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

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. 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) detailed 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 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.'
>
> 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[1], languages[2], populations[3], and concepts[4] (for a review, see Cova 2016), although 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 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

---

[1] 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.

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

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

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

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 participants are given the opportunity to morally censure (e.g.) 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 effect 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 suitably mitigate the pragmatic pressure to say that he harmed the environment intentionally. To remedy this, in the crucial condition of their study participants read the harm version of Knobe's chairman vignette and then rated the following 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.

In contrast with this cancelling condition, participants in L&S's help and harm conditions read the corresponding version of Knobe's chairman vignette and then rated a *simple* statement about it, corresponding to Knobe's original test statement:

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

L&S predicted that the opportunity to express sufficiently 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 served to 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, which corresponds to the harm variant of their simple statement (S):

(~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.

There are, however, other possible explanations of L&S's findings that deny this key assumption. First, their participants might have expressed agreement with (C) for the same sort of reason that L&S themselves appeal to in explaining the original SEE—namely that participants felt *pragmatic pressure* to censure the chairman for harming the environment, and since they were able to do this only by indicating their agreement with (C), they did this even though they disagreed with the first clause of the statement. Second, there are ways of reading a statement of the form 'not-A but B' on which it warrants assent even when the simple statement 'not-A' would not. (To illustrate, consider 'Shaq isn't big, but huge' and 'Jane didn't make dinner, but prepared a sumptuous feast that was a delight for the senses'.) For now we will postpone further discussion of the details and merits of these alternative explanations, as they will be the focus of our concluding section. The crucial thing to emphasize is that the viability of these alternatives undermines the support that L&S's findings are supposed to provide for the pragmatic account of the SEE. This is because, according to these alternative explanations, the fact that L&S's participants tended to agree with the cancelling statement (C) does *not* itself show that they believed that the chairman didn't harm the environment intentionally.

There are, however, several straightforward ways to modify L&S's experimental paradigm in order to probe their core assumption, namely by having participants evaluate judgments of intentionality and responsibility either independently or as parts of a clause 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 assumption, and to the predictions of the pragmatic account. In our concluding section, we discuss what to make of the state of play.

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, with the statements displayed in random order. For clarity, in each statement the connective was emphasized and negations were bolded, as shown below:

> 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, <u>*and*</u> 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, <u>*but*</u> 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, <u>*but*</u> 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, <u>*and*</u> 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, L&S's account generates different predictions from the alternative explanations we laid out above about what should result from the rank-ordering paradigm employed in the present study. In particular, since participants must engage with all four statements in order to produce a preferred ranking, they are able to register their moral censure of the chairman by showing a preference for *either* of the two statements affirming (R). In this context, there should therefore be little to no pragmatic pull to show a preference for the statements affirming (I), as agreement with these statements is no longer needed in order to express moral censure. Given this, the pragmatic account predicts that participants will tend to show a preference for (~I but R) over the other three statements, given that they agree independently with both (~I) and (R). By contrast, the alternative explanations, on which participants agree more with (I) than with (~I), yield the prediction that participants will instead show a preference for (I and R) over (~I but R).

Each study in this paper was conducted online with participants recruited through advertising for a free personality test on Google in North America.[5] 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. In line with the alternative explanations of L&S's findings, but 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

---

[5] One notable benefit of using a "push strategy" like this one (i.e., recruiting participants who were not directly looking to participate in research) is that participants are more likely to be "experimentally naïve" and less likely to be motivated to provide the responses that they think the experimenters are looking for (Haug 2018). Samples collected using the recruitment strategy employed here have been previously compared against samples collected with other methods in replication studies. And the present strategy has been consistently found to generate a diverse sample in terms of geography, socio-economic status, religiosity, political orientation, age, and education. Studies using this strategy have been previously reported in publications including, e.g., Livengood et al. (2010, 2017), Sytsma (2010, 2012), Sytsma & Livengood (2011, 2021), Feltz & Cokely (2011), Murray et al. (2013), Machery et al. (2015), Livengood & Rose (2016), Kim et al. (2016), Sytsma & Reuter (2017), Fischer et al. (2021).

