Explanation classification depends on understanding: extending the epistemic side-effect effect

Daniel A. Wilkenfeld[1] and Tania Lombrozo[2]

Abstract

Our goal in this paper is to experimentally investigate whether folk conceptions of explanation are psychologistic. In particular, are people more likely to classify speech acts as explanations when they cause understanding in their recipient? The empirical evidence that we present suggests this is so. Using the side-effect effect as a marker of mental state ascriptions, we argue that lay judgments of explanatory status are mediated by judgments of a speaker's and/or audience's mental states. First, we show that attributions of both understanding and explanation exhibit a side-effect effect. Next, we show that when the speaker's and audience's level of understanding is stipulated, the explanation side-effect effect goes away entirely. These results not only extend the side-effect effect to attributions of understanding, they also suggest that attributions of explanation exhibit a side-effect effect because they depend upon attributions of understanding, supporting the idea that folk conceptions of explanation are psychologistic.

Keywords: Explanation, understanding, side-effect effect, psychologism

1. Introduction

1.1 Explanation and Understanding

[1] Corresponding author
University of Pittsburgh
Department of History and Philosophy of Science
Email: dawilk@gmail.com
[2] University of California, Berkeley
Department of Psychology

Explanation and understanding often go hand in hand. Explanations can generate understanding, and, on some views, understanding consists in knowledge or grasp of explanations (e.g., Strevens 2013, Khalifa in press). But is understanding essential to explanation? Or is it merely a common and desirable consequence, a symptom of explanations?

On the one hand, philosophers of science for many years took it as a *datum* that there could be explanations without understanding, and the intuition is compelling—clearly there can be explanations of phenomena so complex or cosmic in scope that it beggars belief to suppose the human mind could ever grasp them. Hempel and Oppenheim (1948, p. 17), in particular, argued against 'psychologistic' approaches to explanation based on the following dilemma: either understanding is objective or subjective. Subjective understanding is associated with a 'feeling of empathic familiarity' (ibid.), and so on this horn a requirement that explanations engender understanding is odious (at least to a positivist philosopher of science). Objective understanding is just what one has when one knows a D-N explanation, and so on this horn a requirement that explanation engender understanding is superfluous. (For a discussion of this dilemma and what it leaves out, see Wilkenfeld 2014.)

On the other hand, alternative strands of philosophical work (e.g., Wilkenfeld 2014) have been more favorable to psychologism about explanation—the view that it is at least partially constitutive of an explanation to be conducive to understanding. These strands suggest that understanding is an important tool in our conceptual toolkit when attempting to elucidate the nature of explanation. This perspective is supported by recent empirical work on scientists' and laypeople's conception of explanation: Specifically, work by Waskan, Harmon, Horne, Spino, and Clevenger (2014) suggests that both scientists and laypeople consider an intellectual achievement to be an explanation only when it actually causes understanding in someone.

Our goal in this paper is to apply an existing paradigm in a new area to experimentally investigate whether folk conceptions of explanation are in fact psychologistic. That is, do people only countenance

speech acts as explanations when they cause understanding in their recipient? The empirical evidence that we present suggests this is so. Our main argument will involve a novel deployment of a tool whose full implications are still being explored—the presence of a side-effect effect (SEE) (described in §1.3) as a marker for attributions of the mental.[3]

We begin in §1.2 with an examination of a recent empirical finding that supports psychologism about explanation, while noting some boundaries of that finding that our own research aims to complement. In §1.3 we introduce the SEE as a phenomenon that we can repurpose as a marker for whether the classification of something as an 'explanation' depends on the mental states attributed to the agents involved. This sets up §1.4, in which we utilize our interpretation of the SEE and its applications to motivate our main empirical research questions. In §2-§4 we present our new experimental results which, we argue, jointly provide a proof-of-concept of using the SEE as a means to probe mental state attributions, defend psychologism about explanation, and demonstrate some of the contours of the SEE as well. In §5 we revisit our tentative assumption regarding the explanation of the SEE, and present a dilemma wherein, whatever scope one thinks the SEE has, one has to draw novel conclusions from our data. We also explore alternative explanations for the effects that we observe. In §6 we explicitly compare our results to those of Waskan et al. (2014), highlighting some dimensions along which our research continues to advance their program.

1.2 The Data on Psychologism of Explanation

While philosophers have for some time debated what we have characterized as psychologism about explanation, there have been few studies that attempt to measure whether such psychologism does or does not characterize laypeople's conception of explanation. We contend that understanding laypeople's conception is important whether or not this descriptive psychological project is regarded as continuous or

[3] While there has a great deal of work exploring the side-effect effect itself, using it as a tool to look at the contours of a concept like explanation is a somewhat different approach.

discontinuous with the philosophical project of developing an account of explanation. On the former view, data concerning laypeople's judgments and explanatory practices is an important constraint on theorizing. On the latter view, these data can help us overcome cognitive biases that arise in developing an objective or normative account of explanation.

Waskan et al. (2014) have done the most thorough study of whether laypeople and scientists consider the generation of understanding ('intelligibility') in evaluating what constitutes an explanation. In an initial study, participants read vignettes involving a scientist who makes a new discovery regarding the cause of gamma ray bursts from a distant galaxy. The scientist produces a model, but the stated characteristics of this model vary across conditions. In some cases, the model allows the scientist to 'come to understand' the cause of the gamma ray bursts. In a second case, understanding has not yet been achieved, but the model *could* enable a competent scientist who examines the model to come to understand. In a final case, the model is so complex that it would not allow someone to come to understand, yet it still possesses many of the other, more objective virtues of explanations, such as predictive power. Participants indicated their agreement or disagreement with the statement that the model the scientist produced 'constitutes an explanation' (Waskan et al 2014 p. 1023).[4] Waskan et al. found that people indicated significantly greater agreement with the statement that the model constitutes an explanation when it produced actual understanding as opposed to potential understanding or no understanding.

