 2.34$, norm-conforming: $M = 3.83$, $SD = 2.60$; $t(58) = 0.31$, $p = .76$, $r = .04$, $BF = 4.92$ [Bayes factor for null over alternative]; without social conformity: norm-violating: $M = 3.4$, $SD = 2.28$, norm-conforming: $M = 3.6$, $SD = 2.18$; $t(58) = .35$, $p = .73$, $r = .05$, $BF = 4.87$ [Bayes factor for null over alternative]).

As this pattern differs from the pattern in Study 1, in the form that there was a difference in intentionality ratings for the case of statistical norms with social-conformity but no difference in blame judgments in Study 4, these results suggest that at least for our scenarios a blame judgment is not the (only) source for higher intentionality ratings in norm-violating cases. One should have in mind, though, that the culpable control model (Alicke, 2000) talks about preceding blame judgments, which also could be on an implicit level that is not reflected in the explicit judgments about the blameworthiness in Study 4.

## 4.5 | Study 5: Default-shift

Study 5 was based on the explanation by the default-shift account (Knobe, 2010). It claims that a shift in the default attitude towards a behavior that is against or in line with a norm can account for the different intentionality ratings. Therefore, for those kinds of norms where we find a difference in intentionality ratings for norm-conforming vs. norm-violating behavior, we should also find a difference in attitude ratings such that norm-conforming behavior is judged as positive while norm-violating behavior is judged as slightly negative. No such difference would be expected for the statistical norm without social-conformity according to the theory.

As this default shift is supposed to be independent of the behavior of the agent (see Knobe, 2010), in this study we asked participants to give a judgment about the attitude that inhabitants of the planet would normally have towards a norm-violating/norm-conforming behavior just based on the general norm without presenting the whole vignette.

### 4.5.1 | Methods

We gave 240 participants ($N = 60$ per type of norm) a simple description of the relevant norm (e.g., for the social–conventional norm: "We find ourselves on planet Zeik. Many people on planet Zeik have flowers. On planet Zeik there is the rule that inhabitants are only allowed to have flowers that are yellow in their gardens."). They were then asked to give a rating to the question "What do you think, what attitude do the inhabitants of Zeik normally have towards norm-conforming/norm-violating behavior?" on a scale from 1 to 7 with 1 being "negative attitude" to 7 "positive attitude" with 4 being "neutral."

### 4.5.2 | Results and discussion

The attitude ratings given by the participants are shown in Figure 6. The 2 (valence of side-effect) × 4 (type of norm) ANOVA revealed a main effect for valence $F(1{,}232) = 38.62$, $p < .001$, $\eta_p^2 = .14$, a main effect for type of norm $F(3{,}232) = 6.73$, $p < .001$, $\eta_p^2 = .08$, and an interaction between the two factors $F(3{,}232) = 8.28$, $p < .001$, $\eta_p^2 = .10$. The pattern of results from the planned *t*-tests matches the one found in Study 1: people gave slightly positive ratings to norm-conforming behavior and slightly negative ratings to norm-violating behavior for the moral (norm-violating: $M = 3.17$, $SD = 1.09$, norm-conforming: $M = 5.4$, $SD = 1.33$; $t(58) = 7.13$, $p < .001$, $r = .68$), the social–conventional (norm-violating: $M = 2.83$, $SD = 1.15$, norm-conforming: $M = 4.23$, $SD = 1.94$; $t(58) = 3.40$, $p < .001$, $r = .41$), and the statistical norm with social-conformity (norm-violating: $M = 3.93$, $SD = 1.66$, norm-conforming: $M = 5.63$, $SD = 1.45$; $t(58) = 4.23$, $p < .001$, $r = .49$). No such difference was found for the statistical norm without social-conformity (norm-violating: $M = 4.33$, $SD = 1.92$, norm-conforming: $M = 3.93$, $SD = 1.53$; $t(58) = -.89$, $p = .38$, $r = .12$, $BF = 3.59$ [Bayes factor for null over alternative]).

