Mutual Entailment Between Causation and Responsibility

Justin Sytsma, Pascale Willemsen, and Kevin Reuter

Abstract: The standard view in philosophy is that responsibility entails causation. Most philosophers treat this entailment claim as an evident insight into the ordinary concepts of responsibility and causation. Further, it is taken to be equally obvious that the reversal of this claim does not hold: causation does not entail responsibility. In contrast, the account of ordinary causal attributions put forward by Sytsma and Livengood predicts that "responsible for" and "caused" will generally be taken to apply in the same contexts. If the responsibility account is correct, then the reversal of the entailment claim may hold, and, a fortiori, there would be mutual entailment between the ordinary concepts of responsibility and causation. Using the cancellability test, we report the results of three pre-registered studies providing empirical evidence that causation and responsibility are mutually entailed by each other.

Keywords: Entailment Claim; Causation; Responsibility; Experimental Philosophy; Causal Cognition.

Sartorio (2007, p. 750) writes, "here's a very natural idea about the relation between moral responsibility for outcomes and causation: moral responsibility for an outcome requires causing it". This expresses a common claim about the relationship between responsibility and causation in philosophy. Call this the Entailment Claim. This claim holds that causing an outcome is a necessary, conceptual precondition for being morally responsible for that outcome (see also, Driver 2007, Sartorio 2007, Scanlon 2008, Wolf 1993). The Entailment Claims seems plausible. Afterall, it seems that we can hardly blame a person for a broken window, if that person did not cause the window to break.[1] However, causation is standardly thought to be insufficient for

[1] Note that we often form negative judgments of people for merely *trying* to do something, even if they don't succeed in bringing about the intended outcome. And such judgments might include blame. If so, it could then be argued that in such cases we blame a person in the absence of any causal involvement, as their action didn't cause the intended outcome. We set this aside here, however, as the Entailment Claim we're after is only concerned with moral responsibility for *outcomes*. We therefore also ignore cases that are called conduct crimes in the legal system and are "defined without regard to any result of the defendant's conduct (for example, attempts, conspiracy, burglary) [where] there is no need to face the issue of causation" (Kadish et al. 2012).

responsibility: even if a person caused the window to break, we might not judge that they are responsible for that outcome, e.g., if the person was not in control of what they were doing.

Philosophers have proposed different versions of the Entailment Claim, with some of them including or excluding omissions, making additional grounding claims, and so on.[2] There are four key claims that unite most of these proposals. First, the concept of causation at issue is not supposed to be a technical one, but to reflect the dominant, ordinary concept of causation. Some authors commit explicitly to the idea that a useful concept of causation must capture our ordinary understanding. In defending causation by omissions, for instance, Schaffer (2004, p. 205) states that "to dismiss negative causation is to swallow" that "the folk are wrong that voluntary human action is causal, the law is wrong that negligence is causal, ordinary language is wrong that 'remove', 'release', 'disconnect', and so on are causal"; based on this, he goes on to "submit that no theory so dismissive deserves to be considered a theory of causation". The close connection between theories of causation and our ordinary understanding of causation is further illustrated by the plethora of thought experiments employed in the literature to elicit our causal intuitions.[3]

Second, philosophers often assume, whether implicitly or explicitly, that the ordinary concept of causation is expressed most fundamentally by the attributional use of the lemma "cause" (e.g., Driver 2007, Sartorio 2007, Skow 2019).[4] Third, it is typically taken for granted

[2] Taking omissions to be causally impotent, it has sometimes been suggested that omissions constitute a counterexample to the Entailment Claim. While some metaphysical theories of causation exclude the possibility of causation by omissions, those adopting such a view usually conclude that we cannot be blamed for the outcomes of our omissions. Others, such as Driver, explicitly include omissions in the Entailment Claim: "if an agent A is morally responsible for an event e, then A performed an action or omission that caused e" (2007, p. 423).

[3] It should be noted that not all philosophical work on causation targets the ordinary concept, although much does. Further, even for work that aims to develop a technical concept, the ordinary concept often remains relevant (Hall & Paul 2003).

[4] This includes claims such as "X is the *cause* of Y" and "X *caused* Y". Most experimental work has followed the theoretical literature and investigated people's judgments about these causal attributions. However, researchers have also used other constructions, such as statements employing "made" (Samland et al. 2016) or "because" (Livengood

that there is an important conceptual and metaphysical distinction between causation and moral responsibility, such that it makes sense and seems philosophically fruitful to investigate their relationship. Fourth, and relatedly, it is generally assumed that while normative considerations play a central role in the applicability of the concept of moral responsibility, the concept of causation is independent of such considerations, merely describing the causal chain that led to the outcome.[5] In other words, the concept of causation is taken to be purely descriptive. Indeed, this is a cornerstone in these debates, as causation is often thought of as grounding moral responsibility. The possibility of such a grounding relation, however, requires the independence of the grounding and the grounded (see Sartorio 2007). Thus, whatever norms are at play in (correctly) applying the concept of moral responsibility, they should play no such role for (correctly) applying the concept of causation.

Taken together, these four claims inform the standard view of the relationship between causation and moral responsibility that we are concerned with here:

> The standard view holds that causation is a necessary, but not sufficient, precondition for moral responsibility, and assumes that the concept of causation at issue corresponds with the ordinary concept, is canonically expressed by attributional uses of the lemma "cause", is a distinct concept from moral responsibility, and is purely descriptive.

We are skeptical of this standard view.

One reason to doubt the standard view comes from recent work in the experimental literature on causal attributions. This research has shown that normative considerations, including moral considerations, have a notable influence on the causal attributions people

and Machery 2007, Kominsky et al. 2015, Livengood et al. 2017). Other work has investigated causatives like "break" and "burn" (Rose et al. 2021, Schwenkler & Sievers forthcoming) and potential synonyms of "cause" (Sytsma et al. 2019). The results so far seem to be mixed in that some authors highlight interesting differences, while others suggest that the differences are minor.

