Has the Side-Effect Effect been cancelled? (No, not yet.)[1]

Justin Sytsma, Robert Bishop, and John Schwenkler

Abstract: A large body of research has found that people judge bad foreseen side effects to be more intentional than good ones. While the standard interpretation of this Side-Effect Effect (SEE) takes it to show that the ordinary concept of intentionality is influenced by normative considerations, a competing account holds that it is the result of pragmatic pressure to express moral censure and, thus, that the SEE is an experimental artifact. Attempts to reveal this have previously been unsuccessful, however. That is until recently, when Lindauer and Southwood (2021) presented a study purporting to cancel the SEE. We are not convinced. Here, we detail three studies testing their interpretation. The results indicate that it is the purported cancellation, rather than the SEE, that is an experimental artifact.

1. Introduction

Are intentionality judgments influenced by normative considerations? More specifically, when a person's action brings about a foreseen side effect—a consequence that isn't part of the person's goal in performing the action, but that they are able to anticipate in advance—does the perceived valence of the side effect (whether it is seen as being relatively good or bad) make a difference with regard to whether the person is judged to have brought it about intentionally? While philosophers have tended to answer 'no' on the basis of *a priori* considerations, a large body of research suggests that the commonsense position is different from this. All else being equal, people seem to judge bad foreseen side effects of an action as more intentional than good foreseen side-effects. The classic demonstration of this effect uses Joshua Knobe's chairman case (2003a, p. 191):

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also [harm/help] the environment.'

[1] Forthcoming in *Synthese*. We presented a version of this paper at the 2021 Agency and Intentions in Language workshop organized by Julie Goncharov and Hedde Zeijlstra, and are grateful to the audience for their feedback. Thanks also to Joshua Knobe and to two referees for *Synthese* for their helpful comments.

> 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].

In Knobe's seminal study, participants read either the 'harm' or 'help' version of this vignette, then indicated their agreement with the statement that the chairman had harmed or helped the environment intentionally. Strikingly, large majorities of participants expressed agreement with this test statement in the harm condition but disagreement with it in the help condition. This asymmetry in intention attribution has been dubbed the *Side-Effect Effect* (SEE), a.k.a. the Knobe effect.

Subsequent investigation has replicated the side-effect effect with different cases[2], languages[3], populations[4], and concepts[5] (for a review, see Cova 2016), but disagreement remains about how to understand the results. One persistent disagreement concerns two competing ways in which the SEE might be explained. The *straightforward* account of the SEE holds that participants in these experiments mean just what they say: they are, by their own lights, applying 'intentionally' literally and felicitously in their responses. By contrast, the *pragmatic* account holds that participants say that bad side effects were brought about intentionally only as a way to avoid the appearance of *excusing* the agent for what they did (e.g., Adams & Steadman 2004, 2007). That is, if participants believe that knowingly bringing about a bad outcome makes a

---

[2] See, for example, Knobe (2003b), Nadelhoffer (2004), and Knobe (2006). See also Cushman & Mele (2008) for the effect tested with a battery of 16 cases. Note that Cushman & Mele did find interesting ordering effects on the cases, but their results still displayed the Knobe effect's asymmetry.

[3] See, for example, Knobe & Burra (2006), Cova & Naar (2012), Dalbauer & Hergovich (2013), and Mizumoto (2018).

[4] See, for example, Leslie et al. (2006), Young et al. (2006), Pellizzoni et al. (2009), and Zalla & Leboyer (2011).

[5] See, for example, Knobe (2004, 2010), Pettit & Knobe (2009), Hitchcock & Knobe (2009), and Phillips et al. (2015).

person blameworthy for it while knowingly bringing about a good outcome does not make a person worthy of credit, then the asymmetry in attributions of intentionality might arise as an artefact of this asymmetry in the preconditions for credit versus blame. If the straightforward account is right, then the SEE provides crucial insight into the way that ordinary people understand intentional action. But if the pragmatic account is right, then the SEE does not give us this kind of insight; rather, it is just a matter of an inadequate experimental design in which participants are given insufficient opportunity to say what they really think. The resolution of this disagreement is therefore crucial for understanding the significance of the SEE.

Fortunately, there seems to be a straightforward way of testing pragmatic explanations like the one just outlined. If the SEE is the result of pragmatic pressure to *morally censure* agents for bringing about bad side effects for which they are judged to be blameworthy, then alleviating that pressure by giving participants another way to censure the agents should largely *cancel* the effect. More specifically, if the SEE is an artefact of such a pragmatic effect, then when, for example, participants are given the opportunity to morally censure Knobe's chairman in some other way, they should now tend to *disagree* with the statement that he intentionally harmed the environment. This is because having an alternative way of censuring the chairman should relieve the pragmatic pressure from which the SEE is supposed to arise.

Given the significance of the disagreement that we outlined above, there is a sizable bounty on showing the SEE to be cancellable in the way just described. Yet efforts to do this have so far been largely unsuccessful (e.g., Adams & Steadman 2007, Nichols & Ulatowski 2007). Several cancelling studies have fallen short in print, and it's likely that the data from many other attempts at cancelling the effect will never see the light of day. These failures suggest that the quarry may not be so much elusive as illusory. A recent study, however, seems

to have bagged the prize. Matthew Lindauer and Nicholas Southwood (2021; henceforth 'L&S') take themselves to have successfully cancelled the SEE.