This pattern of results is in line with the proposed explanation from the default-shift account (Knobe, 2010) that the "normal" attitude a person has towards a certain norm-conforming/norm-violating behavior can predict whether or not this behavior is judged as intentional.

## 5 | STEP 3: EXPERIMENTAL MANIPULATION

Studies 1–5 have shown (a) that the SEE arises not only for prescriptive norms but also for descriptive norms if they are confounded with social-conformity and (b) that this pattern of intentionality ratings in different scenarios involving actions in accordance with/in violation of different types of

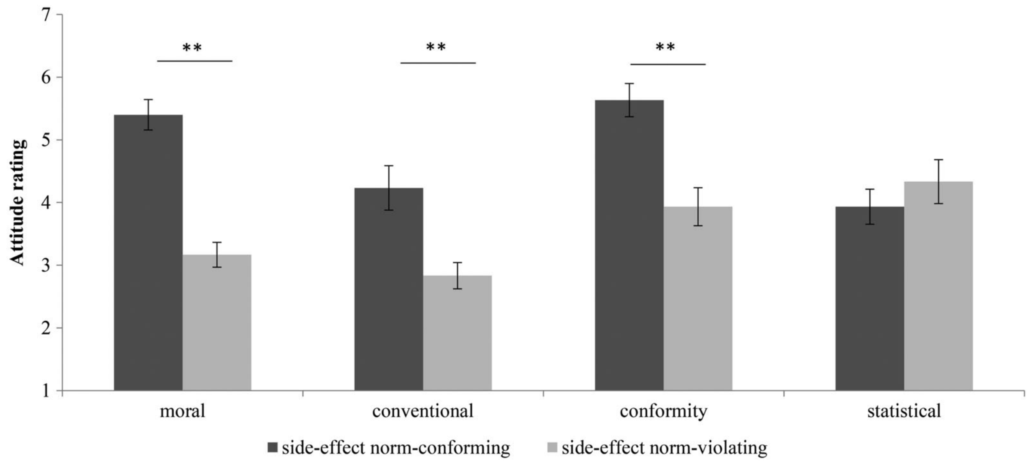

**FIGURE 6** Ratings to the question: "What do you think: what attitude do the inhabitants of Zeik normally have towards norm-conforming/norm-violating behavior?" on a scale from 1 (negative attitude) to 7 (positive attitude). Note. Error bars display standard errors, **$p < .01$

norms is mirrored by the pattern of subjects' default attitudes, but not the pattern of their prescriptiveness, unexpectedness or blame ratings. In a third step, we therefore took the shift in default attitudes and tested directly whether a manipulation of this variable would also lead to a change in intentionality judgments by the participants. The logic here was the following: if the non-existing shift in the default attitudes in the statistical case without social-conformity led to the absence of the typical SEE in our vignette, then by explicitly introducing different default attitudes one should be able to elicit the prototypical difference in intentionality ratings in this vignette, too.

## 5.1 | Study 6: Manipulation of default-shift

### 5.1.1 | Methods

80 participants each received one of two slightly modified versions of the norm-violating case of the statistical norm without social-conformity vignettes described in the general methods section: after the salesman told Snoj about the side-effect of the fertilizer (i.e., blue flowers growing next to his tree) a sentence was added, stating that “Most people on planet Zeik would be reluctant [excited] about blue flowers on the planet.” The rest of the story followed as in the original vignette. Participants were then asked to rate how much they agreed with the statement “Snoj brought about the side-effect intentionally” on a rating scale of 1–7 with 1 being “no agreement at all” to 7 “total agreement”, with 4 being “unsure.”

### 5.1.2 | Results and discussion

Participants' ratings of how much they agreed with the statement that Snoj brought about the side-effect intentionally are displayed in Figure 7. The pattern of results suggests that people's intentionality ratings were sensitive to the manipulation of the default attitude: they gave higher intentionality ratings in the reluctant ($M = 4.35$, $SD = 2.26$) than in the excited ($M = 3.18$, $SD = 2.17$, $t(78) = 2.37$, $p = .02$, $r = .26$) condition.