In subsequent experiments, Waskan et al. presented more complex vignettes with the same basic structure, varying whether a scientific model produced actual understanding, potential understanding, or no understanding. They also varied the way in which they assessed the association between explanation and understanding by employing a variant of a semantic integration task: participants first read a vignette

[4] The authors are at pains to distinguish the question of whether or not understanding-generation is necessary for a model to count as an explanation at all from the further question of whether it is a good-making feature of an explanation.

that did not explicitly classify a scientific model as an explanation, and later answered questions designed to assess whether they had activated the concept of 'explanation' in reading about the model. Specifically, participants were asked whether the following 'claim [was] likely to be true based upon what they read,' that: 'Dr. Brown's paper and the accompanying computer model provide an explanation for why type-B2 stars produce gamma-ray bursts' (Waskan et al 2014 p. 1028). Mirroring their initial results, Waskan et al. found that both laypeople and scientists endorsed the statement that the model provided an explanation when it generated *actual* understanding, but endorsed the statement at lower rates when it was only potentially intelligible or never intelligible.

Waskan et al. (2014) conclude that laypeople's and scientists' conception of explanation is importantly psychologistic: actual intelligibility appears to be a necessary condition for classifying a model as an explanation. This certainly departs from the anti-psychologism advocated by Hempel and others, but Waskan et al. are also careful to distance their characterization of the folk/scientific conception from the kind of psychologism that philosophers most often reject: the idea that explanation is a matter of supporting or achieving some affective state or feeling of familiarity. Instead, they identify explanation with *intelligibility*, which they describe as 'the more intellectual occurrence of understanding how or why, at least possibly, the target of explanation came about' (p. 1019).

While Waskan et al.'s studies provide compelling evidence of a close relationship between explanation and understanding, it's challenging to make the stronger claim that intelligibility is partially constitutive of explanation. Instead, it could be that participants treat intelligibility as evidence for the presence of some objective feature F, where this feature F is what is partially constitutive of explanation. Waskan et al. do several things to help rule out this possibility: across conditions they stipulate that the models all have various objective virtues (such as robustly predicting the explanandum), and they show that the models do not differ in plausibility. At the same time, other features of the studies leave this as a live

option. For instance, the second set of studies indicated that a model was actually intelligible by including the following text:

> After a full year spent doing little else, Dr. Brown finally declared that he was able to detect some high-level (coarse-grained) structures and behavioral patterns that enabled him to make sense of why each distinct new simulation gravitated towards the same end state…a puff of gamma-ray energy, followed by a blast.

When the model was only potentially intelligible, the vignette instead stipulated that 'although no one has yet been able to decipher the relevant high-level variables from the materials Dr. Brown provided, it is only a matter of time before someone does' (p. 1028). If 'high-level (coarse-grained) structures and behavioral patterns' is our feature F, then the differences across vignettes could reflect the difference between actually possessing F, potentially possessing F, and definitely not possessing F.

We raise these worries not to question the value of Waskan et al.'s studies, but to underscore some open questions and to motivate the approach that we adopt in our own studies. Specifically, in an effort to obtain more direct evidence concerning whether understanding is partially constitutive of explanation, our experiments use simpler explanations, thereby preventing participants from making different assumptions about the objective features of the explanations themselves. We also test for effects of mental states (i.e., the presence of understanding) in explanation judgments using a very different approach: by investigating whether attributions of explanations and understanding are susceptible to a 'side-effect effect.' We explain the logic behind this approach in the section that follows.

1.3 Using the SEE as a Marker of the Mental

The side-effect effect (SEE) is by now a well-known result in several areas of philosophy, including the philosophy of action and epistemology. First reported by Knobe (2003), the original finding was that

people's judgments regarding whether a foreseen side-effect of someone's actions was brought about *intentionally* varied depending on whether the foreseen side-effect was good or bad. For example, a businessman who was told his actions in pursuit of profits would bring about environmental harm was judged to have brought about the harm intentionally, whereas a businessman in a symmetrical case of environmental benefit was said not to have brought about the benefit intentionally. Beebe and Buckwalter (2010) extended this finding to the epistemic realm, showing that whether the businessman is judged to have *known* that a consequence would result from his action depends on whether the side-effect is good or bad. Other results (e.g., Knobe 2007, Uttich & Lombrozo 2010) have shown that it is not moral badness per se that leads to higher attributions of intentionality, but rather the violation of some operative norm.

What drives the SEE? Uttich and Lombrozo (2010) develop one account of the effect, according to which it reflects inferences about mental states. They argue that when someone acts in violation of a known, operative norm, one can infer that the person had a reason for action that was sufficiently strong to outweigh the reason to conform to the norm. For instance, from the fact that a businessman was willing to pursue a plan despite its environmental cost, we can infer that he valued money much more highly than the environment. By contrast, an action that is norm-conforming is far less diagnostic of underlying mental states: we have strong evidence that the CEO who harmed the environment values money much more highly than the environment; we have weaker evidence that this is so for the CEO who helps the environment. If this account is right, then the SEE's characteristic asymmetry in attributions reflects differences in the mental states inferred across cases of norm-violation versus norm-conformity. Most importantly for our purposes, the SEE can be used as an index of whether some attribution is sensitive to mental states.

To use the SEE to evaluate psychologism about explanation, we would need to present participants with cases in which a candidate explanation is offered in a norm-conforming versus a norm-violating context. For instance, the businessman could be provided with information about how the proposed

business plan would lead to environmental harm versus benefit, and participants would then be asked to evaluate whether the information that was provided constitutes *an explanation*. If the lay conception of explanation is psychologistic, we would expect greater agreement with the claim that an explanation was provided when the action violated a norm (leading to environmental harm) than when it did not. If the relevant mental state that generates this asymmetry is *understanding*, we would further expect corresponding effects for understanding attribution (i.e., 'the businessman understood how the program would generate environmental harm'). Finally, if the presence or absence of understanding is sufficient to determine whether an explanation has been offered, we would expect that fixing understanding by stipulation should block the effect of norm-conformity versus norm-violation on judgments concerning explanation. These are the predictions that our studies test.