L&S suspected that previous attempts to cancel the effect were unsuccessful because they had not given participants the opportunity to censure the chairman's action strongly enough and, thus, did not sufficiently mitigate the pragmatic pressure to say that he harmed the environment intentionally. To remedy this, L&S ran a new study with three between-participants conditions. In the crucial condition of their study participants read the harm version of Knobe's chairman vignette and then rated the following *cancelling* statement on a 7-point scale:

(C) The chairman didn't intentionally harm the environment, but he knowingly harmed the environment, and he is morally responsible and should be blamed for doing so.

Since there was no appropriate way to phrase it, L&S did not present a corresponding statement for the help version of the chairman vignette; rather, responses in this cancelling condition were compared with those from two other conditions in which participants read either the 'harm' or 'help' version of Knobe's chairman vignette and then rated the relevant version of the following *simple* statement, corresponding to the first clause of (C):

(S) The chairman didn't intentionally [help/harm] the environment.

L&S predicted that the opportunity to express strong moral censure of the chairman through the second and third clauses of (C) would relieve the pragmatic pressure to say that the chairman had harmed the environment intentionally, thus leading participants to agree with (C) overall, even though its first clause denies that the chairman harmed the environment intentionally. They predicted, therefore, that responses to (S) in the help and harm conditions would exhibit the SEE, but that responses to (C) in the cancelling condition would be more similar to responses to (S) in the help condition than in the harm condition, even though participants in the cancelling condition had read the harm vignette.

The results of this experiment were in line with L&S's predictions. While ratings of (S) in the help condition and harm condition exhibited the usual asymmetry, ratings of (C) in the cancelling condition were not significantly different from ratings of (S) in the help condition. Since the primary difference between the cancelling condition and harm condition was the inclusion in (C) of the clauses following 'but', which express strong moral censure of the chairman, L&S take these results to constitute strong evidence for the pragmatic account of the SEE.

Our paper challenges this interpretation of L&S's findings. We begin from the observation that their cancelling statement (C) has two parts, separated by the contrastive conjunction 'but'. The first part of (C) is a denial that the chairman intentionally harmed the environment:

(~I) The chairman didn't intentionally harm the environment.

Further, the second part of (C) is a positive attribution of moral responsibility to the chairman, which gave participants the opportunity to censure his action:

(R) The chairman knowingly harmed the environment, and he is morally responsible and should be blamed for doing so.

As we have seen, L&S assume that their participants expressed agreement with (C) because they agreed independently with *both* (~I) *and* (R), and that the opportunity to express their agreement with the latter statement relieved pragmatic pressure to deny the former.

Yet this assumption is debatable. Given that (C) contains such a complex series of descriptions of the chairman's activity, it is possible that L&S's participants expressed agreement with it even though they did not agree independently with each of its parts. (Indeed, it is possible that they did so because this was the only way to attribute responsibility to the chairman for the harm.) Alternatively, given that (C) differs from the corresponding version of

(S) in that it includes all the descriptions identified under (R) above, it could be that there is something about the presence of these additional clauses that influences how the phrase 'intentionally harmed the environment' is understood. We will discuss these matters in more detail in our concluding section. The crucial thing to emphasize here is the essential role of the following assumption in L&S's interpretation of their findings: that their participants expressed agreement with (C) because they agreed with each one of its independent clauses, and in particular with the clause we have labeled (~I). Absent this assumption, their findings provide no evidence for the pragmatic account.

Fortunately, there are several straightforward ways to modify L&S's experimental paradigm in order to probe this core assumption, namely by having participants evaluate judgments of intentionality and responsibility either independently or as parts of a statement where they are joined with a connective other than 'but'. Below we present the results of three studies in which we did just this. In each case, the results ran counter to L&S's key assumption, and also to the predictions of the pragmatic account. In our concluding section, we discuss what to make of the state of play, and consider three alternative explanations of L&S's results in light of our new findings.

2. *Study 1: Rank Ordering*

In our first study, each participant read the harm version of the chairman case, and then was asked to rank-order a series of statements concerning the intentionality of the chairman's action and his responsibility for the environmental harm. For clarity, in each statement the connective was emphasized and negations were bolded, as shown below. The statements were displayed in random order:

Please rank the following four claims about the chairman of the board in order of how much you agree with them, with (1) being the claim you most strongly agree with and (4) being the claim you most strongly disagree with:

(I and R) The chairman intentionally harmed the environment, ***and*** he knowingly harmed the environment, is morally responsible for doing so, and should be blamed for it.

(~I but R) The chairman did **not** intentionally harm the environment, ***but*** he knowingly harmed the environment, is morally responsible for doing so, and should be blamed for it.

(I but ~R) The chairman intentionally harmed the environment, ***but*** he did **not** knowingly harm the environment, is **not** morally responsible for doing so, and should **not** be blamed for it.

(~I and ~R) The chairman did **not** intentionally harm the environment, ***and*** he did **not** knowingly harm the environment, is **not** morally responsible for doing so, and should **not** be blamed for it.

The labels in parentheses were not displayed to participants, but are shown here for convenience. As can be seen, the four statements that were displayed in this study exhaust the logical possibilities for combining the statements (I) and (R).

