Resituating the Influence of Relevant Alternatives on Attributions

Justin Sytsma[1]

Abstract: Phillips et al. (2015) provide what looks like compelling evidence in favor of explaining the impact of broadly moral evaluations on a range of attributions in terms of the relevance of alternative possibilities. In a series of manipulation studies, they found that asking participants to describe what an agent could have done differently in neutral cases (cases in which information about broadly moral considerations was removed) showed a similar effect to varying the morality of the agent's action. Phillips et al. take this to show that broadly moral evaluations impact the alternative possibilities people see as relevant, which in turn impact their attributions. These studies leave open the possibility that the manipulation impacts people's broadly moral evaluations which in turn impact their attributions, however, rather than directly impacting their attributions. But this alternative model conflicts with Phillips et al.'s account, while being compatible with competing explanations. These two models are tested for causal attributions using the same manipulation method. Against Phillips et al.'s model, the results suggest that the influence of relevant alternative possibilities on causal attributions works primarily through people's broadly moral evaluations.

Broadly moral considerations have been shown to have a notable impact on a range of judgments, including on ordinary causal attributions.[2] Perhaps the most discussed example for causal attributions is the Pen Case from Knobe and Fraser (2008):

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.
>
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly e-mailed them reminders that only administrative assistants are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk.

[1] I want to thank David Rose for helpful comments on a previous version of this paper, as well as his guidance on the Greedy Equivalence Searches reported.

[2] By "ordinary causal attributions" I mean the use of language like "X caused Y". While these effects are often discussed in terms of morality's influence on judgments, the actions in question sometimes fall short of what we might consider "moral"; as such, "broadly moral considerations" is intended to capture a larger class of norms, and specifically to capture injunctive norms, including both prescriptive norms (what should be done) and proscriptive norms (what should not be done).

This scenario describes a situation in which two agents jointly bring about a bad outcome by performing actions that are symmetric outside of Professor Smith violating an injunctive norm (faculty members are prohibited from taking pens). Despite this, when participants were asked to rate agreement with a causal attribution for each agent, ratings were very notably higher for Professor Smith. How are we to explain findings like this?

Knobe and colleagues have developed an intriguing explanation, arguing that the effect of norms works through the counterfactuals that people consider (Hitchcock and Knobe 2009, Halpern and Hitchcock 2015, Kominsky et al. 2015, Phillips et al. 2015, Icard et al. 2017, Kominsky and Phillips 2019). This *counterfactual view* holds that norm violations make the counterfactual on which the norm-violation does not occur seem more relevant to people, such that they are more likely to consider that counterfactual. And, in scenarios like the Pen Case, if the agent did not violate the norm (e.g., if Professor Smith did not take a pen), then the outcome would not have occurred. Knobe and colleagues argue that this then leads participants to judge that the norm-violating agent is more causal since the counterfactual highlights that but for that agent's action, the outcome would not have occurred. Further, this account has been extended to a host of other judgments that are impacted by broadly moral considerations (Phillips et al. 2015, Knobe forthcoming), including judgments of freedom (e.g., Phillips and Knobe 2009), doing versus allowing (e.g., Cushman et al. 2008), and intentionality (e.g., Knobe 2003). For each of these, Knobe and colleagues have argued that the influence of broadly moral considerations can be explained in terms of the alternative possibilities that people consider relevant.

One of the most compelling pieces of evidence for the alternative possibilities account comes from the manipulation experiments performed by Phillips et al. (2015). In these studies, they gave participants neutral versions of the vignettes used to show the impact of broadly moral

considerations on each of the judgments noted above (i.e., vignettes where the broadly moral considerations were removed), varying whether participants were asked to briefly describe what the primary agent could have done differently or simply to describe the scenario they read. Phillips et al.'s idea was that writing about what the agent could have done differently would increase the salience of the alternative actions the agent might have taken, inducing the same effect found previously when information about broadly moral considerations was included. And they interpret their findings as indicating that "independent of morality, the relevance of alternative possibilities to the agent's actions showed the same pattern of influence that morality has been shown to have" (39–40).