[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, compared to 20.9% for (~INT but RESP): 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. (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

intentionally. As such, the present findings suggest that agreement with the cancelling statement in L&S's study does *not* in fact reveal that people tend to believe that the chairman did not intentionally harm the environment.

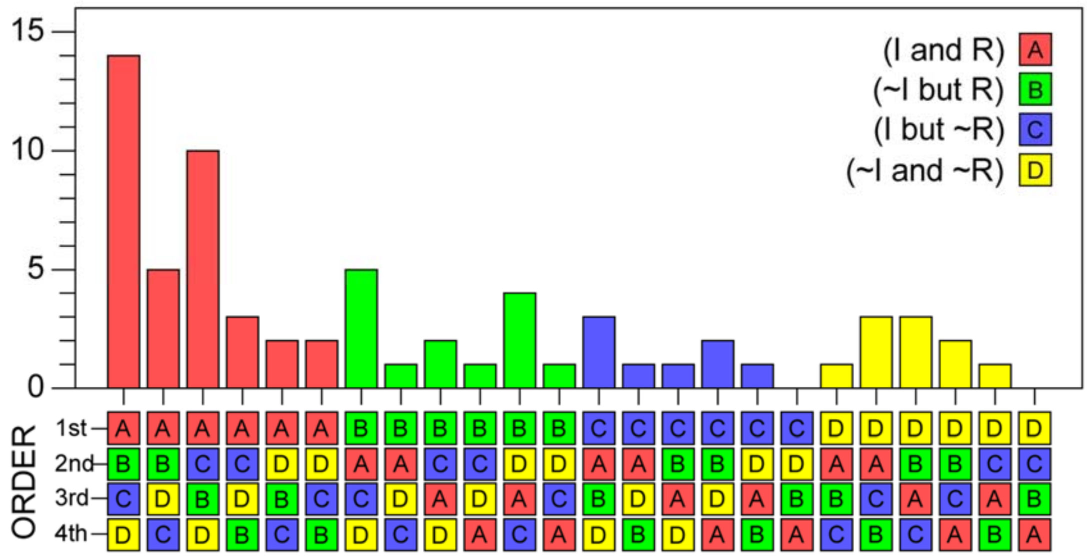

Figure 1: Histogram for each rank ordering.

3. *Study 2: Likert Rankings*

To further test whether agreement with L&S's cancelling statement should be taken to indicate independent agreement with both (~I) and (R), 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'). In each condition participants read the harm version of the chairman case and then rated one or more of the four statements. For comparison, agreement ratings were solicited both within-participants (each participant rating all four statements in random order) and between-participants (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]

As we have seen, the pragmatic account of the SEE contends that people tend to say that the chairman intentionally harmed the environment because they have no other way of expressing strong moral censure of him. In line with this, L&S interpret their results as showing that participants tended to express agreement with their cancelling statement (C) because they agreed with *both* (~I) *and* (R). This account generates several predictions about how the statements (I and R) and (~I but R) should be rated our Study 2. Let us consider the within-participants condition first:

(i) Given that participants in this condition are able to rate both (I and R) and (~I but R), then on the assumption that the first conjunct of the former statement is believed to be literally false, it seems plausible that participants should disagree overall with this statement, as there is no pragmatic pressure to express agreement with it.

(ii) Alternatively, if a defender of the pragmatic account wished to hold that there might still be *some* pressure to agree with (I and R) given the way it expresses moral censure of the chairman, she probably should concede that this statement should at least receive *lower* agreement than (~I but R), as only the latter is literally true.

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

(iii) Finally, if the defender tried to dig in and deny even this much, then she must concede *at the very least* that ratings of (~I but R) should *be no lower* than those of (I and R), since both express moral censure and only the latter is literally true.

A further prediction that follows from the pragmatic account concerns a difference in how the statement (I and R) should be rated in each of our two conditions:

(iv) Since (I and R) is supposed to say something that is literally false and participants agree with it only as a way to express moral censure, agreement with this statement should be higher in the between-participants condition than in the within-participants condition, as the latter condition affords other ways to express moral censure, thus relieving at least some of the supposed pragmatic pressure.

Unfortunately for defenders of the pragmatic account, none of these predictions were borne out by the data.