While prior work has not investigated whether an SEE obtains for attributions of understanding, there's reason to believe that there should be one. After Knobe's original demonstration with 'intentionally,' the effect was extended to a wide range of attributions that involved a component desire or pro attitude, including: 'intended', 'desired', 'decided', 'advocated' [5], and 'was in favor of' (for a summary, see Knobe & Pettit 2009). It was then expanded to include psychological attributions that involve measures of belief, including 'knew' (Beebe & Buckwalter 2010) and 'believed' (Alfano, Beebe, & Robinson 2012, Beebe 2013). For instance, participants indicate that the businessman was more likely to *know* about the program's environmental side effects and to *believe* they would occur when those side effects were negative as opposed to positive.[6] The picture that seems to have emerged is that we exhibit an SEE for a

[5] 'Advocated' proves a bit awkward here, as it does not obviously involve any mental state. Pettit and Knobe (2009) assume that advocating something requires having a pro-attitude towards it; if correct, 'advocated' would fall under the general rubric of pro-attitude attribution. There is reason to suspect that this is right—in an unreported pilot experiment, we measured whether claims of the form 'X said Y' exhibit an SEE, and found no significant effect. This suggests that what is doing the work in the case of 'advocated' is the pro-attitude on top of what is said.

[6] One hypothesis is that many instances of the SEE result from an increased willingness to attribute beliefs (Alfano et al 2012, Beebe 2013, Dalbauer & Hergovich 2013), perhaps because norm-violating behavior gives us more information (Uttich & Lombrozo 2010).

broad range of attributions that depend on mental states; it's a short leap from *knowing* that the side effect would occur to *understanding why* it would occur.[7]

While there is no agreed-upon interpretation of what causes the SEE, for our purposes, we only need to accept the relatively minimal assumption that the SEE reflects mental state ascriptions – and even this relatively minimal assumption will be revisited and weakened in §5. If the asymmetric attributions that characterize the SEE reflect attributions of mental states such as understanding, then the presence of an SEE for explanation would provide especially strong evidence for psychologism: we would have evidence that explanation judgments depend on inferred mental states. Moreover, by holding fixed the objective properties of the speech acts under evaluation, we can be more confident that asymmetric attributions reflect inferences about mental states, not inferences about underspecified but potentially objective properties of the explanation.

One might reasonably ask why we use the SEE rather than more directly testing the relationship between explanation and some particular mental state, such as understanding or belief. The key is the generality of the SEE: we can investigate whether explanation is psychologistic without first identifying the full set of relevant mental states. This yields two benefits. First, while we do ultimately tie explanation to one particular mental state (namely understanding), first establishing the more general point that explanation is tied to *some* mental state attribution(s) makes the broader claim independently acceptable. Second, having a marker for mental state attribution more broadly helps us to determine whether and when we have identified a set of mental states that is sufficient to satisfy the mental state requirements for an explanation: when we have succeeded, we should be able to eliminate the SEE for explanation by

[7] There is (to our knowledge) one possible exception to the general claim that attributions of all-and-only mental states exhibit an SEE. Knobe and Fraser (2008) find that people are more likely to say that X's actions *caused* a certain result if the action was norm-violating. However—as anyone who has taught an introductory ethics class can attest—laypeople often conflate causal responsibility with ethical responsibility. As such, we are not inclined to read too much into this result.

stipulating the relevant mental states in this set. Using the SEE thus allows us to determine both *whether* mental states are involved, and also *which* mental states are sufficient.

1.4 Overview of Experiments

Our general methodology will involve using the SEE as a tool for revealing that some inference to a mental state (such as understanding) underlies the application of "explanation." We proceed in three steps. First, in Experiment 1, we show that (as expected) understanding attributions display an understanding epistemic side-effect effect (UESEE). Next, in Experiment 2, we show that explanation attributions display an explanation epistemic side-effect effect (EESEE). Finally, in Experiment 3, we demonstrate experimentally that stipulating the presence or absence of understanding is sufficient to block the influence of norm-conformity versus norm-violation on judgments of explanation – that is, that stipulating understanding eliminates the EESEE.

2. Experiment 1

In this experiment, we extend the ESEE for knowledge-that to knowledge-why, as well as to understanding-that and understanding-why. The ever-widening scope of the SEE suggests that prior results pertaining to knowledge-why should extend to understanding-why. However, some prior research (e.g., Wilkenfeld, Plunkett, & Lombrozo, 2016) has found that attributions of knowledge and understanding can come apart. It was therefore important to demonstrate, rather than assume, that we would observe an SEE for attributions of understanding. Experiment 1 thus lays the groundwork for Experiment 2, in which we investigated whether we observe an SEE for explanation. If psychologism about explanation is correct, and

the mental state that is necessary for a speech act to count as an explanation is *understanding*, then it's important to show that the SEE indeed tracks attributions of understanding.

2.1 Method

*Participants***:** Two-hundred-and-nine participants (133 male, 75 female, 1 other; mean age 32, SD = 9) were recruited through the Amazon Mechanical Turk marketplace (MTurk) and participated in exchange for monetary compensation. In all experiments, participation was restricted to users with an IP address within the United States, with an approval rating of at least 95% based on at least 50 previous tasks, and who had not completed another experiment in the sequence. An additional 14 participants were excluded prior to analysis for failing to consent, failing to complete the experiment, or giving an incorrect response to one of the reading comprehension or attention-check questions (described below).[8]

Materials and Procedure: At the beginning of the experiment, participants were randomly assigned to read one of eight vignettes describing a chairman of a company in the Gizmo industry deciding whether to start a new program, after his vice president informs him of a potential side effect of the decision. The eight vignettes varied along three dimensions: whether the side effect would comply with or violate some operative norm (2: *conform*, *violate*), whether or not the vice president described a mechanism that brought about the side effect (2: *mechanism absent*, *mechanism present*), and whether the norm in question was moral or conventional (2: *moral*, *conventional*). The *mechanism* variable was introduced to ensure that understanding judgments would not be near floor, and to ensure that any effects were robust across levels of mechanistic detail. The type of norm variable (moral versus conventional) was introduced to ensure the generality of any effects.

