Has the Side-Effect Effect Been Cancelled? (No, Not Yet.)

Justin Sytsma, Victoria University of Wellington
John Schwenkler, Florida State University
Robert Bishop, Florida State University

In the *Side-Effect Effect* (SEE), side effects that are regarded as morally bad are judged as more intentional than side effects that are neutral or morally good (Knobe 2003a). Investigation of the SEE has replicated the effect with different cases (Knobe 2003b, Nadlehoffer 2004, Knobe 2006, Cushman & Mele 2007), languages (Mizumoto 2018, Knobe & Burra 2006), and concepts (Beebe & Buckwalter 2010, Pettit & Knobe 2009).

There is, however, persistent disagreement about the explanation of the SEE. The first, most straightforward account is that the effect arises because participants mean just what they say: the effect reflects their *beliefs* about whether certain things are done intentionally or not. But an alternative account claims that the asymmetry arises due to pragmatic pressure to express moral censure. The pragmatic account holds that, since a person is blameworthy for the harmful side-effects of their actions in a way that they are not praiseworthy for side-effects that are beneficial, participants express agreement with the statement that the harmful side-effects were intentional in order to avoid implying that they are excusing them, whereas the same pressure is not felt in connection with neutral or beneficial side-effects. If the pragmatic account is correct, it should be possible to cancel the SEE by giving participants a different way of attributing moral responsibility.

A forthcoming paper by Lindauer and Southwood (hereafter 'L&S') purports to do just this. In the crucial condition of their study, participants read the harm version of Knobe's (2003a) 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.

In the other conditions, participants read either the help or harm version of the chairman vignette and rated one of the following *simple* statements instead:

(S_{help}) The chairman didn't intentionally help the environment.
(S_{harm}) The chairman didn't intentionally harm the environment.

L&S's main finding was that, while ratings of (S_{help}) and (S_{harm}) exhibited the usual SEE, their participants agreed with (C) to about the same extent as they agreed with (S_{help}). That is, participants who read the version of the vignette in which the chairman had *harmed* the environment agreed just as much with (C), which says that this harm was unintentional, as participants who read the version in which the chairman had *helped* the environment agreed with

(S_{help}). L&S take this finding to provide evidence for the pragmatic account of the SEE. Our paper shows why it does not.

To begin, notice that L&S's cancelling statement (C) has two parts, which are separated by the contrastive conjunction 'but'. The first part of (C) is the same as the statement (S_{harm}) above, which denies that the chairman harmed the environment intentionally. And the second part of (C) is a positive attribution of moral responsibility to the chairman:

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

According to L&S, their participants agreed with (C) because they agreed independently with both (S_{harm}) and (R), and the opportunity to express their agreement with the latter statement relieved what had been merely pragmatic pressure to deny the former. To explore whether this is the correct account of these findings, we conducted several experiments to test a further prediction that follows from L&S's account, namely that participants who are allowed to censure the chairman should prefer saying that he did *not* harm the environment intentionally over saying that he *did* harm the environment intentionally—since the first thing is supposed to be what they really believe. That is, Lindauer and Southwood's pragmatic account predicts that *participants who are given the opportunity to express strong moral censure of the chairman, by affirming a statement like (R), should tend to disagree overall with (S_{harm})*. In each of our experiments, this prediction was not borne out.

In our first study, participants read the harm version of the chairman vignette and then rank-ordered four variants of L&S's cancelling statement (C), including the following:

(C1) The chairman intentionally harmed the environment, *and* he knowingly harmed the environment, is morally responsible for doing so, and should be blamed for it.
(C2) 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.

As we have seen, L&S's pragmatic account predicts that participants should rank (C2) higher than (C1): for the final three clauses of each statement remove any pragmatic pressure to censure the chairman, and in this context participants should feel free to say that he did not harm the environment intentionally. But we found just the opposite, as 65.7% of our participants ranked (C1) higher than (C2). This pattern was statistically significant: χ^2=5.97, p=.015.

In our second study, participants read the same vignette and then rated the four statements from our first study on a 7-point scale. Again, L&S's account predicts that (C2) will elicit higher ratings than (C1), at least when participants are given the opportunity to rate both of the statements. Again, we found the opposite, as mean ratings of (C1) were significantly higher than mean ratings of (C2): t(195.76)=1.80, p=.037, d=.25. This same pattern was observed for both a

Deleted: But there are several reasons to doubt this inference, including that the statement (C) introduces pragmatic pressures of its own and that the inference ignores the various roles that the discourse connective 'but' can play. To explore these matters

Deleted: crucial prediction of L&S's

Deleted: We tested this prediction in three separate studies

Deleted: the following

Deleted: statements (C1) and (C2)

between-participants design, in which each participant rated only one statement, and a within-participants design, in which each participant rated all four.

Finally, in our third study participants read the same vignette and then first indicated their level of agreement with the responsibility-attributing statement (R), after which they indicated their level of agreement with the simple statement (S_{harm}), which denies that the chairman harmed the environment intentionally. Again, L&S's account predicts that in this condition participants should tend to disagree with (S_{harm}), since the opportunity to express agreement with (R) removes pragmatic pressure to censure the chairman and allows them to say what they really think, namely that he did not harm the environment intentionally. And once again we found just the opposite, as mean ratings of (S_{harm}) were significantly above the neutral point: $t(53)=6.61$, $p<.001$, $d=.90$.

Taken together, these findings suffice to refute Lindauer and Southwood's account of their (forthcoming) findings. While participants in their experiment expressed higher agreement with their cancelling statement (C) than with their simple statement (S_{harm}) this was *not* simply because they agreed with what (S_{harm}) says on its own. The Side-Effect Effect remains uncancelled.

Deleted: or the other

Deleted: them both