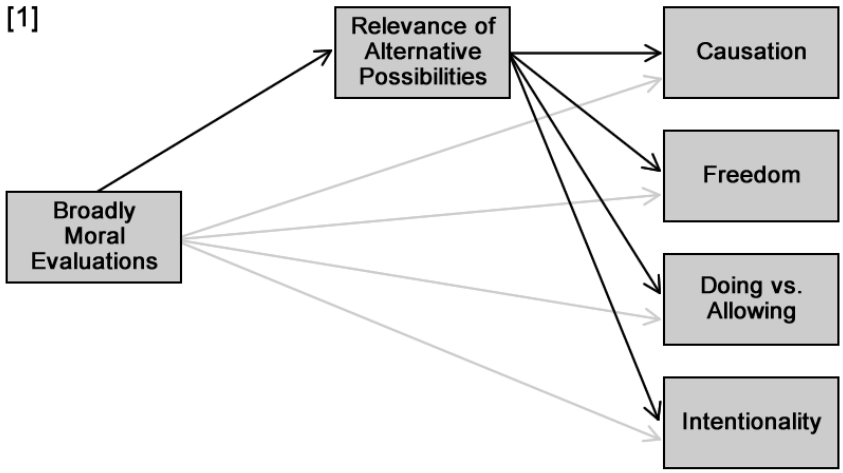

Figure 1: Model proposed by the alternative possibilities account, adapted from Figure 1 in Phillips et al. (2015).

The alternative possibilities account holds that the impact of broadly moral evaluations on the judgments at issue (causation, freedom, doing versus allowing, intentionality) primarily works through the alternative possibilities that people consider relevant. That is, they propose model **[1]** in Figure 1. Previous results show the impact of broadly moral evaluations on the

relevant judgments, but do not show that this impact primarily works through the alternative possibilities that people consider relevant. The manipulation studies aimed to test this mechanism, showing that the same type of effect is produced by manipulating the alternative possibilities people find relevant, and thereby providing evidence for the paths from the relevance of alternative possibilities to the judgments in **[1]**.

There is a potential problem, however: Phillips et al. did not test participants' broadly moral evaluations. As such, their studies leave open the possibility that the alternative possibilities that participants considered relevant impacted their broadly moral evaluations and that their broadly moral evaluations in turn impacted their judgments. In other words, model **[2]** in Figure 2 is also compatible with Phillips et al.'s results. But this model runs counter to the alternative possibilities account, positing that the impact of broadly moral evaluations on the relevant judgments primarily works directly, not through the alternative possibilities that people consider relevant.

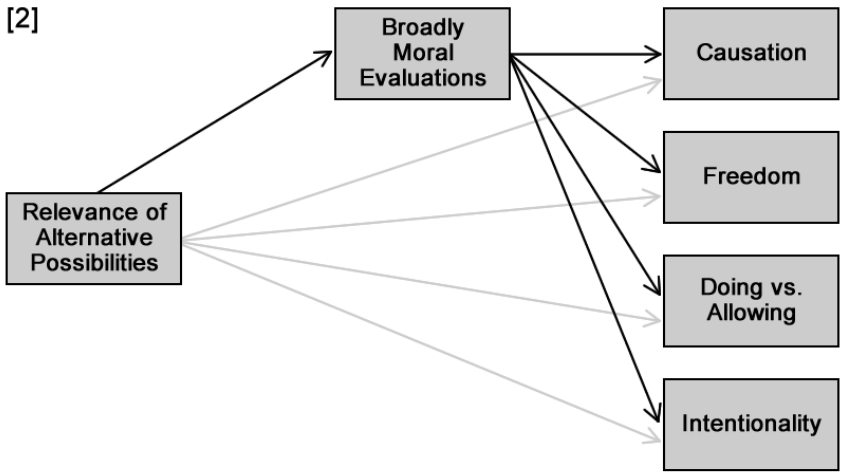

Figure 2: Proposed alternative model.

While model **[2]** runs counter to the alternative possibilities account, it is compatible with a number of competing explanations in the literature. Focusing on accounts of the impact of broadly moral considerations on causal attributions, the model is compatible with our *responsibility view*, Samland and Waldmann's *pragmatic view*, and Alicke's *bias view*. The responsibility view contends that broadly moral evaluations matter for causal attributions because the ordinary concept of causation at play in such attributions is a broadly moral concept, akin to the concepts of responsibility and accountability (Sytsma et al. 2012, Livengood et al. 2017, Sytsma et al. 2019, Livengood and Sytsma 2020). As such, we deny that the impact of broadly moral evaluations on causal attributions works primarily through the counterfactuals people consider relevant. Since our view is silent on how people arrive at their broadly moral evaluations, however, it allows that the alternative possibilities we consider relevant might play a role in those evaluations. As such, our view is compatible with the alternative explanation of Phillips et al.'s findings. And similarly, for the pragmatic and bias views.[3]