[5] Psychological models share these latter two assumptions, taking judgments about causation to precede those of moral responsibility (e.g., Fincham Shultz 1981, Heider 1958, Schleifer et al. 1983, Shaver 1985).

endorse.[6] A number of different explanations of these findings have been put forward, typically attempting to square the effect with the assumption that the ordinary concept of causation is purely descriptive.[7] One leading explanation takes another tact, however. The responsibility account, first put forward in Sytsma et al. (2012), denies that the ordinary concept of causation is purely descriptive, instead contending that the dominant attributional use of the lemma "cause" expresses a normative concept akin to concepts like "responsibility".[8]

Here the responsibility account pushes against another piece of philosophical orthodoxy, beyond the key claims uniting the standard view noted above. It is commonly assumed that there are multiple ordinary concepts of responsibility. Most importantly, philosophers often distinguishing between a normative concept of *moral responsibility* and a purely descriptive concept of *causal responsibility*. Sytsma (forthcoming-b) raises doubts about this distinction, suggesting that the dominant attributional use of "responsible" is normative, but not necessarily moral. The idea is that attributing responsibility ordinarily goes beyond the purely descriptive, but that the norms at play might fall short of what one is inclined to label as moral. For present purposes, the key point is that while the responsibility account focuses on responsibility attributions, its proponents understand such attributions in a way that aligns with the concept of moral responsibility at issue for the Entailment Claim. We return to this issue in Section 2.

There is now a good deal of evidence supporting the responsibility account, including studies indicating that people's judgments about causal attributions (statements like "X caused

[6] See, for example, Alicke (1992), Knobe & Fraser (2008), Hitchcock & Knobe (2009), Sytsma et al. (2012), Reuter et al. (2014), Kominsky et al. (2015), Henne et al. (2017), Willemsen (2017), Livengood et al. (2017), Kominsky & Phillips (2019), and Livengood & Sytsma (2020), among many others.

[7] This includes explanations that focus on how people think about counterfactuals (e.g., Hitchcock & Knobe 2009, Halpern & Hitchcock 2015, Kominsky et al. 2015, Icard et al. 2017, Kominsky & Phillips 2019), explanations that contend the desire to blame or praise biases people's causal judgments (e.g., Alicke 1992, 2000; Alicke et al. 2011; Rose 2017), and explanations that hold that the experimental results are due to pragmatic factors (e.g., Samland et al. 2016, Samland & Waldmann 2016).

[8] See Sytsma (forthcoming-a) for a recent survey of the evidence for this view relative to other accounts.

Y”) are quite similar to their judgments about normative claims like responsibility attributions (statements like “X is responsible for Y”) and blame attributions (statements like “X is to blame for Y”).[9] If this is correct, it indicates that the ordinary concepts of responsibility and causation are much more similar than the standard view supposes. And this in turn suggests not only that people will tend to treat responsibility as entailing causation (the Entailment Claim that the standard view supposes) but that they’ll also tend to treat causation as entailing responsibility (the Reverse Entailment Claim the standard view denies). In other words, the responsibility account suggests that the conceptual relationship between responsibility and causation is one of *mutual entailment*.

In the next two sections, we present the results of three studies designed to test the twin entailment claims comprising mutual entailment. In all three studies, we take a rather novel methodological approach. While most of the experimental work on the ordinary concept of causation features vignette-based studies, we instead use the cancellability test in this paper. Although the cancellability test is widely known in the philosophical and linguistic literature, it has only recently been applied to *empirically* investigating semantic-cum-pragmatic relations between concepts in philosophy (Willemsen & Reuter 2021, Baumgartner et al. 2022, Coninx et al. forthcoming, Almeida et al. ms).

1. Study 1: Testing Mutual Entailment Between Responsibility and Causation

The cancellability test is used to examine whether a feature or component is conversationally implicated by another concept (Grice 1989). For instance, by saying “I tried to publish a book”, a

[9] See Sarin et al. (2017), Murray and Lombrozo (2017), Grinfeld et al. (2020), Sytsma & Livengood (2021), and Sytsma (2021, forthcoming-b).

speaker usually conveys the additional information that they failed to do so. However, canceling this derived piece of information does not result in a contradictory statement: “I tried to publish a book, but by that I am not saying that I failed to do so” sounds perfectly fine since the speaker might simply want to highlight the attempt. However, some pieces of information cannot be canceled in the same way. For instance, saying “Tom is a bachelor, but by that I am not saying that he is unmarried” is contradictory, as being unmarried is semantically entailed by being a bachelor.

In the three studies in this paper, we used the cancellability test to investigate mutual entailment, assessing the competing claims of the standard view and the responsibility account. In our first study, we assume the contention from advocates of the responsibility account noted above that the dominant ordinary concept of responsibility is a normative concept, testing cancellation statements involving “caused” and “responsible”.

2.1 Materials, Hypotheses, and Participants

Each participant in our study judged whether each of five statements was contradictory. This set was comprised of two key test statements—one to test the Entailment Claim (EC) and one to test the Reverse Entailment Claim (REC)—and three comparison statements (Control, Semantic Entailment, Conversation Implicature). Each statement was prefaced by telling participants to “please imagine that Sally said the following sentence”. After each statement they were asked “Does Sally contradict herself?” and answered using a 9-point scale anchored at 1 with “definitely not” and at 9 with “definitely yes”. The five statements were presented in random order and were preceded by a short training round explaining the notion of contradiction and giving participants two practice questions. Methodology and hypotheses were pre-registered at the Open Science Framework and full materials can be accessed through the online repository.[10]

The EC, REC, and Control statements each involve a causal attribution concerning one of three agents bringing about a different outcome:

(i) John caused the file to be deleted
(ii) Brian caused the patient to get worse
(iii) Steve caused the window to break

These attributions were assigned randomly, such that each participant received a statement using each of the three without repetition. For the EC statements, responsibility is asserted and causation is denied. To illustrate, for (i) the corresponding EC statement is:

> John is responsible for the file being deleted, but by that I am not saying that John caused the file to be deleted.