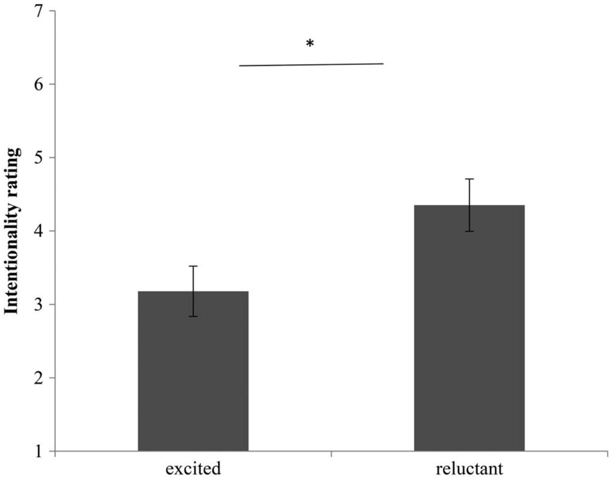

**FIGURE 7** Ratings of how much the person agrees with the following sentence: “Snoj brought about the side-effect intentionally” on a scale from 1 (not at all) to 7 (totally). *Note.* Error bars display standard errors, $*p < .05$

## 6 | GENERAL DISCUSSION

The present set of studies investigated the content-specificity and the cognitive basis of the SEE in three steps. In the first step, we investigated whether the SEE was extendable to descriptive norms. Subjects in Study 1 were confronted with contrast pairs in which an agent did or did not act in accordance with different types of prescriptive (moral, conventional), descriptive (purely statistical norms) and potentially mixed norms (statistical with social-conformity). In a second step, in Studies 2–5, the patterns of intentionality ratings in these cases were related to direct measures of the crucial candidates of the cognitive basis underlying the SEE postulated by leading theoretical accounts (prescriptiveness, [un]expectedness, blame attribution and default shift). The third step experimentally manipulated the most promising candidate for the cognitive foundation derived from Studies 1–5, the shift in subjects' default assumptions.

The main results were the following. First, the patterns of intentionality judgments displayed the typical SEE bias for three of the four types of norms tested (moral, conventional and statistical with social-conformity). Asking people whether they thought the norm in question was action guiding revealed that they judged only the moral and the conventional norm as prescriptive, suggesting the SEE to be actually extendable to descriptive norms. Second, directly testing four proposed candidates for the cognitive basis of the SEE (prescriptiveness, unexpectedness, blame and default shift) showed that the pattern found for intentionality judgments in Study 1 was only reflected by the analogous pattern concerning default shifts. Third, after experimentally manipulating the default attitude, the typical SEE bias was also found in the statistical norm without social-conformity condition. This pattern suggests that it was in fact the shift in default attitudes that elicited subsequent asymmetries in intentionality judgments.

In the current set of studies we used a novel, three-step approach as an attempt to disentangle the different theories under consideration. What are the advantages of this on first glance seemingly indirect and unusual approach? Most importantly, our approach allowed us to give all four candidates the same fair chance as we could test each of them independently and in their best constellation. Take, for instance, the case of the default shift. According to Knobe (2010) the default is supposed to be independent of the behavior of the agent. Consequently, default attitudes should be assessed without any action conducted or at least before participants hear about the actual action. A similar configuration arises also for the prescriptiveness ratings which are also best assessed independently of the agents' actions. Such a constellation of different optimal positions of test questions opens up problems for many experimental designs, as it restricts the required counterbalancing. However, this was not the case for our stepwise approach. And additionally, as we used the same basic vignette throughout the experiments, results remained comparable across the different studies. Furthermore, the experimental manipulation of default attitudes in the third step allowed us to draw causal inferences.