[8] Excluding participants on the basis of an attentional or comprehension check is common practice in psychology, and our exclusion rate is not out of line with prior research. Given the probability of internet participants multi-tasking or otherwise not devoting full attention, it is standard practice to recruit a large number of participants (Crump, McDonnell, & Gureckis, 2013; Oppenheimer, Meyvis, & Davidenko, 2009), with the expectation that surveys tracking more nuanced differences will eliminate upwards of 40% (Downs, Holbrook, & Sheng, 2010).

To illustrate, the vignette involving a conventional violation with the mechanism present read as follows [with the text for the norm-conforming variant in brackets]:

> The convention in the Gizmo industry is for Gizmos to be a light [dark] color. Specifically, the convention is to make Gizmos that are colored lighter [darker] than navy blue. This color preference had been the standard for decades, and it was well-supported by sales and tradition. The original founder of the Gizmo industry insisted on the light [dark] color as a characteristic feature of the product.
>
> The vice-president of a company in the Gizmo industry 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 result in our Gizmos being colored black. The convention is to make Gizmos colored lighter [darker] than navy blue, so we would be violating [complying with] the convention.’
>
> The vice-president went on to describe in painstaking detail the mechanism by which the Gizmos would appear black. Among other things, he told the chairman that the new program would involve making a product that would require a new form of sealant, the application of which would lead to oxidation of the metal the Gizmo is made out of, which would lead to the whole thing taking on a black color.
>
> The chairman of the board answered, ‘‘I don’t care at all about the color of the Gizmos. I just want to make as much profit as I can. Let’s start the new program.’
>
> They started the program. As it happened, the Gizmos were black, darker than navy blue and violating [complying with] the convention.

In the *moral* condition, the norm specified an acceptable risk for increasing carbon emissions, and the proposed program either conformed to or violated (exceeded) the norm. The *mechanism absent* vignettes

were identical to the corresponding *mechanism present* vignettes, except that the paragraph describing the mechanism (beginning with 'The vice-president went on to describe…') was omitted entirely.

After reading one of the eight vignettes, all participants were asked to rate their agreement with each of the following six claims, presented in randomized order. The first five were treated as dependent variables of interest, whereas the sixth was used as a measure of reading comprehension for screening participants:

a. The chairman *knew why* the new program would [harm the environment/lead to the Gizmos being black].
b. The chairman *knew that* the new program would [harm the environment/lead to the Gizmos being black].
c. The chairman *understood why* the new program would [harm the environment/lead to the Gizmos being black].
d. The chairman *understood that* the new program would [harm the environment/lead to the Gizmos being black].
e. The chairman intentionally [harmed the environment/caused the Gizmos to be black].
f. [The environment was harmed as a result of the new program./The Gizmos produced as part of this program were black.]

Answers were given on a 7-point Likert scale ranging from *Strongly Disagree* to *Strongly Agree* (coded as 1 and 7, respectively). Participants then had to complete an attention check modeled after Oppenheimer, Meyvis, and Davidenko (2009). To conclude, they were asked demographic and debriefing questions. Participants who made errors on the comprehension question (by not responding above the scale midpoint) or on the attention check were excluded from further analysis.

2.2 Results

First, we verified that we could replicate the SEE with our modified vignettes by testing whether there was a significant effect of norm status on intentionality attributions. A univariate ANOVA with norm status (2: *conform, violate*), norm type (2: *moral*, *conventional*) and mechanism (2: *mechanism absent, mechanism present*) as between-subjects factors revealed the predicted significant effect of norm status on intentionality judgments, $F(1, 201) = 9.425$, $p = .002$, $\eta_p^2 = .045$, with significantly lower ratings in the *conform* condition (N = 101, M = 4.40, SD = 1.955) than in the *violate* condition (N = 108, M = 5.24, SD = 1.884), and no significant interactions. This replicates prior research, and corresponds to a small-to-medium effect size.

Having established the standard SEE, we tested whether we would observe an effect of norm conformity on attributions of knowledge and understanding. Knowledge and understanding attributions were analyzed with a mixed ANOVA with attribution type (2: *understanding*, *knowledge*) and attribution object *(*2: *why, that*) as within-subjects factors, and with norm status (2: *conform, violate*), norm type (2: *moral*, *conventional*) and mechanism (2: *mechanism absent, mechanism present*) as between subjects factors (see Figure 1).

The mixed ANOVA revealed several main effects. Most importantly, participants provided significantly higher ratings in the *violate* conditions than in the *conform* conditions, $F(1, 201) = 12.531$, $p < .001$, $\eta_p^2 = .059$. This effect was moderate in size, and it was not qualified by interactions with attribution type, $F(1, 201) = .054$, $p = .817$, $\eta_p^2 = .000$, nor attribution object, $F(1, 201) = .190$, $p = .663$, $\eta_p^2 = .001$. This suggests that we succeeded in replicating the ESEE found in prior work involving knowledge-that, and also found a comparable SEE for knowledge-why, understanding-that, and understanding-why.

There were additional significant effects that do not bear on our central hypotheses. The analysis revealed a main effect of mechanism, $F(1, 201) = 29.365$, $p < .001$, $\eta_p^2 = .127$, with higher ratings in the *mechanism present* conditions than in the *mechanism absent* conditions, and a main effect of attribution object, $F(1, 201) = 78.156$, $p < .001$, $\eta_p^2 = .280$, with higher ratings for '*that*' statements than for '*why*'

statements. In addition, there were two significant interactions. First, there was a significant interaction between norm type and attribution, $F(1, 201) = 4.133$, $p = .043$, $\eta_p^2 = .020$: participants always gave higher ratings for *moral* than for *conventional* norms, but the difference was greater for *knowledge* than for *understanding*. Second, there was a three-way interaction between attribution object, conformity, and mechanism, $F(1, 201) = 4.298$, $p = .039$, $\eta_p^2 = .021$: when the mechanism was present, it disproportionately increased the effect of norm-violation on ratings of why-statements. (We hypothesize that the presence of a mechanistic description made participants more sensitive to any difference between thinking about *why* something happened versus thinking *that* something happened.)