Crucially, since participants in the present study must engage with all four statements in order to produce a preferred ranking, they are able to register their moral censure of the chairman while still showing a preference for *either* (I) or (~I). In this context, there should therefore be little to no pragmatic pull to show a preference for the statements affirming that the chairman harmed the environment intentionally, as agreement with these statements is no longer needed in order to express moral censure. Given this, the pragmatic account should predict that participants will tend to show a preference for (~I but R) over the other three statements, given that they are supposed to agree independently with both (~I) and (R).

Each study in this paper was conducted online with participants recruited through advertising for a free personality test on Google in North America. Prior to considering the

philosophical scenario, participants answered basic demographic questions. At the end of the experiment they took a 10-item Big Five personality inventory. Results for Study 1 were collected from 67 participants who reported that they were 16 years of age or older and hadn't taken the survey previously.[6]

Histograms of rank orderings are shown in Figure 1. Contrary to the predictions of the pragmatic account, participants showed a clear preference for (I and R) over (~I but R). Indeed, not only did a significantly larger proportion of participants rank (I and R) highest compared to (~I but R),[7] but the former was ranked higher by a significant majority of participants[8] and had a significantly higher mean rank.[9] In short, despite being able to express strong moral censure of the chairman by giving a high ranking to either one of (I and R) or (~I but R), participants tended to indicate greater agreement with the statement that affirmed that the chairman harmed the environment intentionally. As such, the present findings suggest that agreement with the cancelling statement in L&S's study does *not* reflect a belief that the chairman did not intentionally harm the environment.

---

[6] Participants were 68.7% women, with an average age of 33.5 years.

[7] We use two-tailed tests throughout, except where indicated otherwise. 52.2% ranked (I and R) in first position, compared to 20.9% for (~I but R): χ^2=37.947, p<.001

[8] 65.7% ranked (I and R) above (~I but R): χ^2=5.97, p=.015

[9] We'll use Student's t-tests for one-sample or paired-sample comparisons (as here), and Welch's t-tests for independent-sample comparisons. In this study, (I and R) had a mean rank of 1.90 compared to 2.34 for (~I but R): t(66)=2.11, p=.038, d=.43. Arguably, t-tests aren't appropriate for comparing mean ranks, as rank orderings are very plausibly ordinal rather than interval. A similar result holds using a Wilcoxon signed rank test, however: V=833, p=.051.

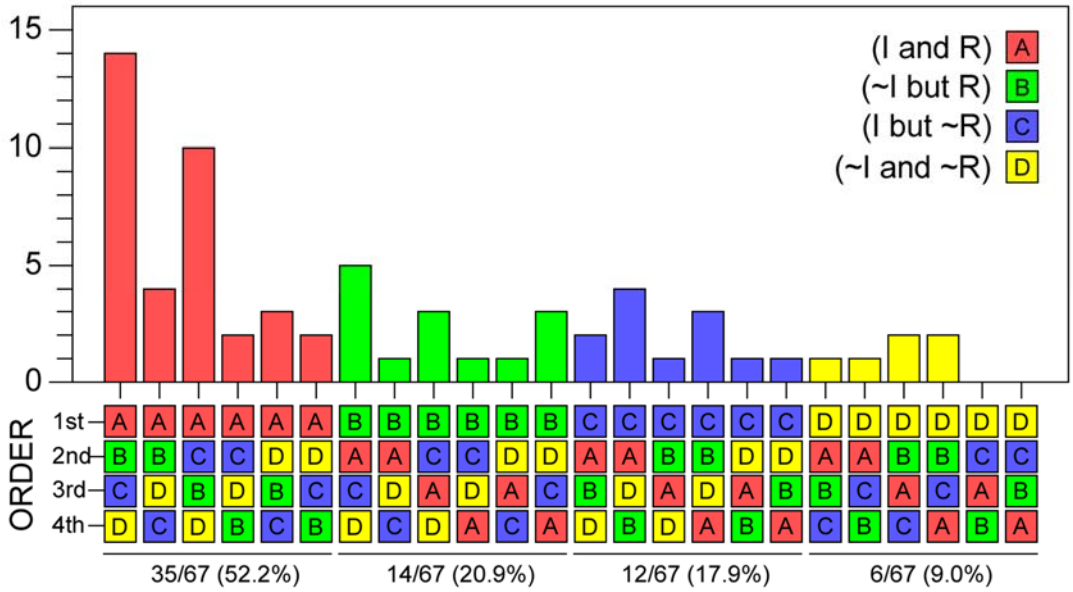

Figure 1: Histogram for each rank ordering.

3. Study 2: Likert Rankings

Our first study elicited relative preferences for the four logical combinations of the statements (I) and (R) and their negations, finding that contrary to the prediction of the pragmatic account, participants tended to prefer (I and R) over (~I but R). But the pragmatic account also makes further predictions about the conditions under which participants will tend to agree or disagree with each of these four compound statements, including how this should vary based on whether the statements are presented together or independently. Our second study aimed to test these predictions.