The REC statements reverse this, with causation being asserted and responsibility denied. To illustrate, for (ii) the corresponding REC statement is:

> Brian caused the patient to get worse, but by that I am not saying that Brian is responsible for the patient getting worse.

The purpose of the Control statements is to test whether participants treat just any attribution as being entailed by the causal claim. To do this, the Control statements assert causation (like the REC statements), but now deny that the agent *wanted* that outcome to occur. To illustrate, for (iii) the corresponding Control statement is:

> Steve caused the window to break, but by that I am not saying that Steve wanted the window to break.

Since an agent causing an outcome does not necessarily mean that the agent wanted that outcome to occur, we predicted that responses to the Control statements should be low.

[10] Preregistration at https://osf.io/hx735?view_only=f2134153eaa147f486bbdd7a1791956e; online repository at https://osf.io/kv9rt/?view_only=951000ffc7764e11a13c129104cc6d53

The other two statements—Semantic Entailment and Conversational Implicature—were used to set a high and low baseline for comparison, respectively. Semantic Entailment reads as follows:

> This is a lake, but by that I am not saying that it consists of water.

This statement gives an example where the assertion and denial are contradictory on the dominant use of the terms at issue. As such, we expected contradiction ratings to be high. By contrast, Conversational Implicature gives an example where what is denied is implied by the person making the assertion, but is not entailed by the statement:

> This chocolate is good value-for-money, but by that I am not saying that we should buy it.

Here we expected that contradiction ratings should be low.

There are three key null hypotheses for the EC and REC test statements that are relevant to testing the standard view and the responsibility account:

> **Hypothesis 1:** *Average contradiction ratings for EC statements are not significantly above the midpoint.*
>
> **Hypothesis 2:** *Average contradiction ratings for REC statements are not significantly above the midpoint.*
>
> **Hypothesis 3:** *No significant difference in contradiction ratings for EC statements and REC statements.*

As discussed above, both the standard view and the responsibility account hold that responsibility entails causation on the dominant ordinary concepts. This means that both views predict that people will tend to judge that the EC statements are contradictory, since responsibility is asserted while causation is denied. As such, both views expect Hypothesis 1 to be rejected. By contrast, the standard view and the responsibility account diverge with regard to whether causation entails responsibility: the responsibility account holds that it does, while the standard view denies this. This means that while the responsibility account predicts that people will tend to judge that the REC statements are contradictory, the standard view predicts that people will tend to judge that they are non-contradictory. As such, the responsibility account predicts that Hypothesis 2 will be rejected while the standard view predicts that it will not be rejected. It follows from these predictions that advocates of the standard view expect rather different responses for EC and REC statements, while advocates of the responsibility account expect these to be similar. As such, the standard view predicts that Hypothesis 3 will be rejected while the responsibility account predicts that it will not be rejected.

Participants were recruited through Prolific and reimbursed for their participation (pre-selection criteria: Approval Rate >90%, Native Language English, Age 18). Results were collected from 71 participants (62.0% women, one non-binary, average age 31.0 years).

2.2 Results

Responses for the two practice questions indicated that participants understood the idea of a speaker contradicting herself.[11] The mean contradiction ratings for the five questions in the main study are depicted in Figure 1 along with the distribution of responses. The baseline and control conditions worked as expected, with Semantic Entailment having a mean contradiction rating significantly above the mid-point of 5 [*t*(70)=7.33, *p*<.001, *d*=.87], while Conversational Implicature and Control had mean ratings significantly below the mid-point [*t*(70)=-11.80, *p*<.001, *d*=1.40; *t*(70)=-15.32, *p*<.001, *d*=1.82].

[11] The mean rating for the contradictory statement were significantly above the midpoint [M=8.31, SD=1.86, *t*(70)=14.96, *p*<.001 (one-tailed), *d*=1.78], while the mean rating for the non-contradictory statement was significantly below the mid-point [M=1.99, SD=1.92, *t*(70)=-13.03, *p*<.001 (one-tailed), *d*=1.57].

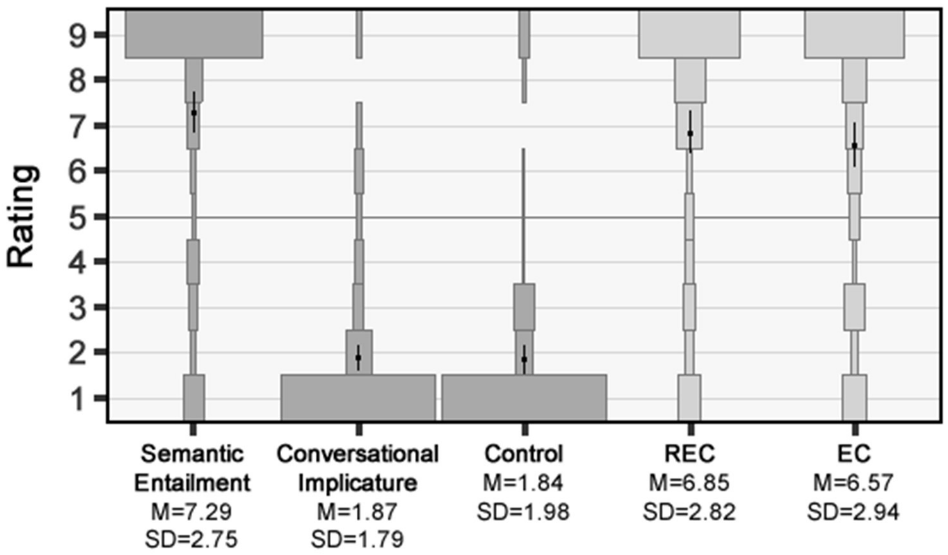


Figure 1: Results of Study 1. Plots show the relative percentage of participants selecting each response option, with means (dots) and 95% confidence intervals overlaid (error bars).