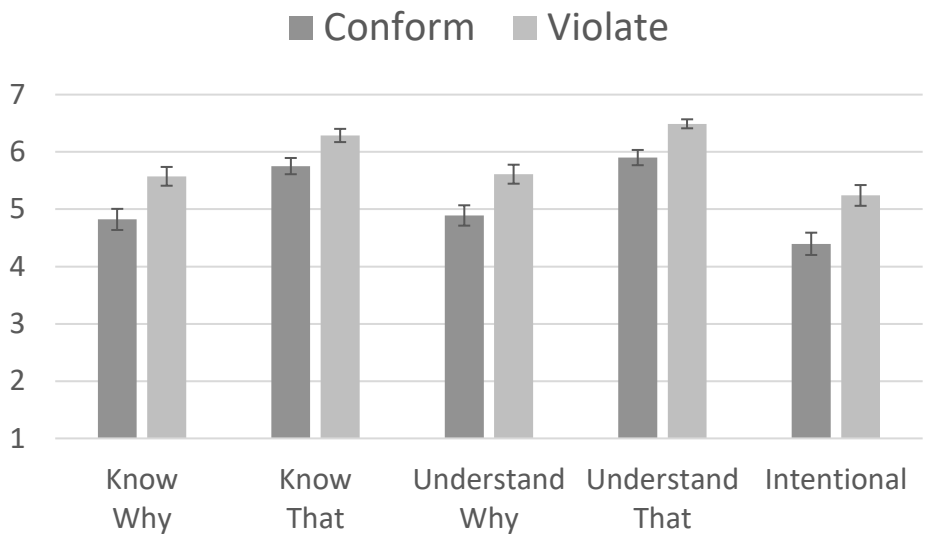

Figure 1: Mean attribution ratings from Experiment 1 as a function of norm status, attribution type, and attribution object, Error Bars ± 1 SEM

2.3 Discussion

The main finding of Experiment 1 was that there was an observed SEE for understanding, and, outside of a three-way interaction that indicates mechanistic detail somehow exacerbates the effect for 'why'

statements only, no significant interactions between the SEE and any other variables. Most importantly, there was no interaction between conformity and attribution type nor between conformity and attribution object—the SEE manipulation affected knowledge and understanding—as well as that- and why-statements—alike. This suggests that the SEE can reflect attributions of understanding as well as attributions of knowledge.[9]

3. Experiment 2

In Experiment 1, we extended the range of the SEE to include knowledge-why, understanding-why, and understanding-that. As pretheoretically it is plausible that these are constituted by or at least partially supervene on mental states, the findings of Experiment 1, while novel, were not necessarily surprising. Conversely, explanation is standardly taken to be an objective state, devoid of reference to individuals' mental states (for exceptions to this position, see §1). It would thus be relatively surprising to show that the SEE influenced judgments of the form 'X offered an explanation why P', which is precisely what we set out to demonstrate in Experiment 2. Since attribution object and mechanism for the most part did not interact with the SEE manipulation of interest (norm status) in Experiment 1, they were dropped from further study. We were also concerned that somehow being in a *violate* condition prompted participants to give higher ratings for every possible question—to that end, a control question that clearly did not involve anything mental was added.

3.1 Method

[9] As an added boon, the results of Experiment 1 support the contention that understanding states are evaluated by the same mechanisms that we use to evaluate (other) mental states. This result is a prediction of some (more mentalistic) views of understanding (e.g., Wilkenfeld 2013, Kelp 2015), but a surprising (though not necessarily inconsistent) result on others (e.g., de Regt & Dieks 2005), and directly opposed to still others (e.g., Ylikoski 2009).

Participants: Two-hundred-and-ninety-two participants (176 male, 113 female, 2 other, 1 blank; mean age 33, SD = 9) were recruited through the Amazon Mechanical Turk marketplace (MTurk) as in Experiment 1 and participated in exchange for monetary compensation. An additional 41 participants were excluded prior to analysis following the same procedure used in Experiment 1. Participants who had participated in Experiment 1 were not eligible to participate in Experiment 2.

Materials and Procedure: At the beginning of the experiment, participants were randomly assigned to read one of four vignettes similar to those from Experiment 1. The vignettes varied along two dimensions: whether the side-effect would comply with or violate some operative norm (2: *conform*, *violate*), and whether the norm in question was moral or conventional (2: *moral*, *conventional*).

To allow for variation in participants' judgments about whether an explanation had been offered, each vignette included a minimal description of the mechanism. For example, in the conventional violation condition, participants read a passage including the following [text for corresponding *conform* variant in brackets]:

> The vice-president of a company in the Gizmo industry 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 result in our Gizmos being black. The convention is to make Gizmos lighter [darker] than navy blue, so we would be violating [complying with] the convention.'
>
> The vice-president went on to give a brief description of the mechanism by which the Gizmos would appear black. He told the chairman that the new program would involve making a product that would require a new chemical procedure, which would lead to the whole thing taking on a blackish color.

After reading one of the four vignettes, all participants were asked to rate their agreement with the following four claims (in randomized order). The first three were treated as dependent variables of interest, and the fourth was used as a measure of reading comprehension to screen participants:

a. The vice president offered the chairman *an explanation why* the new program would [harm the environment/lead to the Gizmos being black].
b. The chairman *understood why* the new program would [harm the environment/lead to the Gizmos being black].
c. The chairman *intentionally* [harmed the environment/caused the Gizmos to be black].
d. [The environment was harmed as a result of the new program./The Gizmos produced as part of this program were black.]

Participants then advanced to a new page on which they were presented with the following information:

> As it happens, the secretary in the next room was testing a new audio recorder that would record all interactions in the chairman's office. It wasn't working very well, and only successfully recorded about half the time. How likely do you think it is that the following statement is true?
>
> e. The audio recorder captured *the vice president's description of* the new program.