Generally, as noted in the discussion of our first study, presenting the four statements together should reduce, if not eliminate, any purely pragmatic pressure to agree with (I and R) as a means of expressing moral censure of the chairman. As such, if the pragmatic account is correct, then we would expect participants' ratings of this statement to be affected by whether it is presented individually or in conjunction with the other three statements. In light of this, L&S's

account makes the following prediction about how ratings of (I and R) should be affected by how it is displayed:

(i) Since (I and R) is supposed to say something that participants believe is literally false, any tendency to agree with it should arise only as a way to express moral censure. Therefore, *agreement with (I and R) should be higher when it is presented alone than when it is presented alongside the other three statements*, as the latter condition affords other ways to express moral censure, thus relieving at least some of the supposed pragmatic pressure to agree with it.

Additionally, since presenting the four statements together should serve to reduce the pragmatic pressure to agree with (I), the pragmatic account makes further predictions about how participants should rate (I and R) in a condition that presents it alongside the other three statements. These predictions vary depending on how strongly one expects this condition to relieve the supposed pragmatic pressure. The strongest such prediction is the following:

(ii) In a condition where participants are able to rate both (I and R) and (~I but R), *participants should disagree overall with (I and R)*, as there will be no pragmatic pressure to express agreement with it.

If, however, a defender of the pragmatic account resists this prediction on the grounds that there might still be *some* pressure to agree with (I and R) given the way it expresses moral censure of the chairman, then the following weaker prediction still seems to follow:

(iii) In a condition where participants rate both of the statements in question, *(I and R) should at least receive lower agreement than (~I but R)*, as only the latter is believed to be literally true.

Finally, if the defender tried to dig in and deny even this much, perhaps on the grounds that participants are so inclined to censure the chairman that they will take absolutely any opportunity to do so, then she must at the very least concede the following:

(iv) In a condition where participants rate both of the statements in question, *ratings of (I and R) should be <u>at least as high as</u> those of (~I but R)*, since both express moral censure and only the former is believed to be literally true.

To test these predictions, in our second study we had participants indicate their agreement with each of the four compound statements from Study 1 using a 7-point Likert scale anchored at -3 ('Strongly Disagree'), 0 ('Neither Agree nor Disagree'), and 3 ('Strongly Agree'). All participants read the harm version of the chairman case and then rated one or more of the four statements. In order to test prediction (i) above, agreement ratings were solicited both within-participants (with each participant rating all four statements in random order) and between-participants (with each participant rating just one of the statements). Sample size was selected to correspond with that used by L&S for their cancelling condition, with 100 participants rating each of the four statements, evenly split between the within-participants condition (N=50) and the between-participants conditions (N=50 per condition). In total, results were collected from 250 participants using the same recruitment strategy and restrictions as in Study 1.[10]

The results of this study are shown in Figure 2. The first thing to note is that ratings for each statement are remarkably similar across the two designs, suggesting against the general claim that responses are being notably affected by pragmatic pressure to morally censure the chairman. A linear mixed effects analysis indicates that whether the statements were presented

[10] Participants were 74% women (four non-binary), with an average age of 34.4 years.

individually (between-participants) or as a group (within-participants) did not make a statistically significant difference in the ratings of them.[11] Indeed, ratings were not significantly different between the designs for any of the four statements, including (I and R)—contrary to prediction (i) above.[12] Further, the analysis indicates that across conditions the statement of moral responsibility had a significant effect, with the mean rating being higher for statements affirming (R) than for statements denying it. And while we also find a result for intentionality that is close to the .05 significance threshold using a two-tailed test, this runs in the opposite direction to what we would expect on the pragmatic account: the mean rating was higher for statements affirming (I) than statements denying it.

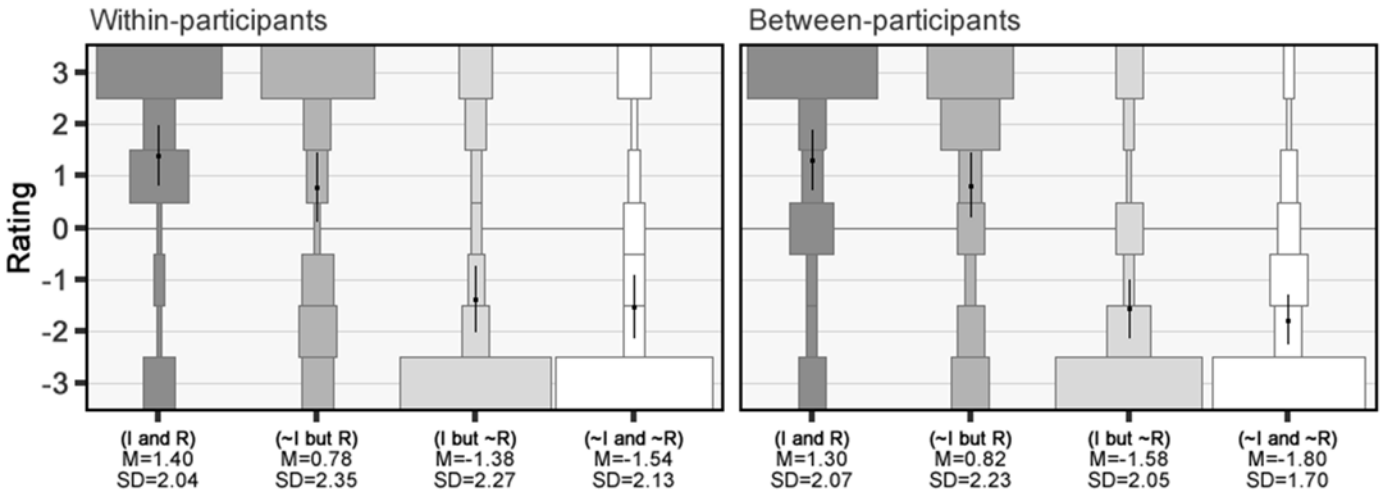

Figure 2: Results for Study 2, showing relative percentage of participants selecting each response option for each statement, with means and 95% confidence intervals overlaid, and split between the within-participants condition (left) and between-participants conditions (right).