One-way ANOVAs did not show a significant effect for the different causal attributions—(i), (ii), and (iii) from above—for either REC [$F(2,68)=1.24$, $p=.30$, $\eta^2=.035$] or EC [$F(2,68)=.33$, $p=.72$, $\eta^2=.01$]. As such, we collapsed the data for the different outcomes in the subsequent analysis. In line with the predictions of both the standard view and the responsibility account, the EC statements had a mean contradiction rating significantly above the mid-point [$t(70)=3.44$, $p<.001$ (one-tailed), $d=.41$]. Thus, we can reject Hypothesis 1. More importantly, in line with the prediction of the responsibility account, but contrary to the prediction of the standard view, the mean contradiction rating for the REC statements was also significantly above the mid-point [$t(70)=4.67$, $p<.001$ (one-tailed), $d=.55$]. Thus, we can also reject Hypothesis 2. Indeed, ratings for the REC statements were not significantly different from the upper baseline given by Semantic Entailment [$t(70)=1.52$, $p=.13$, $d=.25$]. Finally, in line with the prediction of the responsibility account, but contrary to the prediction of the standard view, we found no significant difference between the mean ratings for the EC and REC statements [*t*(70)=.86, *p*=.40, *d*=.13]. Thus, we cannot reject Hypothesis 3.

2.3 Discussion

The responsibility account suggests that the relationship between the dominant ordinary concepts of responsibility and causation is one of mutual entailment. This generates the predictions that people should tend to treat the EC statements in our study as contradictory (rejecting Hypothesis 1), that they should tend to treat the REC statements as contradictory (rejecting Hypothesis 2), and further that the responses for these types of statements should be largely similar (affirming Hypothesis 3). In contrast, the standard view holds that the relationship is instead just one of unidirectional entailment, with responsibility entailing causation, but not the reverse. As such, the standard view makes competing predictions to the responsibility account for Hypothesis 2 and Hypothesis 3. The results of our first study bore out all three predictions of the responsibility account and, thus, ran counter the second and third predictions of the standard view. This pattern of findings supports mutual entailment.

As noted above, however, this conclusion assumes that people will tend to understand the responsibility attributions in our EC and REC statements as being normative. And while there is reason to expect this to be the case (Sytsma et al. 2019, Sytsma forthcoming-b), the claim is controversial. As we've seen, advocates of the standard view typically distinguish between a normative concept of responsibility ("moral responsibility") and a descriptive concept ("causal responsibility"). This distinction sets up a ready response to our first study: if participants did not interpret the responsibility attributions in terms of a relevant normative concept, then our results would not in fact support mutual entailment. Our final two studies address this concern.

3. Studies 2a & 2b: Testing Mutual Entailment Between Blame/Fault and Causation

It might be objected that our first study suffers from a major flaw, namely that we leave the responsibility attributions underspecified. The result, so the objection goes, is that the REC statements from our first study might have triggered readings of "responsible" that are irrelevant to the discussion at hand.

The first issue is that normative responsibility can be forward- and backward-looking. In a backward-looking sense, an agent might be morally responsible for something she did. In this case, we are referring to responsibility in the sense of blame and praiseworthiness. However, she might also be normatively responsible in a forward-looking sense that is not concerned with blame and praise but with future duties, such as when we say, e.g., "it is my responsibility as faculty member to attend the departmental meetings". Note, however, that duties are usually not specified for outcomes, and especially not for negative outcomes like making a patient get worse, breaking windows, or deleting files. Instead, duties are usually connected to *activities* that bring about *desired* outcomes. As such, it seems quite unlikely that participants interpreted the responsibility attributions in our REC statements in terms of duties.

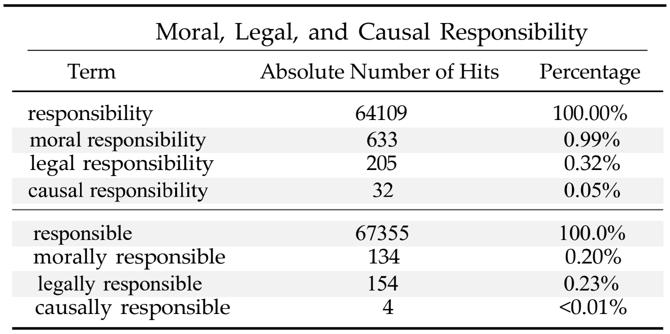
But even if we can be certain that participants took a backward-looking approach to responsibility, a second issue arises due to the potential ambiguity above: defenders of the standard view standardly distinguish between causal responsibility and moral responsibility. A person is considered causally responsible if "she is the (or a) salient cause of—some occurrence or outcome" (Talbert 2019). In other words, causal responsibility is a close cousin, if not even synonymous (at least on some accounts) with the notion of causation (see also Willemsen 2018 for discussion). If the participants in our first study interpreted responsibility as *causal* responsibility, then the results indeed would not be surprising at all and would not suggest against the standard view or in favor of the responsibility account.[12]

A final issue is that a third sense of "responsible" is sometimes distinguished from the concepts of causal responsibility and moral responsibility—that of *legal* responsibility. Thus, a critic might contend that in our previous study participants might well have interpreted the responsibility attributions in the REC statements as saying that the agents are liable for compensating someone for the broken window, the deleted file, or the patient's condition. It should be noted, here, that while there is clearly a distinction to be drawn between legal responsibility and moral responsibility, both would seem to be normative concepts that involve further considerations beyond a purely descriptive notion of causation. As such, even if participants read the responsibility attributions in our first study in terms of legal responsibility, our results would arguably still suggest against the standard view and in favor of the responsibility account.