Undoubtedly, there are also disadvantages arising from the indirectness of the approach, such that we cannot directly relate patterns from one candidate to another and consequently cannot test their mediation on intentionality judgments. However, we want to emphasize here, that, to our knowledge, no theory directly predicts a correlation between the ratings for their candidate with intentionality ratings. On the contrary, the theories mainly imply some kind of threshold (for example, when there is a default shift, the SEE should occur) rather than a direct influence (for example, the bigger the default shift, the bigger the differences in intentionality ratings). And exactly these kinds of thresholds are captured in our approach.

### 6.1 | Theoretical implications

What are the theoretical implications of the present findings? First of all, the current findings seem incompatible with accounts that view the SEE exclusively rooted in *prescriptive* norms that are directly action-guiding (in contrast to descriptive norms): in the case of the statistical norm with social-conformity, participants showed the typical SEE bias for intentionality judgments while denying that the norm should regulate the agent's actions.

Second, the present data are not easily compatible with the idea that the effect should be extendable to all kinds of statistical norms or generalizations, since participants did not show the SEE in the case of statistical norms without social-conformity. Are the data then strictly incompatible with the ideas of the rational scientist view (Uttich & Lombrozo, 2010)? Not necessarily. The rational scientist view considers the difference in diagnostic quality in help/harm cases as the basis for the bias in intentionality ratings. The account claims that people derive information from different factors, including norms, to infer default mental states, which can be updated after a behavior, if that behavior provides new information about, for example, the desires or intentions of the actor. One important factor in this process is the unexpectedness of the behavior observed (see Uttich & Lombrozo, 2010). Looking at unexpectedness ratings in Study 3, people were surprised by Snoj's norm-violating action even in the statistical norm without social-conformity. The most obvious interpretation of these findings would thus be that the two statistical norms (with and without social-conformity) should not differ in their quality to serve as the basis for the mental state ascription and that behavior against both norms should lead to a change in mental state ascription.

However, from the point of view of the rational scientist account, (un-)expectedness of a given behavior might be an important manifestation of diagnostic quality with regard to the underlying mental states of the agent, but this may by far not be the only way of diagnostically inferring mental states from behavior that may explain the SEE. Rather, people might derive information about a person's mental state from various sources of information, including such more complex matters as the willingness of the manager in the original vignette to go ahead with the program despite knowing about the harming consequences. Therefore, it might be that subjects, though being surprised about the agent's behavior, did not ascribe intentionality in the statistical norm without social-conformity scenario since they thought the norm in question did not have enough predictive power and therefore did not see a difference in diagnostic quality between norm-conforming and norm-violating behavior. Empirically, future research is necessary to pinpoint the different factors (other than unexpectedness) that lead to these differences in diagnostic quality. And theoretically, the rational scientist view would then need to incorporate and spell out some additional assumptions to explain which factors beyond (un-)expectedness differentially influence diagnostic validity of norm-violating vs. norm-conforming behavior and why and in which ways they apply to the social-conformity but not the purely statistical scenario.

Third, the present results do not support morality accounts that emphasize differential blame attribution towards the agent as the basis of the SEE either. In the case of the statistical norm with social-conformity, subjects refused to blame the agent for his behavior but nevertheless showed the SEE concerning intentionality judgments.

Fourth, while the current data turned out to be in tension with these three accounts in one or the other way, they are clearly compatible with a more general default-shift account. The basic idea of such an account is that a shift in the default attitude towards a certain behavior combined with the actual attitude someone expresses in a given situation can explain the differential intentionality ratings given constituting the SEE. In the present study, when considering subjects' explicit attitude ratings, the proposed default shift towards slightly negative attitudes for norm-violating behavior and slightly positive attitudes for norm-conforming behavior was found for exactly the same situations in which the SEE appeared: moral, social–conventional and statistical norms with social-conformity. Furthermore, offering more direct and causal information, the experimental manipulation of the default shift, giving people information about either a positive or a negative default attitude, led to a difference in intentionality ratings in the predicted manner.