[11] The analysis was run on participant responses with *design* (within, between), *intention* (I, ~I), and *responsibility* (R, ~R) as fixed factors and *participant* as a random factor. An analysis of variance for the linear mixed model's fixed effects showed a significant main effect for *responsibility* [$F(1,400)=160.44$, $p<.001$], and a main effect that was close to the .05 significance threshold for *intention* [$F(1,400)=3.13$, p=.078], but no effect for *design* [$F(1,400)=.39$, p=.53], and there were no significant interaction effects. We want to thank an anonymous reviewer for recommending this analysis.

[12] (I and R) $t(97.98)=.24$, $p=.81$, $d=.049$; (~I but R) $t(97.73)=.087$, $p=.93$, $d=.017$; (I but ~R) $t(97.04)=.46$, $p=.64$, $d=.093$; (~I and ~R) $t(93.45)=.67$, $p=.50$, $d=.13$

Turning to the within-participants condition, against the strong prediction (ii) participants not only didn't tend to disagree with (I and R), but they tended to *agree* with it, with the mean rating being significantly above the neutral point.[13] Further, contrary to the weaker prediction (iii), ratings were not lower for (I and R) than for (~I but R); in fact, they were *higher*.[14] Thus, contrary to the weakest prediction (iv), the mean rating for (I and R) was *not* at least as high as for (~I but R). All this provides further strong evidence against the pragmatic account.

Finally, we want to highlight a puzzling finding from this study: participants not only tended to agree with (I and R), as noted above, but they also tended to agree with (~I but R)—replicating L&S's key finding—and did so even in the within-participants condition.[15] Indeed, exactly half of our participants expressed agreement with both (I and R) and (~I but R) even when they were presented simultaneously![16] This suggests that many participants somehow treated the statements 'The chairman harmed the environment intentionally' and 'The chairman did not harm the environment intentionally', as they appear in (I and R) and (~I but R)

[13] $t(49)=4.85$, $p<.001$ (one-tailed), $d=.69$; a similar finding held for the between-participants condition: $t(49)=4.44$, $p<.001$ (one-tailed), $d=.63$

[14] Indeed, the mean response for (I and R) was higher than for (~I but R) and this difference was close to the .05 significance threshold: $t(49)=1.52$, $p=.068$ (one-tailed), $d=.28$. Further, in line with the result of the linear mixed effect analysis reported above, across conditions a partially paired t-test showed a significant difference: $t(125.63)=1.84$, $p=.034$ (one-tailed).

[15] Within-participants: $t(49)=2.35$, $p=.012$ (one-tailed), $d=.33$; between-participants: $t(49)=2.60$, $p=.0061$ (one-tailed), $d=.37$

[16] Agreement was counted as giving a rating above the neutral point (1, 2, or 3). We found that 80% of participants agreed with (I and R) in the within-participants condition, with 62.5% of these participants also agreeing with (~I but R). By contrast, just 18% of participants agreed with both (I and R) and (I but ~R), while only 10% agreed with (I and R) and (~I and ~R). Interestingly, while responses to these items showed a positive correlation, it was not significant ($r=.14$, $p=.32$), with only responses to (I but ~R) and (~I and ~R) showing a significant correlation ($r=.47$, $p<.001$).

respectively, as not saying contradictory things about what the chairman did. We return to this puzzling finding in the concluding discussion.

4. Study 3: Separate Statements

According to Lindauer and Southwood, while people tend to find Knobe's chairman to be deserving of strong moral censure for harming the environment, they do not tend to believe that he harmed the environment intentionally. On their account, the persistent tendency to say that the chairman intentionally harmed the environment is the result of pragmatic pressure to censure the chairman for doing this. And L&S take the strong agreement with their cancelling statement (C) to provide evidence for this account. But the results of our first two studies undermine this reasoning by challenging a core assumption that it depends on. As a final test, in our third study we considered whether it was possible to cancel the proposed pragmatic effect in another way.

Participants were given the harm version of Knobe's chairman case and then asked to rate separately, using the same 7-point scale as in the previous studies, both a simple statement attributing responsibility to the chairman and a simple statement saying that he had harmed the environment intentionally:

(R) The chairman knowingly harmed the environment, and he is morally responsible and should be blamed for doing so.

(I) The chairman intentionally harmed the environment.

Importantly, in this study the two statements were shown in a fixed order, with all participants rating (R) before (I), in order to relieve any pragmatic pressure to express agreement with the latter statement. On the pragmatic account, it seems that participants should therefore tend to disagree overall with (I), given that they have already been able to censure the chairman by expressing their agreement with (R).