How can we address these concerns? A first thing to note is that while *philosophers* might treat "responsible" simpliciter as being ambiguous between "causal responsibility", "legal responsibility", and "moral responsibility", it is unclear that lay people find it similarly ambiguous. And there is some reason to suspect that they do not. Thus, Sytsma et al. (2019) present corpus evidence suggesting that ordinary responsibility attributions are normative, while Sytsma (forthcoming-b) presents corpus evidence indicating that "morally responsible" is very rarely used. Expanding on this, we collected data from the Corpus of Contemporary American

[12] Note, that discussions about the entailment claim often use "being a cause of X", "having caused X", and "being causally responsible for X" interchangeably. Sartorio, for instance, defines the entailment claim in the following way: "An agent A is morally responsible for an outcome O only if A is *causally responsible* for O, i.e., only if one of A's actions or omissions *caused* O; moreover, the fact that one of A's actions or omissions *caused* O (partly) explains the fact that A is morally responsible for O" (Sartorio 2007, p. 750, own emphasis).

English (COCA). Table 1 lists the number of hits for a range of phrases plausibly expressing different forms of responsibility. As can be seen from the table, ordinary people do not speak of causal responsibility: Of the mere 32 hits for “causal responsibility” and 4 for “causally responsible”, all but 5 come from academic texts. Uses of “morally responsible” and “legally responsible” are more frequent, although they remain rather uncommon, with only 0.2% of all uses of “responsible” being modified with the adverb “morally”.



Moral, Legal, and Causal Responsibility		
Term	Absolute Number of Hits	Percentage
responsibility	64109	100.00%
moral responsibility	633	0.99%
legal responsibility	205	0.32%
causal responsibility	32	0.05%
responsible	67355	100.0%
morally responsible	134	0.20%
legally responsible	154	0.23%
causally responsible	4	<0.01%

Table 1: List of the absolute hits and relative percentage for various responsibility phrases on COCA.

What we find is that non-academics seldom, if ever, use the phrases that philosophers employ to distinguish between concepts of responsibility. This does not necessarily mean that ordinary people lack such concepts, however; it simply means that they don’t express them in this way. Nonetheless, these findings are congruent with the contention that scholars have stipulated new technical notions that go beyond the ordinary sense of “responsible”. Coupled with previous results showing that normative considerations matter for people’s responsibility attributions, this suggests against the worries we’ve raised concerning our first study.

The most direct way of addressing these concerns would be to replace “responsible” with “morally responsible” in the statements from our first study, bringing them in line with standard statements of the Entailment Claim. Given the rather infrequent use of this phrase, however, there is a serious risk that participants would read too much into the use of "morally" if it were included in our test sentences. Luckily, there are alternative terms that can be used to resolve the potential ambiguity of "responsibility" and yet remain congruent with the standard view. This was done in our final two studies: In Study 2a we replaced "responsible" with "to blame" and in Study 2b we replaced "responsible" with "at fault".

3.1 Studies 2a: Entailment between Blame and Causation

3.1.1 Materials, Hypotheses, and Participants

The design of Study 2a strictly followed the design of Study 1. The only difference was that the term "responsible" was replaced with "to blame" in the EC and REC statements. For instance, the revised statements for causal attribution (i) now read as follows, with bolding added here to highlight differences and not included in the study:

> EC: John is **to blame** for the file being deleted, but by that I am not saying that John caused the file to be deleted.
>
> REC: John caused the file to be deleted, but by that I am not saying that John is **to blame** for the file being deleted.

The same null hypotheses that were posited in Study 1, were investigated in Study 2a: that the average contradiction ratings for EC statements are not significantly above the midpoint (Hypothesis 1), that the average contradiction ratings for REC statements are not significantly above the midpoint (Hypothesis 2), and that there is no significant difference in contradiction ratings between the EC and REC statements (Hypothesis 3).

The same recruitment method and pre-selection criteria were used as in Study 1. Results were collected from 72 participants (50 women, two non-binary persons, average age 31.8 years). As before, the study was pre-registered on OSF.[13]

3.1.2 Results

Once again, responses for the two practice questions indicated that participants understood the idea of a speaker contradicting herself.[14] The mean contradiction ratings for the five questions in the main study are depicted in Figure 2 along with the distribution of responses. The baseline and control conditions again worked as expected, with Semantic Entailment having a mean contradiction rating significantly above the midpoint [$t(71)=6.69$, $p<.001$ (one-tailed), $d=.79$], while Conversational Implicature and Control had mean ratings significantly below the midpoint [$t(71)=-19.78$, $p<.001$ (one-tailed), $d=2.33$; $t(71)=-12.04$, $p<.001$ (one-tailed), $d=1.42$]. As there were once again no significant effects of the different outcomes for either REC [$F(2,69)=.035$, $p=.97$, $\eta^2=.001$] or EC [$F(2,69)=2.28$, $p=.11$, $\eta^2=.062$], we collapsed the data for the different outcomes in the subsequent analysis.