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. Whether the statements were presented individually (between-participants) or as a group (within-participants) did not make a statistically significant difference in the rating of any of the four statements—including for ratings of (I and R), contrary to prediction (iv) above.[11] Focusing on the within-participants condition, against the strong prediction (i) participants not only didn't tend to disagree with (I and R), but tended to agree with it, with the mean rating being significantly above the neutral point.[12] Further, contrary to the weaker prediction (ii), ratings were higher for (I and R) than for

---

[11] (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) $t(93.45)=.67$, $p=.50$, $d=.13$

[12] $t(49)=4.85$, $p<.001$ (one-tailed), $d=.69$

(~I but R), and significantly so when combining conditions.[13] Thus, contrary to the weakest prediction (iii), the mean rating for (I and R) was *not* at least as high as for (~I but R). This is, then, further strong evidence against the pragmatic account.

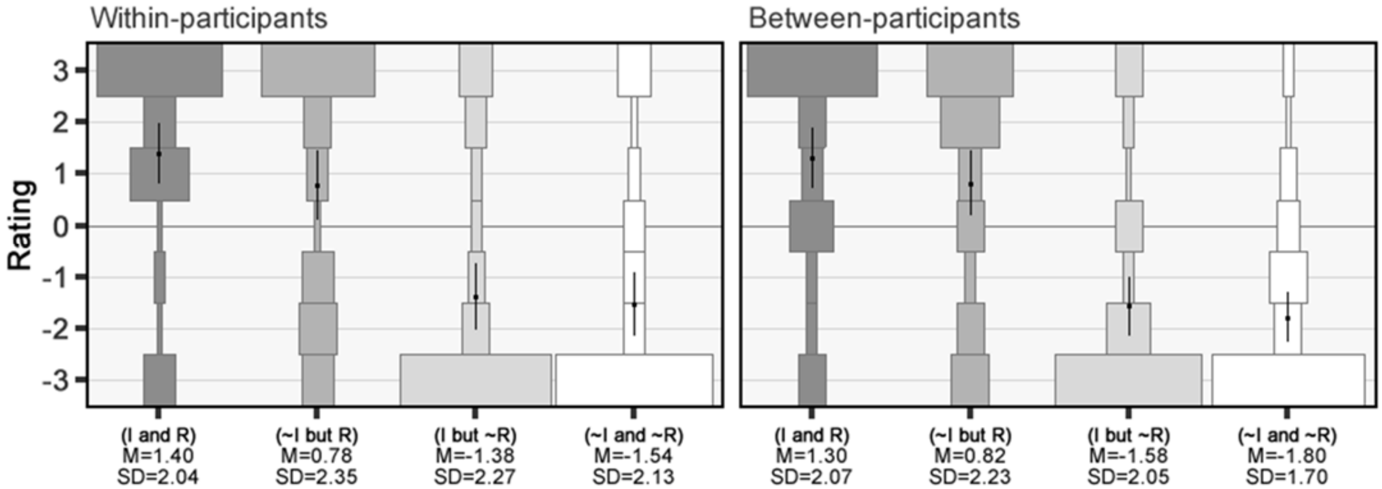

Figure 2: Results for Study 2, showing relative percentage of participants selecting each response option for each statement, with means and 95% confidence intervals overlaid, and split between the within-participants condition (left) and between-participants conditions (right).

4. *Study 3: Separate Statements*

According to Lindauer and Southwood, while people find Knobe's chairman to be deserving of strong moral censure for harming the environment, they do not 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 suggest against this. As a final

[13] We used a partially paired t-test to compare (I and R) with (~I but R) across conditions: $t(125.63)=1.84$, $p=.034$ (one-tailed) [within-participants: $t(49)=1.52$, $p=.068$ (one-tailed), $d=.28$]

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 Study 2, both a simple statement that attributed responsibility to the chairman and a simple statement that said 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 L&S's account, it seems that participants should therefore tend to disagree overall with (I), since they had already been able to censure the chairman by expressing agreement with (R).