Such a more general idea of default shift can be seen as a friendly amendment to Knobe's (2010) original explanation for the typical side-effect effect vignette. The basic claim in Knobe's account was that *morally* valenced actions lead to a shift of the default on the attitude continuum. In a morally negative scenario, the default attitude is to judge the immoral action as rather negative, whereas in a morally positive scenario the default shifts towards the pro side of the continuum. As the intentionality judgment is anchored to the default (actions on the pro side in relation to the default are seen as intentional), this default shift directly affects intentionality judgments: if the default shifts towards the negative side of the continuum a neutral attitude expression is already judged as intentional; a default shift towards the positive side of the continuum has as a consequence that only really positive attitude expressions are seen as intentional.

The more general extension of this default shift idea proposed here has it that those kinds of default shifts are not limited to moral norm situations. Rather, default shifts also appear for other kinds of prescriptive or even descriptive norms. And it is this more general default shift idea that is supported by the data of the present studies. This extension broadens the explanatory scope of the account remarkably. Crucially, it transcends the group of morality accounts as the emphasis is no longer on specifically *moral* considerations but on *normality* considerations more general. The main question to be asked is thus what people normally think about a certain behavior and how this attitude differs from the attitude behind the actually executed action—independent of whether the behavior has *moral* valence.

Now, this extended account itself makes several interesting novel predictions worth testing in future studies using naturally occurring variation as well as experimental manipulations of normative defaults (as has been done for the morally valenced cases for example, by Shepherd, 2012). Such more systematic future studies will then help to elucidate whether, in fact, and if so in which ways the shifting of default attitudes towards action outcomes is the crucial cognitive basis of the side-effect effect.

### 6.2 | Limitations and outlook

Potential limitations stem from both the methods and design of the current study. We asked subjects to give explicit judgments for what might be (at least partly) implicit processes. In particular, this might have affected the explicit prescriptiveness judgments in the case of statistical norms with social conformity. Though subjects might have thought that the agent in the scenario should have done what everyone else did, they might not have felt comfortable expressing this explicitly. However, it is unlikely that this had a severe impact on the general pattern of results, since in Study 2b, where subjects were asked to not give their own but another person's evaluation, the prescriptiveness ratings also did not fit intentionality ratings from Study 1. Relatedly, subjects' blame judgments might have been underestimated by the explicit judgments. The culpable control model postulates "spontaneous negative reactions" leading to blame judgments (Alicke, 2000), and such negative reactions could be more subtle-feelings not adequately tapped by explicit questions. However, participants did answer according to the theory in the moral and social–conventional scenarios, implying that at least some notion of this reaction can be captured in explicit judgments.

Additionally, the current conclusions are clearly limited to the comparisons between the four accounts tested and, bearing in mind the amount of untested theories (e.g., Adams & Steadman, 2004a, 2004b; Machery, 2008; Nadelhoffer, 2006; Sripada, 2010), cannot be seen as a general solution concerning the cognitive foundation of the side-effect effect.

That said, between the four accounts tested here, the current body of evidence as a whole clearly favors the default-shift account. But, this does not mean that a shift in default attitudes is necessarily the *only* factor driving people's intentionality judgment. In fact, it is highly plausible and likely that for a large class of cases it will be the interplay of the different factors that best explains intentionality ratings. For instance, in Study 6, where we found the SEE after manipulating people's default attitudes, this manipulation might have led to an increase in people's blame ratings, too, which then also influenced intentionality ratings. A fundamental goal for future research in this area is thus to investigate the dynamic cognitive interplay of representations of blame, prescriptiveness, default-shifts, unexpectedness, and other potentially influencing factors underlying the side-effect effect using a larger variety of vignettes.

## ACKNOWLEDGMENTS

Thank you to Taina Gabriel, Joscha Franke, Tabea Müller, Daniel Freund and Carina Neumann for help with data collection and coding. Special thanks go to Joshua Knobe for useful comments and ideas on earlier drafts of this paper and to Ralf Mayerhofer for help with computing the Bayesian analyses.