Thus, determining whether the effect of Phillips et al.'s manipulation reflects the direct or the indirect impact of the relevance of alternative possibilities on causal attributions is important for determining whether their results support the counterfactual view or the competitors. And similar points could be made for accounts of the impact of broadly moral considerations on other judgments, such as intentionality. To test the mechanism behind the effect shown by Phillips et al., I replicated their manipulation study for causal attributions,

[3] The pragmatic view holds that ordinary concept of causation is non-normative, but that for pragmatic reasons participants read the questions in probes like that given by Knobe and Fraser (2008) to instead be asking about a normative concept like responsibility or accountability (Samland and Waldmann 2016, Samland et al. 2016). Their view is also silent about how we arrive at our broadly moral evaluations, however, and as such is compatible with those evaluations being influenced by alternative possibilities. See Sytsma et al. (2019) for a discussion of the pragmatic view relative to the responsibility view. The blame view holds that people's judgments tend to be biased by their desire to blame or praise the agent (Alicke 1992, 2000; Alicke et al. 2011; Rose 2017). And this view is compatible with considerations of alternative possibilities playing a role in our broadly moral evaluations and those evaluations then impacting blame and praise judgments. See Sytsma (under review) for a discussion of the bias view relative to the responsibility view.

adding two further questions—one about a broadly moral attribution and one about a responsibility attribution. The results suggest that the impact of the relevance of alternative possibilities on causal attributions primarily works through people's broadly moral evaluations.

1. Study

1.1 Methods

Participants were recruited through advertising for a free personality test on Google. In addition to answering the questions reported below, participants were asked basic demographic questions and after the test questions were given a 10-item Big Five personality inventory. Responses were collected from 210 native English-speakers, age 16 or older.[4] The sample size was selected to give a power greater than 90% to detect an effect of the size reported by Phillips et al. (d=0.42) using a one-tailed test.

Each participant read the vignette for the neutral version of the Pen Case from Phillips et al.'s second study:

> The receptionist in the philosophy department keeps her desk stocked with pens. Both the administrative assistants and the faculty members are allowed to take the pens, and both the administrative assistants and the faculty members typically do take the pens. The receptionist has repeatedly e-mailed them reminders that both administrators and professors are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later, that day, the receptionist needs to take an important message... but she has a problem. There are no pens left on her desk.

Participants were then either asked to describe the events of the story (control) or to write about what other decision Professor Smith could have made (alternatives). Finally, they rated their

[4] 63.3% women (three non-binary), average age 41.4 years, ranging from 16 to 80. Given the higher percentage of women, ANOVAs were run for each question with condition and gender as between-participant factors. No significant gender effects were found.

agreement or disagreement with each of three claims on a scale from 1 ("completely disagree") to 7 ("completely agree"):

> (Wrong) It was wrong for Professor Smith to take a pen.
> (Responsible) Professor Smith is responsible for the problem.
> (Caused) Professor Smith caused the problem.

The three claims were presented in random order.

1.2 Results

Results are shown in Figure 3. The original study successfully replicated. Analyzing Phillips et al.'s results together with those for the present study, an ANOVA looking at ratings for Caused with study and condition as between-participant factors showed no significant effects for study, but did show a significant main effect for condition, $F(1, 440)=15.0$, $p<0.001$, $\eta^2=0.033$. This suggests that including the two additional questions did not have a notable effect on responses to Caused.[5]

In line with the effect reported by Phillips et al., for the present study the mean rating for Caused in the alternatives condition was significantly higher than the mean rating in the control condition, $t(206.18)=2.29$, $p=0.01$, d=0.32; $W=6341.5$, $p=0.021$.[6] A similar effect was found for Responsible, $t(207.5)=1.67$, $p=0.048$, d=0.23; $W=6246.5$, $p=0.038$, and Wrong, $t(208)=3.13$, $p=0.0010$, d=0.43; $W=6887.5$, $p<0.001$. In line with previous results (e.g., Sytsma and Livengood under review), there was a strong correlation between Responsible and Caused,

[5] It is worth noting that this potentially suggests against Samland and Waldmann's pragmatic view. If pragmatic factors were leading participants to interpret the causal attribution as instead asking about a normative concept like responsibility, we would expect the effect of the manipulation to notably dissipate when also asking participants about a responsibility attribution; but this is not what was found.