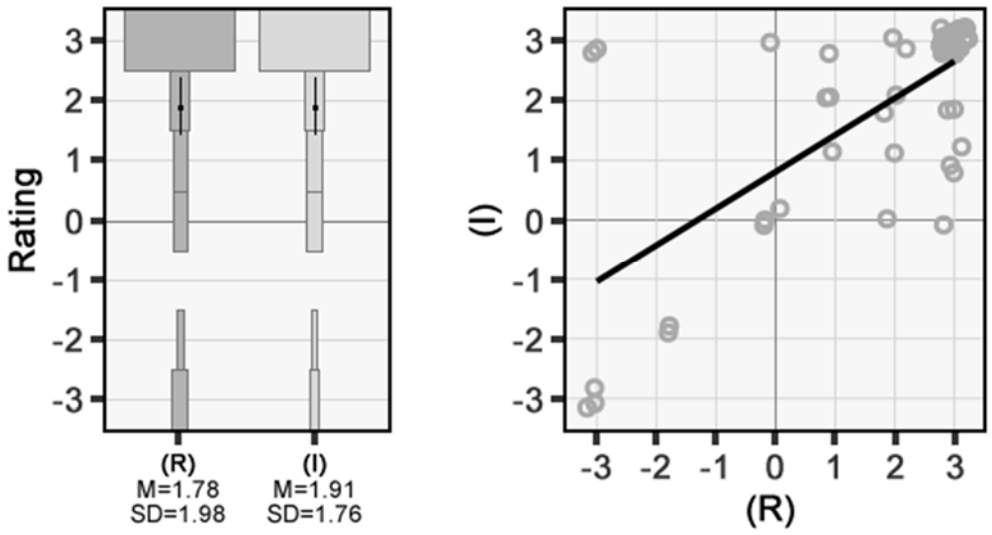

Figure 3: Results for Study 3, showing relative percentage of participants selecting each response option for each statement on the left, with means and 95% confidence intervals overlaid. The scatterplot on the right show points with jitter and regression line calculated without jitter.

Responses were collected from 54 participants using the same recruitment strategy and restrictions as in the previous studies.[17] The results are shown in Figure 3. Against the prediction of the pragmatic account, participants continued to agree with (I) even when it was presented following (R), with the mean rating significantly above the neutral point.[18] In fact, the mean rating for (I) was actually (though non-significantly) *higher* than the mean rating for (R)! Finally, as can be seen in Figure 3, there wasn't a single participant who showed the pattern of responses predicted by the pragmatic account: not a single person affirmed (R) and denied (I).[19] All of this is exactly the opposite of what is predicted by the pragmatic account of the SEE.

---

[17] Participants were 70.4% women (two non-binary), with an average age of 46.7 years.

[18] $t(53)=6.61$, $p<.001$, $d=.90$

[19] In line with this, there was an extremely strong positive correlation between ratings of (R) and (I): $r=0.69$, $t(52)=6.92$, $p<.001$.

5. *Discussion*

Lindauer and Southwood (2021) purport to have *finally* cancelled judgments of intentionality for the harm version of Knobe's (2003a) chairman case, thereby vindicating the pragmatic account of the Side-Effect Effect. This conclusion rests on the assumption that their participants' tendency to agree with their cancelling statement (C)—the statement that 'The chairman didn't intentionally harm the environment, but he knowingly harmed the environment, and he is morally responsible and should be blamed for doing so'—reflects independent agreement with the claims on *each* side of the connecting 'but'. However, the three new studies reported in this paper provide ample evidence that this key assumption is unfounded. Despite tending to agree with (C), most participants nonetheless expressed agreement with claims saying that the chairman intentionally harmed the environment, even when they were also able to express strong moral condemnation of the chairman in some other way. In other words, our results indicate that L&S's supposed cancellation of the SEE was an artefact of their limited experimental design.

At the same time, these results also raise a further question that is interesting in its own right, namely: Why do people tend to agree with L&S's cancelling statement despite their apparent belief, as revealed in our studies, that the chairman intentionally harmed the environment? It is beyond the scope of the present paper to settle this question; indeed, the authors of this paper are divided on the matter. Nevertheless, in closing we want to consider three possible explanations in the light of the results of our studies.

Our studies were initially motivated by the possibility that L&S's original finding might be explained in just the way that the pragmatic account attempts to explain the SEE: perhaps participants expressed agreement with their cancelling statement (C) simply because doing so was the only way to express moral censure of the chairman. While this possibility seems

plausible on its face, the results of our second study suggest that it can't be the whole story. Specifically, as noted above, we found that in the within-participants condition of Study 2, participants tended to agree overall with (~I but R) even though they could express moral disapproval by agreeing with (I and R) instead. Indeed, as we have noted, about half of the participants expressed agreement with *both* of these two statements at once. Further, ratings for (~I but R) in the within-participants condition were not significantly different from ratings of it in the between-participants condition. All of this provides compelling evidence against this simple pragmatic explanation of L&S's finding.