## REFERENCES

Adams, F. & Steadman, A. (2004a). Intentional action and moral considerations: Still pragmatic. *Analysis*, *64*(3), 268–276. https://doi.org/10.1111/j.0003-2638.2004.00496.x

Adams, F. & Steadman, A. (2004b). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, *64*(282), 173–181. https://doi.org/10.1111/j.1467-8284.2004.00480.x

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574. https://doi.org/10.1037/0033-2909.126.4.556

Alicke, M. D. & Rose, D. (2010). Culpable control or moral concepts? *Behavioral and Brain Sciences*, *33*(04), 330–331. https://doi.org/10.1017/S0140525X10001664

Alicke, M. D., Rose, D. & Bloom, D. (2011). Causation, norm violation, and culpable control. *Journal of Philosophy*, *108*(12), 670–696.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Guglielmo, S., Monroe, A. E. & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry*, *52*(5), 449–466. https://doi.org/10.1080/00201740903302600

Holton, R. (2010). Norms and the Knobe effect. *Analysis*, *70*(3), 1–8. https://doi.org/10.1093/analys/anq037

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194. https://doi.org/10.1093/analys/63.3.190

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*(2), 309–325. https://doi.org/10.1080/09515080307771

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, *9*(8), 357–359. https://doi.org/10.1016/j.tics.2005.06.011

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315–329. https://doi.org/10.1017/S0140525X10000907

Knobe, J. & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture*, *6*(1–2), 113–132. https://doi.org/10.1163/156853706776931222

Leslie, A. M., Knobe, J. & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, *17*(5), 421–427. https://doi.org/10.1111/j.1467-9280.2006.01722.x

Lewis, D. (1969). *Convention: A philosophical study*. Cambridge: Harvard University Press.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language*, *23*(2), 165–189. https://doi.org/10.1111/j.1468-0017.2007.00336.x

Michelin, C., Pellizzoni, S., Tallandini, M. & Siegal, M. (2009). Evidence for the side-effect effect in young children: Influence of bilingualism and task presentation format. *European Journal of Developmental Psychology*, *7*(6), 641–652. https://doi.org/10.1080/17405620902969989

Nadelhoffer, T. (2006). On trying to save the simple view. *Mind & Language*, *21*(5), 565–586. https://doi.org/10.1111/j.1468-0017.2006.00292.x

Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, *84*(2), 221–236. https://doi.org/10.1016/s0010-0277(02)00048-3

Pellizzoni, S., Siegal, M. & Surian, L. (2009). Foreknowledge, caring, and the side-effect effect in young children. *Developmental Psychology*, *45*(1), 289–295. https://doi.org/10.1037/a0014165

Pettit, D. & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, *24*(5), 586–604. https://doi.org/10.1111/j.1468-0017.2009.01375.x

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Shepherd, J. (2012). Action, attitude, and the Knobe effect: Another asymmetry. *Review of Philosophy and Psychology*, *23*(2), 171–185. https://doi.org/10.1007/s13164-011-0079-7

Smetana, J. G., Rote, W. M., Jambon, M., Tasopoulos-Chan, M., Villalobos, M. & Comer, J. (2012). Developmental changes and individual differences in young children's moral judgments. *Child Development*, *83*(2), 683–696. https://doi.org/10.1111/j.1467-8624.2011.01714.x

Sripada, C. S. (2010). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*(2), 159–176. https://doi.org/10.1007/s11098-009-9423-5

Uttich, K. & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87–100. https://doi.org/10.1016/j.cognition.2010.04.003

**How to cite this article:** Proft M, Dieball A, Rakoczy H. What is the cognitive basis of the side-effect effect? An experimental test of competing theories. *Mind Lang*. 2019;34:357–375. https://doi.org/10.1111/mila.12197