[6] Welch's t-tests are used throughout. Since the effect at issue is directional—mean causal rating being higher in the alternatives condition—I use one-tailed tests unless specified otherwise. And while it is typical to test for effects on causal attributions using parametric statistics, the distributions are often non-normal; as such, I will report non-parametric tests (either Wilcoxon signed rank tests, *W*, or Wilcoxon rank sum tests, *V*) in addition to the standard t-tests. Similarly for correlations, reporting Pearson's r and Spearman's ρ.

r=0.76, t(208)=17.1, p<0.001; ρ=0.77, S=361990, p<0.001. Further, no significant difference was found between them for either the control condition, t(102)=0.18, p=0.85 (two-tailed), d=0.011; V=213, p=0.56, or the alternatives condition, t(106)=1.17, p=0.24 (two-tailed), d=0.08; V=397.5, p=0.25.[7] Somewhat weaker correlations were found between Wrong and Responsible, r=0.54, t(208)=9.28, p<0.001; ρ=0.54, S=702820, p<0.001, and between Wrong and Caused, r=0.47, t(208)=7.61, p<0.001; ρ=0.49, S=789990, p<0.001.

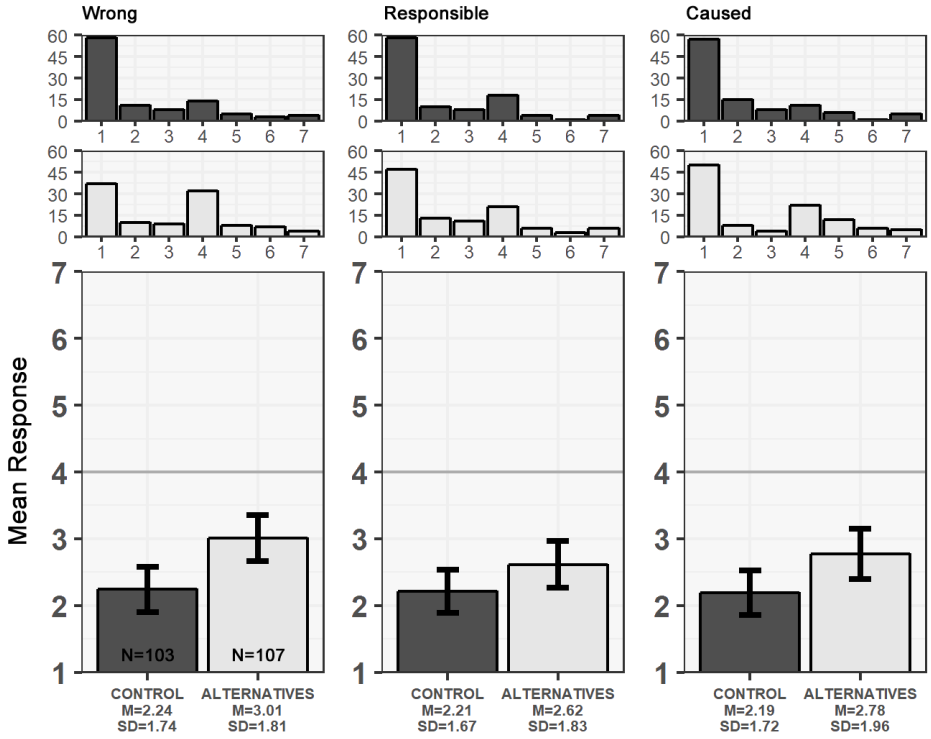

Figure 3: Results showing 95% confidence intervals; histograms above.

[7] Again, the close correspondence between causal attributions and responsibility attributions suggests against the pragmatic view: if the correspondence was due to pragmatic factors leading participants to interpret the causal attribution as instead asking about responsibility, we would expect this to be cancelled-out by including the responsibility attribution; but it wasn't.