Responses were collected from 54 participants using the same recruitment strategy and restrictions as in the previous studies.[14] The results are shown in Figure 3. Against the prediction of L&S's account, participants continued to agree with (I) even when it was presented following (R), with the mean rating being significantly above the neutral point.[15] In fact, the mean rating for (I) was actually *higher* than the mean rating for (R)! Finally, there was a very high correlation between participants' ratings of the two statements, indicating that there was no trade-off between attributing responsibility to the chairman and saying that he harmed the environment

---

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

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

intentionally.[16] Indeed, 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). All of this is exactly the opposite of what is predicted by the pragmatic account of the SEE.

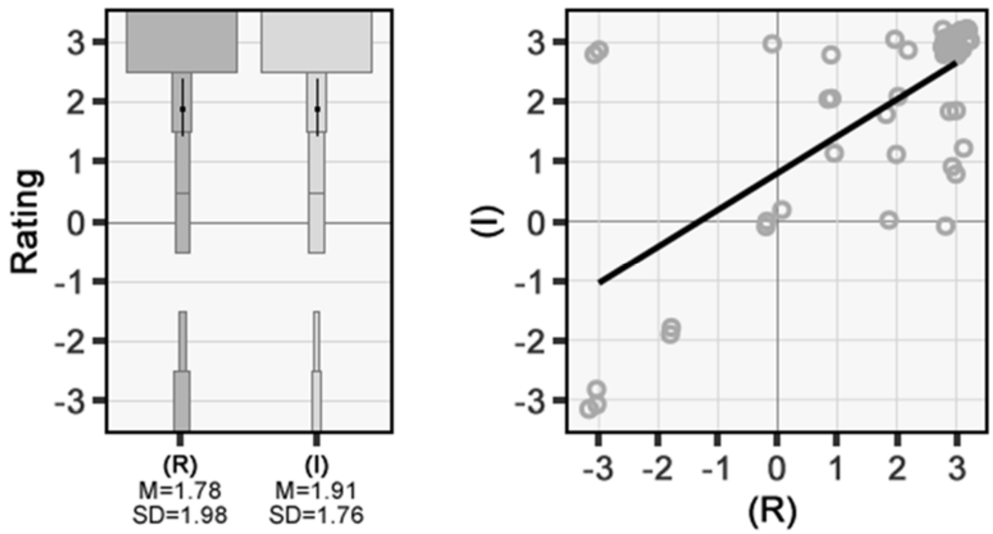

Figure 3: Results for Study 3, showing relative percentage of participants selecting each response option for each statement on the left, with means and 95% confidence intervals overlaid. The scatterplot on the right show points with jitter and regression line calculated without jitter.

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 core assumption that agreement 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

---

16 r=0.69, $t(52)$=6.92, p<.001

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 above provide ample evidence that this key assumption does not hold. Despite tending to agree with (C), most participants nonetheless agree that the chairman intentionally harmed the environment, even when they are 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 ourselves divided on the matter. Nonetheless, in closing we want to return to the alternative positions that we outlined very briefly in the introduction, considering them in the light of the results of our findings.

According to the first of the alternative explanations we outlined, L&S's original finding can be explained in just the way that the pragmatic account attempts to explain the SEE: the explanation says that participants expressed agreement with their cancelling statement (C) because doing so was the only way to express moral censure of the chairman. While this account may seem quite plausible—and indeed was the initial motivation for our studies—the results of our second study suggests that this can't be the whole story. Specifically, in the within-participants condition of Study 2 respondents had the opportunity to express moral censure by agreeing with (I and R), yet 60% of participants nevertheless gave (~I but R) a response above the neutral point. Indeed, exactly half of the participants in this condition expressed agreement with *both* (I and R) *and* (~I but R)! This finding appears to be incompatible with the simple 'pragmatic' explanation of why L&S's participants agreed with their statement (C). Yet it is also a puzzling finding, which itself calls out for explanation: it suggests that many participants 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) respectively, as *not* saying contradictory things about what the chairman did. How could that be? Below we will explore a couple of possible answers, each focusing on how the connective 'but' might influence the interpretation of a statement like (~I but R).

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 in a statement like 'Shaq isn't big, but huge' or 'Einstein wasn't a smart guy, but a genius'. This use of the phrase 'not … but', which the linguist Larry Horn (1985) refers to as *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 statements above while believing that Shaq is big, and Einstein was a smart guy.