Participants rated the likelihood on a 7-point scale. This question was included as a control to ensure that participants were not providing higher ratings in the *violate* conditions indiscriminately. The content of the item was designed to match the provision of an explanation in terms of its physical characteristics (i.e., a spoken signal transmitted from one source to another), but without the potential role for mental states that explanation could involve.

Finally, participants completed an attention check as in Experiment 1, and were asked demographic and debriefing questions. Participants who made errors on the comprehension question or attention check were excluded from further analysis.

3.2 Results

First, we analyzed the intentionality question to ensure that we replicated the traditional SEE, and the control 'recorder' question to ensure that participants were not showing an SEE indiscriminately. For the intentionality variable, a univariate ANOVA with norm status (2: *conform*, *violate*) and norm type (2: *moral*, *conventional*) as between-subjects factors revealed the predicted effect of norm status, $F(1, 288) = 20.467$, $p < .001$, $\eta_p^2 = .066$, with significantly lower ratings in the *conform* condition (N = 149, M = 4.21, SD = 1.971) than in the *violate* condition (N = 143, M = 5.13, SD = 1.900). This involved a medium effect size, and no interactions with other variables. For the recorder question, a *t*-test comparing responses to the recorder question for *conform* (N = 149, M = 4.19, SD = 1.170) versus *violate* (N = 143, M = 4.39, SD = 1.114) revealed no significant effect, $t(290) = -1.522$, $p = .129$, $d = .18$. These results jointly indicate that the SEE manipulation was successful, but that it did not cause *all* ratings to shift indiscriminately.

We next turned to the dependent variables of greatest interest: explanation and understanding. Responses were analyzed with a mixed ANOVA with attribution type (2: *explanation*, *understanding*) as a within-subjects factor and norm status (2: *conform*, *violate*) and norm type (2: *moral*, *conventional*) as between-subjects factors (Fig. 2). This analysis revealed a main effect of norm status, $F(1, 288) = 9.151$, $p = .003$, $\eta_p^2 = .031$, with significantly higher ratings in the *violate* condition than in the *conform* condition. This main effect was not qualified by an interaction with attribution type, $F(1, 288) = 1.750$, $p = .187$, $\eta_p^2 = .006$, which suggests that the effect of conformity applied to understanding and explanation judgments alike, and with a small-to-medium effect size.

The analysis revealed two additional significant main effects, neither of which bears on our central hypotheses. Ratings for *explanation* were significantly higher than for *understanding, F*(1, 288) = 39.026, $p < .001$, $\eta_p^2 = .119$, and participants gave significantly higher ratings for *moral* norms than for *conventional* norms, $F(1, 288) = 4.029$, $p = .046$, $\eta_p^2 = .014$**.** There was also an interaction between attribution type and norm type: participants were more likely in the moral case to say that the chairman understands, but less likely to say that the VP explained, $F(1, 288) = 59.897$, $p < .001$, $\eta_p^2 = .117$.

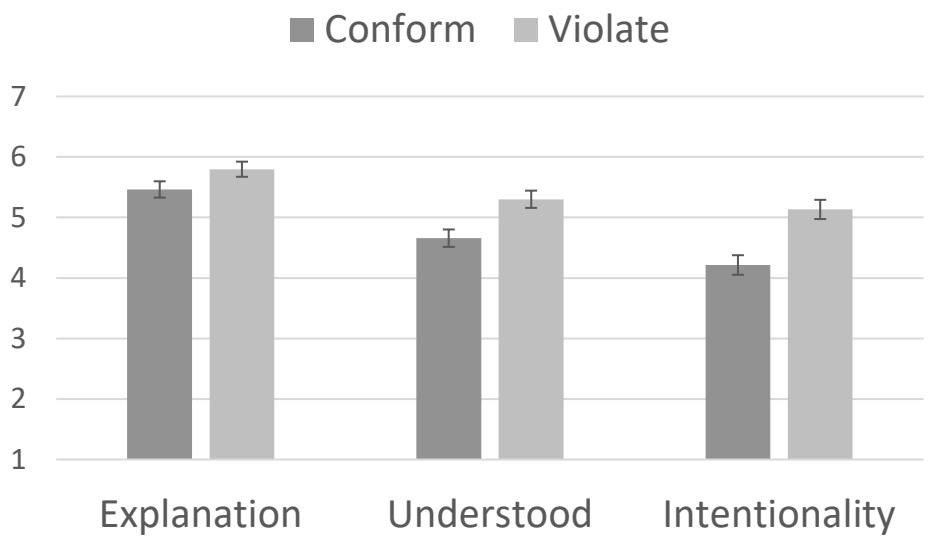

Figure 2: Mean attribution ratings from Experiment 2 as a function of norm status and judgment. Error Bars ± 1 SEM

Table 1

Explanation and Understanding Attributions in Experiment 2 by Norm Status and Norm Type

	Norm Status	Norm Type	Mean	Std. Deviation	N
Explanation Attribution	Conform	Conventional	5.64	1.646	77
		Moral	5.28	1.638	72
		Total	5.46	1.646	149
	Violate	Conventional	5.94	1.371	70
		Moral	5.66	1.601	73
		Total	5.80	1.494	143
	Total	Conventional	5.78	1.524	147
		Moral	5.47	1.625	145
		Total	5.63	1.580	292
Understanding Attribution	Conform	Conventional	4.30	1.606	77
		Moral	5.04	1.850	72
		Total	4.66	1.762	149
	Violate	Conventional	4.70	1.688	70
		Moral	5.88	1.527	73
		Total	5.30	1.708	143
	Total	Conventional	4.49	1.653	147
		Moral	5.46	1.740	145
		Total	4.97	1.762	292

While these results indicate that attributions of explanation are susceptible to a SEE, consistent with psychologism, they do not establish the more specific claim that *understanding* is the mental state that partially constitutes explanation. As initial support for this stronger claim, we can investigate whether norm conformity affected explanation attributions because norm conformity influenced attributions of understanding, which in turn influenced attributions of explanation. In other words, we can test for whether the effect of norm conformity on explanation attributions was *mediated* by attributions of understanding. To do so, we used the SPSS PROCESS macro written by Andrew Hayes to test for an indirect effect of norm status on explanation *via understanding*; this test indeed revealed significant mediation

(95% confidence interval .0768 to .3743). A hierarchical regression revealed that once attributions of understanding were included in the model, incorporating norm status did not significantly improve predictions of attributions of explanation, with a change in R-squared of .002, p = .470. Conversely, the PROCESS macro revealed that the indirect effect of explanation on understanding was not significant (95% confidence interval -.0065 to .2735), suggesting that there was no significant mediation in the other direction. In other words, we failed to find support for an alternative hypothesis according to which attributions of understanding depend on attributions of explanation. This provides some additional statistical confirmation that participants' judgments of whether an explanation had been offered were not merely correlated with judgments of understanding or driven by a shared cause, but actually depended on understanding judgments themselves, as predicted by psychologism.