The mean contradiction rating for the EC statements was again above the midpoint [M=6.90, SD=2.81, $t(71)=-12.04$, $p<.001$ (one-tailed), $d=1.42$], as was the mean rating for the REC statements [M=7.06, SD=2.70, $t(71)=6.47$, $p<.001$ (one-tailed), $d=.76$]. Indeed, once again the mean rating for the REC statements was not significantly different from the upper baseline given by Semantic Entailment [$t(71)=0.45$, $p=0.65$, $d=0.065$]. Finally, as in Study 1, no

[13] https://osf.io/3rgxs?view_only=917761f934c6450d96659a48990866e6

[14] Contradictory [M=8.47, SD=1.68, $t(71)=17.56$, $p<.001$ (one-tailed), $d=2.07$], Noncontradictory [M=2.28, SD=2.33, $t(71)=-9.93$, $p<.001$ (one-tailed), $d=1.17$].

significant difference was found between the mean ratings for the REC and EC statements [*t*(71)=0.42, *p*=0.67, *d*=0.055].

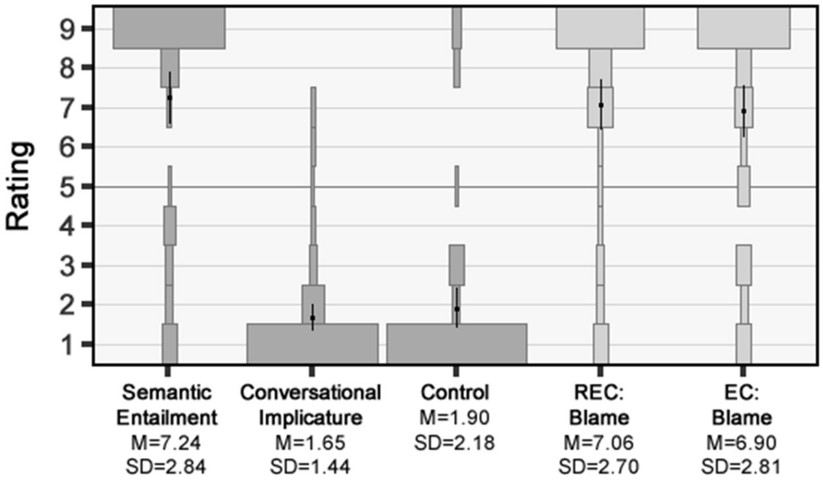


Figure 2: Results of Study 2a. The plots show the relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid.

3.2 Study 2b: Entailment between Fault and Causation

3.2.1 Materials, Hypotheses, and Participants

Study 2b again followed the same structure as Study 1, but this time we replaced "responsible" with "at fault". For instance, the revised statements for causal attribution (i) now read:

> John is **at fault** for the file being deleted, but by that I am not saying that John caused the file to be deleted.
>
> John caused the file to be deleted, but by that I am not saying that John is **at fault** for the file being deleted.

Once again, the same recruitment method and pre-selection criteria were used. Results were collected from 71 participants (64.8% women, two non-binary, average age 28.8 years). And, as before, the study was pre-registered on OSF.[15]

[15] https://osf.io/f3zud?view_only=6950ebb73a2c4024ad363674d75bed88

3.2.2 Results

Results are shown in Figure 3 and were once again in line with those from Study 1, including for the practice questions, Semantic Entailment, Conversational Implicature, and Control.[16] As there were no significant effects of the different outcomes for either REC [*F*(2,68)=.58, *p*=.56, η^2=.02] or EC [*F*(2,68)=.088, *p*=.92, η^2=.003], we again collapsed the data in the subsequent analysis. As in the previous two studies, the mean contradiction rating for the EC statements was above the midpoint [*t*(70)=4.79, *p*<.001 (one-tailed), *d*=0.57], as was the mean rating for the REC statements [*t*(70)=4.16, *p*<.001 (one-tailed), *d*=0.49]. Indeed, once again the mean rating for the REC statements was not significantly different from the upper baseline given by Semantic Entailment [*t*(70)=0.45, *p*=.66, *d*=0.068]. Finally, as in the previous studies, no significant difference was found between the mean ratings for the REC and EC statements [*t*(70)=0.36, *p*=0.72, *d*=0.044].

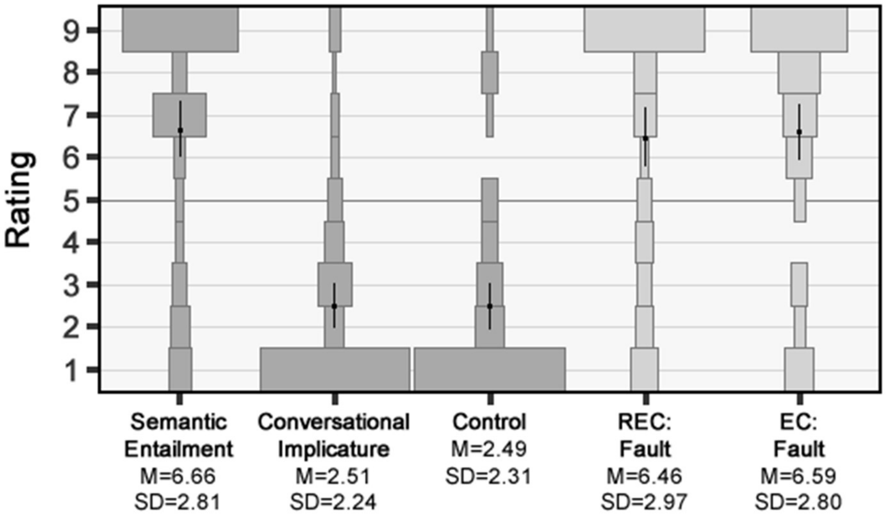


Figure 3: Results of Study 2b. Plots show the relative percentage of participants selecting each response option, with means and 95% confidence intervals overlaid.