As we noted above, the fact that so many participants expressed agreement with both (I and R) and (~I but R) when they were presented together is itself puzzling: it suggests that many participants treated the statements 'The chairman harmed the environment intentionally' and 'The chairman did **not** harm the environment intentionally' as somehow *not* saying contradictory things about what the chairman did. It seems, then, that a full explanation of L&S's original finding will also need to account for how these statements could be interpreted in this way. Below we will explore two further explanations of this phenomenon, each of which focuses on how the connective 'but' might influence the interpretation of a statement like (~I but R). While these explanations are speculative, each is arguably compatible with our findings in Study 2.

The first of these explanations focuses on how 'but' can be used to introduce a phrase that *intensifies* what is said in the one that precedes it, such as a statement like

(D) Jane didn't make dinner, but prepared a sumptuous feast that delighted the senses, and for this she deserves our deepest gratitude.

This use of the construction 'not … but', which the linguist Larry Horn (1985) describes as a type of *metalinguistic negation*, treats the negated phrase as objectionable 'on the grounds that

the predication it yields, though true, is too weak' (Horn 1985, p, 166). The suggestion, then, is that some participants might have read (~I but R) as saying something like: 'It's not *just* that the chairman harmed the environment intentionally, as he *also* did it knowingly and in a way that makes him morally responsible and blameworthy for doing so'. Read in this way, it is possible to agree with (~I but R) while believing (what would be expressed by a standalone sentence saying) that the chairman intentionally harmed the environment—just as it is possible to agree with the statement above while believing that Jane made dinner.

However, one serious problem with this proposal is that it is debatable whether (~I but R) has the appropriate syntax to suggest that the 'but' is introducing metalinguistic negation, rather than functioning as a true sentential connective (see Horn 1985, p. 166). For instance, making the following bolded addition to (D) seems to render it puzzling, as it suggests that the negated phrase is not just understated, but simply false:

(D*) Jane didn't make dinner, but **she** prepared a sumptuous feast that delighted the senses, and for this she deserves our deepest gratitude.

Since, however, (~I but R) has a syntax corresponding to (D*) rather than to (D), it might be found implausible that it could be given the relevant reading. While the authors are divided about the severity of this worry, all of us are open to the possibility, subject to further investigation, that some participants may have nonetheless interpreted (~I but R) as an instance of metalinguistic negation.

A second problem for this proposal was raised by an anonymous reviewer for *Synthese*. They suggested that if participants read (~I but R) as a metalinguistic negation, they would not have been expected to prefer (I and R) over (~I but R) as we found in our first study. This is because, if (~I but R) is read in the way in question, then it expresses essentially the same pair of

beliefs as (I and R), just with a bit of rhetorical flourish and while emphasizing that (R) expresses a stronger claim than (I). Because of this, the most natural prediction might be that we should find no preference for one statement over the other. We find this to be an important point to consider, but a defender of the metalinguistic negation hypothesis might reply that participants could have favored (I and R) in this condition because it appeared to be a more direct and less rhetorically laden description of the case as they understood it. Again, we see this as a matter worthy of additional investigation and are open to further evidence that would prove this supposition to be wrong.

A second possible strategy for resolving our puzzle appeals to the way that the connective 'but' can imply a conceptual *contrast* between the statements that flank it. As an illustration consider the following, which modifies an example due to Grice (2001, p. 25):

(E) He is an Englishman, but he's brave.

The most natural reading of (E) is as saying, not only that the person it refers to is both an Englishman and brave, but that the person's being brave is somehow *unexpected* given that he is an Englishman. This sentence therefore invites treating the concept 'Englishman' as one with which bravery is not commonly associated. Other examples of this phenomenon abound. Here is one more, which implies that the traits listed following 'but' *are* ones that philosophers are especially likely to possess:

(P) Sarah's not a philosopher, but she's extremely intelligent, with a wealth of great ideas and the ability to articulate them clearly and argue persuasively for them.

It seems plausible that (~I but R) can be read as implying the kind of *contrastive* relationship suggested in (E) and (P), now between the concepts of *doing something intentionally* and *doing something knowingly and in a way that makes one morally responsible and blameworthy for it.*

This would explain how participants could agree with this statement while also treating (I and R) as saying something true, as the latter might have been read as trading on a different concept of intentional action.

The most plausible way this could have happened is if (~I but R) invoked a *narrower* reading of 'intentionally' than the reading of 'intentionally' invoked by (I and R). That is to say, (~I but R) may have been read as implying that doing something knowingly and in a way that makes one morally responsible and blameworthy for it is not *sufficient* for doing this intentionally in the relevant narrow sense, perhaps because doing something intentionally in this narrow sense requires either desiring to do it or choosing it as a means to an end. This narrow reading of 'intentionally' would contrast with a *wide* reading, plausibly invited by the phrasing of (I and R), on which it suffices to do something intentionally if one does it knowingly and in a way that incurs moral responsibility and blame.

Along these lines, a number of authors have argued that 'intentionally' is polysemous (e.g., Nichols and Ulatowski 2007, Cushman and Mele 2008, Lanteri 2013; see Cova 2016, Section 8.4, for discussion). And this includes evidence that the term can be read in either of the two ways just noted. For example, Nichols and Ulatowski (2007) asked participants to explain their reasons for saying that Knobe's chairman either had or had not intentionally helped or harmed the environment, and found that those who denied intentionality tended to focus in their explanations on the chairman's *intent* or *motivation*, while those who attributed intentionality tended to focus on his having *known* about the effect that his policy was going to have. While Nichols and Ulatowski interpret these data as revealing individual differences in ordinary concepts of intentionality, it could be that in fact many people are able to use 'intentionally' in either of these two ways depending on the wider context, and that the difference between (~I but

R) and (I and R) supplies a difference in context that makes for the necessary difference in how 'intentionally' is understood.