3.3 Discussion

Experiment 2 found that manipulating whether a norm was adhered to or violated affects not only whether someone is said to have understanding, but also whether someone *else* is taken to have provided an explanation.[10] This finding supports a psychologistic account of laypeople's conception of explanation:

[10] One commenter raised the question of whether the SEE for explanation appeared not because participants were psychologistic about *explanation*, but rather because they were psychologistic about whether something had been *offered*. To test this hypothesis, we ran a supplementary study (focusing on the moral case) in which we varied whether participants evaluated the statement "The vice-president offered the chairman an *explanation of why* the new program would harm the environment" or "What the vice-president said amounted to *an explanation of why* the new program would harm the environment." We predicted that we would continue to find an SEE, and that it would not interact with wording choice. This is what we found in an initial study with 639 participants (post-exclusion). Due to a typo found in that survey (one statement included "help/harm the environment" rather than simply "harm the environment"), we then ran another 614 participants (post-exclusion) with a corrected copy. We analyzed the combined sample with an ANOVA on explanation rating (i.e., people's agreement with whichever statement they saw) with norm status (2: *conform, violate*), wording (2: *offered, amounted*) and survey number (2: *survey 1, survey 2*) as between-subjects factors. This analysis revealed significantly higher ratings in the *violate* condition (N = 625, M = 5.26, SD = 1.799) than in the *conform* condition (N = 628, M = 4.94, SD = 1.837), F(1, 1245) = 9.79, p = .002, η_p^2 = .008), with no interaction between norm status and either wording, F(1, 1245) = 1.37, p = .242, or survey number, F(1, 1245) = 2.071, p =.150, and no three-way interaction, F(1, 1245) = .409, p = .523. Interestingly, there *was* a main effect of wording, with participants giving higher ratings for *offered* (N = 626, M = 5.43, SD =1.674) than *amounted* (N = 627, M = 4.78, SD = 1.913), F(1, 1245) = 39.850, p < .001, η_p^2 = .031. It is perhaps surprising that it is easier to *offer* an

explanation judgments were susceptible to an SEE, which we take as evidence for an influence of mental state inferences. Moreover, mediation analyses suggest that explanation judgments were affected by understanding judgments, and not conversely. Finally, a new control item further demonstrated that the SEE for explanation did not result from a promiscuous attribution-increasing tendency when faced with a story about norm violation, but instead reflects something particular about attributions involving mental states.

4. Experiment 3

Experiments 1 and 2 demonstrate that judgments concerning explanation and understanding are susceptible to an SEE. However, several questions remain open. Most crucially, our interpretation of these results (as evidence for psychologism about explanation) rests on two assumptions: that the SEE for explanation reflects an effect of mental state inferences, and that the relevant mental state is understanding. In Experiment 3 we verify our interpretation by experimentally testing whether stipulating the presence or absence of understanding affects judgments concerning explanation, and whether doing so fully blocks the effect of norm status on explanation judgments. If explanation judgments instead support an SEE for some reason related to mental states other than understanding, or via some different mechanism altogether, we would instead expect an explanation SEE to manifest even though understanding has been stipulated.

A second aim of Experiment 3 was to investigate *whose* understanding is relevant to explanation judgments: the person providing an explanation or the recipient of the explanation. In Waskan et al.'s vignettes, the scientist who produced the explanatory model was also its recipient. However, their predictions and interpretation suggest that the crucial feature of their vignette is that actual understanding

---

explanation than to have what one says *amount to* an explanation, but this finding is orthogonal to the present concern.

is achieved; not that it is achieved by any particular recipient. In Experiment 3, we further manipulate —by stipulation – whether the *explanation provider* achieved understanding and/or whether the *recipient* achieved understanding to test whether the explanation's intended recipient is somehow privileged.

4.1 Method

Participants: Three-hundred-and-eighty-two participants (229 male, 151 female, 2 other; mean age 32, SD = 9) were recruited through the Amazon Mechanical Turk marketplace (MTurk) as in Experiments 1-2 and participated in exchange for monetary compensation. An additional 67 participants were excluded prior to analysis following the same procedure used in Experiments 1 and 2. Participants in any other experiment in the sequence were not eligible to participate in Experiment 3.

Materials and Procedure: At the beginning of the experiment, participants were randomly assigned to read one of eight vignettes similar to those in Experiment 1 and 2. The vignettes varied along three dimensions: whether the side-effect would conform with or violate an operative environmental norm (2: *conform*, *violate*), whether the vice-president understood why it would have the effect that it would have (2: *VP understood*, *VP did not understand*), and whether the chairman understood why it would have that effect (2: *Chairman understood*, *Chairman did not understand*). Given the number of conditions involved, we dropped a manipulation of norm type; all vignettes involved environmental/moral norms.