To test the two mechanisms noted above, I used the Greedy Equivalence Search (GES) algorithm in TETRAD 6.4.0.[8] Since the alternative possibilities account does not make a specific prediction about responsibility attributions, I began by excluding this variable. The search returned model **[A]** in Figure 4, which was a good fit for the data, df=1, χ^2=0.95, p=0.33, *CFI*=1.00, *RMSEA*=0.00, *BIC*=-12.1. This model corresponds with model **[2]**—the alternative to Phillips et al.'s model shown in Figure 2—with the effect of condition on Caused working through Wrong, rather than directly as predicted by the counterfactual view. To test for alternative models, I ran a second GES specifying that condition has a direct effect on Caused. The search returned model **[B]** in Figure 4, df=1, χ^2=5.67, p=0.017, *CFI*=0.95, *RMSEA*=0.15, *BIC*=0.33. Following the rules of thumb given by Schermelleh-Engel et al. (2003), **[B]** is an "acceptable fit" based on p value (>0.01) and *CFI* (>0.95) but is not an acceptable fit based on χ^2 (>3df) or *RMSEA* (>0.08). In contrast, **[A]** is a "good fit" on all four measures (p value > 0.05, *CFI* > 0.97, χ^2 > 3df, *RMSEA* < 0.05).[9]

[8] GES searches over equivalence classes of models by assigning an information score (in this case BIC) to each. The algorithm starts with the null graph, then adds the edge that most improves the information score and applies the edge-orientation rules in Meek (1997). This is iterated until no additions further improve the information score. The algorithm then considers deletions, removing edges while doing so increases the information score, and again applying Meek's orientation rules. As Danks (2016, 467) notes, GES "is now the most widely used [score-based search algorithm] since it has proven to be the most reliable." For discussion see Chickering (2002) and Danks (2016). For applications in experimental philosophy, see Rose et al. (2011), Rose and Nichols (2013), Rose (2017).

[9] As expected, bootstrap mediation analyses (Preacher and Hayes 2008) showed comparable results. First, I tested whether Wrong mediates the effect of condition on Caused using 5000 resamples. There was a significant indirect effect (95% CIs [0.13, 0.63]). I then tested whether Caused mediates Wrong using 5000 resamples. Again there was a significant indirect effect (95% CIs [0.04, 0.50]). While both indirect effects are significant, Wrong mediates a notably larger proportion of the effect of condition on Caused (0.60) than Caused does on Wrong (0.33), and while the direct effect of condition on Caused when controlling for Wrong is not significant (p=0.34), the direct effect of condition on Wrong when controlling for Caused is significant (p=0.014).

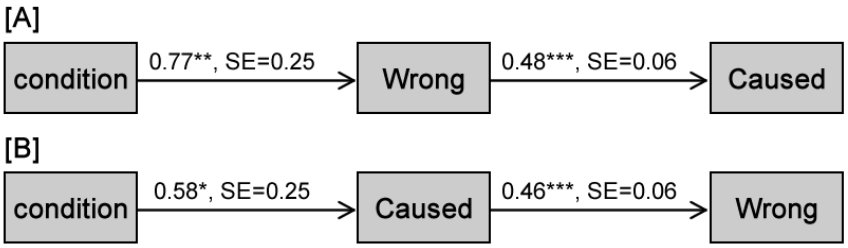

Figure 4: GES outputs from TETRAD excluding Responsible with standardized path coefficients and standard errors. Model **[A]** shows search without forcing a direct path from condition to Caused, **[B]** forcing a direct path. Asterisks indicate significance, **p*<0.05, ***p*<0.01, ****p*<0 .001.

I then ran a GES including Responsible. The search returned model **[C]** in Figure 5, which is again a good fit for the data based on each of the four measures, *df*=3, χ^2=3.99, *p*=0.26, *CFI*=1.00, *RMSEA*=0.040, *BIC*=-12.1. This model also has the effect of condition on Caused working through Wrong, although in addition it now runs through Responsible.[10] To test for alternative models, I first ran a GES specifying that condition has a direct effect on Caused. The search returned model **[D]** in Figure 5, *df*=3, χ^2=8.95, *p*=0.030, *CFI*=1.00, *RMSEA*=0.097, *BIC*=-7.09. This model is a good fit based on *CFI* and an acceptable fit based on *p* value and χ^2 but is not an acceptable fit based on *RMSEA*. Finally, I ran a GES specifying that condition has a direct effect on Responsible. The search returned model **[E]** in Figure 5, *df*=3, χ^2=10.3, *p*=0.016, *CFI*=0.99, *RMSEA*=0.11, *BIC*=-5.78. This model is a good fit based on *CFI* and an acceptable fit based on *p* value but is not an acceptable fit based on χ^2 or *RMSEA*.