By our lights, the most serious difficulty facing this proposal is that in light of the within-participants results from Study 2, it requires the *prima facie* surprising assumption that people are willing to use 'intentionally' in both of these ways side-by-side, employing the narrow interpretation on which doing something intentionally requires motivation or intent in reading (~I but R), and the wider interpretation on which it just requires knowledge of what one is doing in reading (I and R), even when both statements are presented as part of a single display. We are not, however, aware of any experimental evidence that rules out this possibility or even counts strongly against it, and so we leave this as a matter for further investigation.

We conclude that the studies presented in this paper clearly indicate that people are willing to agree with Lindauer and Southwood's cancelling statement (C) despite holding that the chairman intentionally harmed the environment. What is less clear is why. We've considered three speculative explanations, each with some initial plausibility and two that we think are worthy of serious consideration, but nevertheless facing potential problems as general stories. And it is also possible that both of these accounts—or perhaps some further ones we haven't discussed—are jointly at work, with some participants who hold that the chairman intentionally harmed the environment agreeing with (C) for one reason, others for another. Deciding between these accounts awaits further research. For now, we content ourselves with the conclusion that the Side-Effect Effect remains uncancelled.

References

Adams, Fred and Annie Steadman (2004). "Intentional action in ordinary language: core concept or pragmatic understanding?" *Analysis*, 64(2): 173–181.

Adams, Fred and Annie Steadman (2007). "Folk concepts, surveys, and intentional action." In C. Lumer and S. Nannini (eds.), *Intentionality, Deliberation, and Autonomy: The Action-Theoretic Basis of Practical Philosophy*, pp. 17–33, Aldershot: Ashgate Publishers.

Cova, Florian (2016). "The Folk Concept of Intentional Action: Empirical Approaches." In J. Sytsma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*, Wiley Blackwell, pp. 121–141.

Cova, Florian and Hichem Naar (2012). "Side-Effect Effect Without Side Effects: The Pervasive Impact of Moral Considerations on Judgments of Intentionality." *Philosophical Psychology*, 25: 837–854.

Cushman, Fiery and Alfred Mele (2008). "Intentional action: Two-and-a-half folk concepts?" In J. Knobe and S. Nichols (eds.), *Experimental Philosophy*, Oxford University Press, pp. 171–188.

Dalbauer, Nikolaus and Andreas Hergovich (2013). "Is What Is Worse More Likely? The Probabilistic Explanation of the Side-effect Effect." *Review of Philosophy and Psychology*, 4: 639–657.

Grice, Paul (2001). "Logic and conversation." In *Studies in the Way of Words*, pp. 22–40, Cambridge: Harvard University Press.

Hitchcock, Christopher and Joshua Knobe (2009). "Cause and Norm." *Journal of Philosophy*, 11: 587–612.

Horn, Larry (1985). "Metalinguistic negation and pragmatic ambiguity." *Language*, 61(1): 121–174.

Knobe, Joshua (2003a). "Intentional action and side-effects in ordinary language." *Analysis*, 63(3): 190–194.

Knobe, Joshua (2003b). "Intentional action in folk psychology: an experimental investigation." *Philosophical Psychology*, 16(2): 309–323.

Knobe, Joshua (2004). "Intention, intentional action, and moral considerations." *Analysis*, 64(2): 181–187.

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

Knobe, Joshua (2010). "Person as scientist, person as moralist." *Behavioral and Brain Sciences*, 33(4): 315–329.

Knobe, Joshua and Arudra Burra (2006). “Intention and Intentional Action: A Cross-cultural Study.” *Journal of Culture and Cognition*, 6(1-2): 113–132.

Lanteri, Alessandro (2013). “Three-and-a-half Folk Concepts of Intentional Action.” *Philosophical Studies*, 158: 17–30.

Lindauer, Matthew and Nicholas Southwood (2021). “How to Cancel the Knobe Effect: The Role of Sufficiently Strong Moral Censure.” American Philosophical Quarterly, 58(2): 181–186.

Mizumoto, Masaharu (2018). “A Simple Linguistic Approach to the Knobe Effect, or the Knobe Effect without any Vignette.” *Philosophical Studies*, 175: 1613-1630.

Nadelhoffer, Thomas (2004). “On praise, side effects, and folk ascriptions of intentional action.” *Journal of Theoretical and Philosophical Psychology*, 24(2): 196–213.

Nichols, Shaun and Joseph Ulatowski (2007). “Intuitions and individual differences: the Knobe effect revisited.” *Mind and Language*, 22(4): 346–365.

Pettit, Dean and Joshua Knobe (2009). “The Pervasive Impact of Moral Judgment.” *Mind & Language*, 24: 586–604.

Phillips, Jonathan, Jamie Luguri, and Joshua Knobe (2015). “Unifying Morality’s Influence on Non-moral Judgments: The Relevance of Alternative Possibilities.” *Cognition*, 145: 30–42.