[16] Contradictory [*t*(70)=10.80, *p*<.001 (one-tailed), *d*=1.28]; Non-contradictory [*t*(70)=4.40, *p*<.001 (one-tailed), *d*=.52]; Semantic Entailment [*t*(70)=4.98, *p*<.001 (one-tailed), *d*=.59]; Conversational Implicature [*t*(70)=-9.37, *p*<.001 (one-tailed), *d*=1.11]; Control [*t*(70)=-9.14, *p*<.001 (one-tailed), *d*=1.08].

3.3 Discussion

The results of Study 1 suggest that mutual entailment holds between responsibility and causation, providing evidence against the standard view and in line with the responsibility account. Against this interpretation, one might raise the following worry: The term “responsible” is notoriously ambiguous between a duty reading, a descriptive notion, and two backward-looking normative readings. We therefore decided to run two further studies in which we replaced “responsible” with terms that do not readily lend themselves to a non-normative interpretation—“blame” and “fault”. Despite these changes, the results matched those of our first study, and we rejected Hypothesis 1 and Hypothesis 2, but not Hypothesis 3: In line with both views, participants tended to treat each version of the EC statements as contradictory; but, in line with the responsibility account and against the standard view, they also tended to treat the REC statements as contradictory and ratings for these statements were not significantly different from those for the EC statements. Together, the findings across these three studies provide strong evidence that the relationship between the dominant ordinary concepts of responsibility and causation is one of mutual entailment.

4. Conclusion

The standard view holds that responsibility entails causation but that the reverse does not hold. In this paper, we have challenged this view empirically by investigating how the dominant ordinary concepts of causation and responsibility are related. In line with the responsibility account, we hypothesized that not only does causation entail responsibility, but that responsibility also entails causation.

Congruent with both the standard view and the responsibility account, our results suggest that causation is necessary for normative responsibility. Assigning responsibility, blame, or being at fault for one outcome, but immediately denying the agent's causal involvement is considered highly contradictory. Against the standard view, but in line with the responsibility account, however, participants also considered it to be highly contradictory to attribute causation to an agent, then immediately deny that the agent is responsible, to blame, or at fault for the outcome. This suggests that normative responsibility is also taken to be necessary for causation. As such, these findings jointly indicate that the relationship between responsibility and causation is one of mutual entailment.

To conclude, we consider a final worry that a critic might raise against the evidence we have provided for mutual entailment. The worry is that our methodology is unfit to provide *conclusive* evidence for an "entailment relation". And this is a reasonable point. The high contradiction ratings observed indicate that responsibility is not merely conversationally implicated by causation, but there remain alternative possibilities besides semantic entailment: Causation could merely presuppose responsibility or could conventionally implicate responsibility. While these are possible explanations of our findings, we believe that they are far less likely.

The first thing to note is that not only were the contradiction ratings for the REC statements very high across our three studies, but they were not statistically significantly different from those for the EC statements. Indeed, the distributions of responses for EC and REC are extremely similar. The most parsimonious explanation of such a pattern is that whatever relationship holds between responsibility and causation also holds between causation and responsibility (see Sytsma 2021 for a related argument). A quite plausible candidate for this relationship is mutual entailment. By contrast, that causation and responsibility mutually presuppose one another seems far less plausible. In fact, it is difficult to conceptualize what such *mutual* presupposition would even mean.

Perhaps a more viable alternative is to argue that responsibility and causation stand in a relationship of mutual conventional implicature to one another. The basic idea, here, is that the relation is something that is inferred, but not based on features of the conversational context; rather it is something that is inferred from the meaning of the sentence used. For instance, from "the queen is English and therefore brave" we infer that being brave follows from being English (Davis 2019, Section 2, based on Grice 1989, p. 25), even though being brave would not seem to be part of the meaning of "English". It is not clear how straightforwardly conventional implicature applies to our EC and REC statements. The most tempting option is perhaps to argue that our results reflect the meaning of the causal attributions (e.g., "John caused the file to be deleted") rather than telling us about the ordinary concept of causation. The difficulty here is that both the standard view and the responsibility account are focused on causal attributions, as noted in Section 1. As such, this objection would seem to at best result in a pyrrhic victory, with our findings still providing evidence against the standard view and for the responsibility account.

At this point, a critic might be unsatisfied with our "inference to the best explanation", arguing that despite the similarity in ratings for the EC and REC statements, these reflect different relations. Indeed, the fact that response patterns are statistically indistinguishable does not establish that we are dealing with the same relation. A presupposition relation and a semantic entailment relation, for instance, *might* happen to generate the same pattern of responses. But let us consider this example. An advocate of the standard view would need to argue that while responsibility entails causation, causation merely presupposes responsibility. This does not seem like a viable option, however, since a presupposition of a concept is necessarily poorer in information than the concept presupposing it and, thus, cannot at the same time entail the more information-rich statement. Nor would this seem to work any better if the critic were to instead claim that responsibility attributions conventionally implicate causal attributions. Finally, for the reasons discussed above, it does not seem that the critic would fare any better by explaining one pattern of results in terms of semantic entailment and the other in terms of conventional implicature, given the focus on causal attributions for specifying the concept of causation at issue for the Entailment Claim.

References

Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology*, 63(3), 368.

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126(4), 556–574.

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

Almeida, G., Struchiner, N., & Hannikainen, I. (2021). Rule is a dual character concept. *Available at SSRN 4018823*.

Coninx, S., Willemsen, P., & Reuter, Kevin (forthcoming). An experimental-linguistic study of the folk concept of pain: Implication, projection, deniability. In J. Culbertson, A. Perfors, Hugh Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Davis, Wayne (2019). Implicature. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition).