[10] Interestingly, this is the reverse of the relationship often assumed in philosophical discussions, where causation is taken to be necessary for responsibility, but responsibility is not taken to be necessary for causation (see, e.g., Driver 2008).

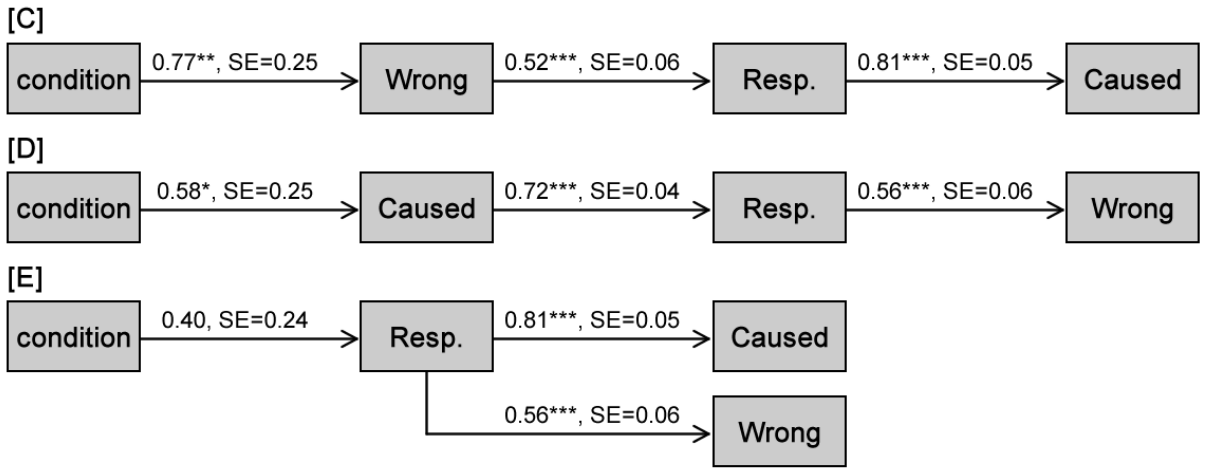

Figure 5: GES outputs from TETRAD including Responsible with standardized path coefficients and standard errors. Model **[C]** shows search without forcing a direct path, **[D]** forcing a direct path from condition to Caused, and **[E]** forcing a direct path to Responsible. Asterisks indicate significance, *p<0.05, **p<0.01, ***p<0 .001.

1.3 Discussion

The manipulation studies conducted by Phillips et al. (2015) show that asking participants to write about what an agent could have done differently impacts ratings for a range of attributions, including causal attributions, and does so in a way similar to that previously found for including information bearing on broadly moral considerations. They take these studies to provide direct empirical evidence for a link between the relevance of alternative possibilities and attributions that is independent from broadly moral evaluations. But these studies did not test whether the manipulation also impacts people's broadly moral evaluations, leaving open the possibility that the effect on attributions primarily works through such evaluations. I tested these competing models for causal attributions, following Phillips et al.'s procedure except for also having participants rate a broadly moral evaluation and a responsibility attribution. The predicted effect was again found for causal attributions but was also seen for the other two ratings. This indicates that the mechanism suggested by Phillips et al. cannot be presumed. Further, causal search

indicates that the best fitting model has the impact of the manipulation on causal attributions working through broadly moral attributions, and in turn responsibility attributions. And this model was a notably better fit than the best alternative model with the manipulation directly affecting causal attributions.

The preferred model produced by TETRAD differs from that previously suggested for the responsibility view (Sytsma under review), where broadly moral evaluations were modeled as a common cause for both causal attributions and responsibility attributions. Model **[C]** instead indicates a chain, with broadly moral evaluations impacting responsibility attributions, which in turn impact causal attributions. This model remains consistent with the key insight of the responsibility view, however, which is that the ordinary concept of causation at play in causal attributions is a normative concept, just as responsibility and accountability are normative concepts. Further, the strong correlation between causal attributions and responsibility attributions found in the present study lends further credence to our view.