However, a series problem with this proposal is that it is debatable whether (~I but R) has the syntax that is necessary to read 'but' as introducing metalinguistic negation. This is because, according to Horn (1985, p. 166), the phrase 'not … but' can serve to introduce metalinguistic negation only if 'but' is not followed by a sentential connective: thus 'Shaq isn't big, but he's huge', and 'Einstein wasn't a smart guy, but he was a genius', both strike the ear as puzzling. If this is correct, then in order for (~I but R) to receive the relevant reading, we would need to drop the 'he' following the 'but', as in:

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

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 read the 'but' in (~I but R) as an instance of metalinguistic negation, despite the sentential connective.

The other 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): 'He is an Englishman, but he's brave.' The most natural reading of this statement 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: '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.' And it seems plausible that (~I but R) can be read as implying this kind of *contrastive* relationship between the concepts of doing something intentionally and doing something knowingly and in a way that makes one morally responsible and blameworthy for it, thus explaining how participants could agree with this statement while also agreeing with (I and R), as the latter might have been read as trading on a different concept of intentional action.

The most likely 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). This could have happened if (~I but R) was 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' contrasts with a *wide* reading on which it suffices to do something intentionally if one does it knowingly and in a way that incurs moral responsibility and blame. And there is some experimental evidence supporting the hypothesis that 'intentionally' can be read in either of these two ways. For example, a study by Nichols & 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 intention, 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.

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 contend 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 offered three possible explanations, each with some initial plausibility, but all of them facing potential problems as general explanations. And it is also possible that 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 concluding 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.

Fischer, Eugen, Paul Engelhardt, and Justin Sytsma (2021). “Inappropriate stereotypical inferences? An adversarial collaboration in experimental ordinary language philosophy.” *Synthese*,

Feltz, Adam and Edward Cokely (2011). “Individual differences in theory-of-mind judgments: Order effects and side effects.” *Philosophical Psychology*, 24(3): 343–355.

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

Haug, Matthew (2018). “Fast, Cheap, and Unethical? The Interplay of Morality and Methodology in Crowdsourced Survey Research.” *Review of Philosophy and Psychology*, 9(2): 363–379.

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.

Kim, Hyo-eun, Nina Poth, Kevin Reuter, and Justin Sytsma (2016). “Where is your pain? A Cross-cultural Comparison of the Concept of Pain in Americans and South Koreans.” *Studia Philosophica Estonica*, 9(1): 136–169.

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.

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.

Livengood, Jonathan and David Rose (2016). “Experimental Philosophy and Causal Attribution.” In J. Sytsma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*, Wiley Blackwell, pp. 434–449.

Livengood, Jonathan, Justin Sytsma, Adam Feltz, Richard Scheines, and Edouard Machery (2010). “Philosophical Temperament.” *Philosophical Psychology*, 23(3): 313–330.

Livengood, Jonathan, Justin Sytsma, and David Rose (2017). “Following the FAD: Folk Attributions and Theories of Actual Causation.” *Review of Philosophy and Psychology*, 8(2): 274–294.

Machery, Edouard, Justin Sytsma, and Max Deutsch (2015). “Speaker’s Reference and Cross-cultural Semantics.” In A. Bianchi (ed.), *On Reference*, Oxford University Press, pp. 62–76.

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

Murray, Dylan, Justin Sytsma, and Jonathan Livengood (2013). “God Knows (But does God Believe?)” *Philosophical Studies*,166: 83–107.

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.

Sytsma, Justin (2010). “Dennett’s Theory of the Folk Theory of Consciousness.” *Journal of Consciousness Studies*, 17(3-4): 107–130.

Sytsma, Justin (2012). “Revisiting the Valence Account.” *Philosophical Topics*, 40(2): 179–198.

Sytsma, Justin and Jonathan Livengood (2011). “A New Perspective Concerning Experiments on Semantic Intuitions.” *Australasian Journal of Philosophy*, 89(2): 315–332.

Sytsma, Justin and Jonathan Livengood (2021). “Causal Attributions and the Trolley Problem.” *Philosophical Psychology*, 34(8): 1167–1191.

Sytsma, Justin and Kevin Reuter (2017). “Experimental Philosophy of Pain.” *Journal of Indian Council of Philosophical Research*, 34(3): 611–628.