Driver, J. (2007). Attribution of causation for moral responsibility. In W. SinnottArmstrong (ed.), *Moral Psychology (Volume 2)*. Cambridge, Mass: MIT Press.

Fincham, F. D., Shultz, T. R. (1981). Intervening causation and the mitigation of responsibility for harm. *British Journal of Social Psychology*, 20(2), 113–120.

Grice, P. (1989). Logic and conversation. In P. Grice (1989), *Studies in the Way of Words* (pp. 22-40). Cambridge, Mass.: Harvard University Press.

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J., & Usher, M. (2020). Causal responsibility and robust causation. *Frontiers in Psychology*, 11, 1069.

Hall, N., & Paul, L.A. (2003). Causation and preemption. in Clark & Hawley (eds.) *Philosophy of Science Today*, Oxford: Oxford University Press.

Halpern, J., & Hitchcock, C. (2015). Graded causation and defaults. *British Journal for the Philosophy of Science*, 66, 413–457.

Heider, F. (1958). *The psychology of interpersonal relations.* Eastford: Martino Fine Books.

Henne, P., Pinillos, Á., & De Brigard, F. (2017). Cause by omission and norm: Not Watering Plants. *Australasian Journal of Philosophy*, 95(2), 270–283.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, 106, 587–612.

Icard, T., Kominsky, J., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, 161, 80–93.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psychology*, 2, 441-8.

Kominsky, J., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137, 196–209.

Kominsky, J., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science*, 43(11), e12792.

Livengood, J., & Machery, E. (2007). The folk probably don't think what you think they think: Experiments on Causation by Absence. *Midwest Studies in Philosophy*, 31, 107–127.

Livengood, J., & Sytsma, J. (2020). Actual causation and compositionality. *Philosophy of Science*, 87(1), 43-69.

Livengood, J., Sytsma, J., & Rose, D. (2017). Following the FAD: Folk attributions and theories of actual causation. *Review of Philosophy and Psychology*, 8(2), 273294.

Moore, M. S. (2010). *Causation and responsibility: An Essay in Law, Morals, and Metaphysics.* Oxford University Press.

Murray, D., & Lombrozo, T. (2017). Effects of manipulation on attributions of causation, Free Will, and Moral Responsibility. *Cognitive Science*, 41, 447–481.

Reuter, K., Kirfel, L., van Riel, R., & Barlassina, L. (2014). The good, the bad, and the timely: How temporal order and moral judgment influence causal selection. *Frontiers in Psychology*, 5, 1336.

Rose, D. (2017). Folk intuitions of actual causation: A two-pronged debunking explanation. *Philosophical Studies*, 174(5), 1323–1361.

Rose, D., Sievers, E., & Nichols, S. (2021). Cause and burn. *Cognition*, 207(104517).

Samland, J., Josephs, M., Waldmann, M., & Rakoczy, H. (2016). The role of prescriptive norms and knowledge in children's and adults' causal selection. *Journal of Experimental Psychology: General*, 145(2), 125–130.

Samland, J., & Waldmann, M. (2016). How prescriptive norms influence causal inferences. *Cognition*, 156, 164–176.

Sarin, A., Lagnado, D., & Burgess, P. (2017). The intention-outcome asymmetry effect: How incongruent intentions and outcomes influence judgments of responsibility and causality. *Experimental Psychology*, 64(2), 124–141.

Sartorio, C. (2007). Causation and responsibility. *Philosophy Compass*, 2(5), 749765.

Schaffer, J. (2004). Causes need not be physically connected to their effects: The case for negative causation. In C. Hitchcock (ed.) *Contemporary debates in philosophy of science*, 197-216.

Schleifer, M., Shultz, T. R., Lefebvre-Pinard, M. (1983). Children's judgements of causality, responsibility and punishment in cases of harm due to omission. *British Journal of Developmental Psychology*, 1(1), 87–97

Schwenkler, J., & Sievers, E. (forthcoming). Cause, "cause", and norm. In P. Willemsen and A. Wiegmann (eds.) *Advances in Experimental Philosophy of Causation*.

Scanlon, T. (2008). *Moral dimensions: permissibility, meaning, blame*. Cambridge: Belknap Press.

Shaver, K. G. (1985). *The attribution of blame*. New York, NY: Springer New York

Sytsma, J. (2021). Causation, Responsibility, and Typicality. *Review of Philosophy and Psychology*, 12, 699–719.

Sytsma, J. (forthcoming-a). The responsibility account. In P. Willemsen and A. Wiegmann (eds.), *Advances in Experimental Philosophy of Causation*, London: Bloomsbury Press. http://philsci-archive.pitt.edu/19180/

Sytsma, J. (forthcoming-b). Crossed wires: Blaming artifacts for bad outcomes. *The Journal of Philosophy*. http://philsci-archive.pitt.edu/19040/

Sytsma, J., Bluhm, R., Willemsen, P., & Reuter, K. (2019). Causal attributions and corpus linguistics, In Fischer & Curtis (eds.) *Methodological Advances in Experimental Philosophy*. London: Bloomsbury.

Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4), 814-820.

Sytsma, J., & Livengood, J. (2021). Causal attributions and the trolley problem. *Philosophical Psychology*, 34(8), 1167–1191.

Talbert, M. (2019). Moral responsibility. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition).

Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass*, 14(1), e12562.

Willemsen, P., & Reuter, K. (2016). Is there really an omission effect? *Philosophical Psychology*, *29*(8), 1142-1159.

Willemsen, P., & Reuter, K. (2021). Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought*.

Wolf, S. (1993). The real self view. In J. Fischer & M. Ravizza (eds.), *Perspectives on Moral Responsibility*, (pp. 151–69). Ithaca, London: Cornell University Press.