In contrast, the results suggest against the model given by Phillips et al., putting pressure on the counterfactual view of the impact of broadly moral considerations on causal attributions: the present findings suggest that it is not that broadly moral evaluations impact the counterfactuals we find relevant, but the other way around. Finally, although Phillips et al.'s manipulation was not tested for judgments of freedom, doing versus allowing, or intentionality, the finding that the manipulation impacted broadly moral evaluations puts pressure on their interpretation of the results for these as well, and with it their general alternative possibilities account.

References

Alicke, M. (1992). "Culpable causation." *Journal of Personality and Social Psychology*, 63: 368–378.

Alicke, M. (2000). "Culpable Control and the Psychology of Blame." *Psychological Bulletin*, 126(4): 556–574.

Alicke, M., D. Rose, and D. Bloom (2011). "Causation, Norm Violation, and Culpable Control." *Journal of Philosophy*, 108: 670–696.

Chickering, D. (2002). "Optimal structure identification with greedy search." *Journal of Machine Learning Research*, 3: 507–554.

Cushman, F., J. Knobe, and W. Sinnott-Armstrong (2008). "Moral appraisals affect doing/allowing judgments." *Cognition*, 108: 281–289.

Danks, D. (2016). "Causal Search, Causal Modeling, and the Folk." In J. Sytsma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*, pp. 463–471, West Sussex: Wiley Blackwell.

Driver, J. (2008). "Attributions of Causation and Moral Responsibility." In W. Sinnott-Armstrong (ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 423–439, Cambridge: MIT Press.

Halpern, J. and C. Hitchcock (2015). "Graded Causation and Defaults." *British Journal for the Philosophy of Science*, 66: 413–457.

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

Icard, T., J. Kominsky, and J. Knobe (2017). "Normality and Actual Causal Strength." *Cognition*, 161: 80–93.

Knobe, J. (2003). "Intentional action and side effects in ordinary language." *Analysis*, 63: 190–193.

Knobe, J. (forthcoming). "Morality and Possibility." In J. Doris and M. Vargas (eds.), *The Oxford Handbook of Moral Psychology*. Oxford: Oxford University Press.

Knobe, J. and B. Fraser (2008). "Causal judgments and moral judgment: Two experiments." In W. Sinnott-Armstrong (ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 441–447, Cambridge: MIT Press.

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

Kominsky, J. and J. Phillips (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., J. Sytsma, and D. Rose (2017). “Following the FAD: Folk attributions and theories of actual causation.” *Review of Philosophy and Psychology*, 8(2): 274–294.

Livengood, J. and J. Sytsma (2020). “Actual causation and compositionality.” *Philosophy of Science*, 87(1): 43–69.

Meek, C. (1997). *Graphical Models: Selecting Causal and Statistical Models*. PhD Thesis, Carnegie Mellon University.

Phillips, J. and J. Knobe (2009). “Moral judgments and intuitions about freedom.” *Psychological Inquiry*, 20: 30–36.

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

Preacher, K. and A. Hayes (2008). “Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models.” *Behavior Research Methods*, 40: 879–891.

Rose, D. (2017). “Folk Intuitions of Actual Causation: A Two-pronged Debunking Explanation.” *Philosophical Studies*, 174(5): 1323–1361.

Rose, D., J. Livengood, J. Sytsma, and E. Machery (2011). “Deep trouble for the deep self.” *Philosophical Psychology*, 25: 629–646.

Rose, D. and S. Nichols (2013). “The lesson of bypassing.” *Review of Philosophy and Psychology*, 4: 599–619.

Samland, J., M. Josephs, M. Waldmann, and H. Rakoczy (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. and M. R. Waldmann (2016). “How prescriptive norms influence causal inferences.” *Cognition*, 156: 164–176.

Schermelleh-Engel, K. and H. Moosbrugger (2003). “Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures.” *Methods of Psychological Research Online*, 8(2): 23–74.

Sytsma, J. (under review). “The Character of Causation: Investigating the Impact of Character, Knowledge, and Desire on Causal Attributions.” http://philsci-archive.pitt.edu/16739/

Sytsma, J. and J. Livengood (under review). “Causal Attributions and the Trolley Problem.” http://philsci-archive.pitt.edu/16200/

Sytsma, J., J. Livengood, and D. Rose (2012). “Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions.” *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43: 814–820.

Sytsma, J., R. Bluhm, P. Willemsen, and K. Reuter (2019). “Causal Attributions and Corpus Analysis.” In E. Fischer and M. Curtis (eds.), *Methodological Advances in Experimental Philosophy*, London: Bloomsbury Press.