The materials were exactly the same as those in the moral conditions of Experiment 2, with the exception that explicit statements were added stipulating whether the vice-president and/or the chairman understood why the program would harm the environment (while being clear that in every case the person *believed* that it would, to avoid inadvertently manipulating the agents' beliefs, as well).[11] For example, in

[11] While the finding of an UESEE on Alfano et al's (2012) interpretation does suggest that belief is a component of understanding, it does not imply that it is the only component—therefore, understanding could still be manipulated separately.

the both-understand-violate condition, participants read the following [text for corresponding comply variant in brackets]:

> There is a regulatory agency for the Gizmo industry that exists in order to provide environmental standards, even though it does not have the authority to ensure compliance with these standards. The regulatory agency has established an environmental standard, which states that a company may only start new programs if the company's total increase of carbon emissions would be less than 5% [45%] of the emissions from the previous year, since carbon emissions cause environmental harm. This 5% [45%] limit has been the standard for decades.
>
> The vice-president of a company in the Gizmo industry 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 increase our carbon emissions by 25%. The industry standard is only to start programs of this type when they would increase carbon emissions by less than 5% [45%], so we would be violating [complying with] the standard.'
>
> The vice-president went on to give a brief description of the mechanism by which the program might potentially harm the environment. He told the chairman that the new program would involve increasing production, which would lead to a 25% increase in carbon emissions. He was speaking from personal understanding—the vice-president understood why the new program would lead to the environment being harmed (and he was completely confident in the prediction that the new program would in fact lead to a 25% increase in carbon emissions).
>
> The chairman of the board listened to what the vice-president had to say, and understood why the new program would lead to the environment being harmed. He was also completely confident in the prediction that the new program would in fact lead to a 25% increase in carbon emissions. The chairman answered, ''I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'

They started the program. As predicted, the environment was harmed.

In the versions in which the chairman and VP were stipulated not to understand why the new program would lead to the environment being harmed, the corresponding paragraphs read as follows:

> The vice-president went on to give a brief description of the mechanism by which the program might potentially harm the environment. He told the chairman that the new program would involve increasing production, which would lead to a 25% increase in carbon emissions. He was not speaking from personal understanding—the vice president did not himself understand why the new program would lead to the environment being harmed (but he was completely confident in the prediction that the new program would in fact lead to a 25% increase in carbon emissions).
>
> The chairman of the board listened to what the vice-president had to say, but did not really understand why the new program would lead to the environment being harmed. Nonetheless, he was also completely confident in the prediction that the new program would in fact lead to a 25% increase in carbon emissions. The chairman answered, ‘‘I don’t care at all about harming the environment. I just want to make as much profit as I can. Let’s start the new program.”

After reading one of the eight vignettes, all participants were asked to rate their agreement with the following claim:

a. The vice president offered the chairman an explanation of why the new program would harm the environment.

To ensure that this initial rating was not affected by first rating understanding, all other dependent variables were moved to a second page. The other statements were:

b. The chairman *understood why* the new program would harm the environment.

c. The vice-president *understood why* the new program would harm the environment.

d. The chairman *intentionally* harmed the environment.

e. The environment was harmed as a result of the new program.

As before, b-d were treated as dependent variables or primary interest, whereas e was treated as a reading comprehension question to screen participants.

Finally, participants completed an attention check and were asked demographic and debriefing questions. Participants who made errors on the comprehension question or attention check were excluded from further analysis.

4.2 Results

We first analyzed participants' attributions of understanding to the chairman and the vice-president to verify that our experimental manipulation of understanding was effective, and in particular that participants accepted the stipulated levels of understanding. To do so, we analyzed understanding attributions with target as a within-subjects factor (2: VP, chairman), and the stipulated understanding for the VP (2: present, absent) and chairman (2: present, absent), as between-subjects factors. This analysis revealed the predicted main effects: participants gave higher mean attributions when VP understanding was stipulated, $F(1, 378) = 196.113$, $p < .001$, $\eta_p^2 = .342$, and when the chairman's understanding was stipulated, $F(1, 378) = 92.288$, $p < .001$, $\eta_p^2 = .196$. There was also a main effect of target of attribution, with participants attributing higher understanding to the VP than to the chairman, $F(1, 378) = 61.065$, $p < .001$, $\eta_p^2 = .139$.

Table 2

Attributed Understanding in Experiment 3 as a Function of Stipulated Understanding

	Stipulated VP Understanding	Stipulated Chairman Understanding	Mean Understanding Attribution	SD	N
Chairman's Understanding	Yes	Yes	5.86	1.378	99
		No	3.02	1.834	89
		Total	4.52	2.143	188
	No	Yes	4.32	2.065	96
		No	2.37	1.689	98
		Total	3.34	2.120	194
	Total	Yes	5.10	1.908	195
		No	2.68	1.785	187
		Total	3.92	2.209	382
VP's Understanding	Yes	Yes	6.27	.901	99
		No	5.96	1.215	89
		Total	6.12	1.070	188
	No	Yes	3.41	2.014	96
		No	3.06	2.065	98
		Total	3.23	2.042	194
	Total	Yes	4.86	2.112	195
		No	4.44	2.241	187
		Total	4.65	2.183	382

These main effects were qualified by the expected two-way interactions: when it was stipulated that the VP understood, attributions of understanding for the VP rose more sharply than did attributions of understanding for the chairman, $F(1,378) = 79.737$, $p < .001$, $\eta_p^2 = .174$, and when it was stipulated that the chairman understood, attributions of understanding for the chairman rose more sharply than did attributions of understanding for the VP, $F(1, 378) = 106.678$, $p < .001$, $\eta_p^2 = .220$. Finally, there was a small but significant three-way interaction whereby it did not matter for attributions of understanding to the VP whether it was also stipulated that the chairman understood, whereas stipulating that the VP understood

affected participants who were told that the chairman understood more than it affected those who were told that the chairman did not understand, $F(1, 378) = 5.158$, $p = .024$, $\eta_p^2 = .013$). (We hypothesize that people showed a particular resistance to the stipulation that the chairman understood when it had already been stipulated that the VP did not.) Moreover, the levels of attributed understanding were above the scale-midpoint when understanding was stipulated to be present, and below the scale mid-point when stipulated to be absent.

As in previous experiments, we also confirmed that the traditional SEE was replicated: a univariate ANOVA of intentionality ratings with the stipulated understanding of the VP (2: *present, absent*), the stipulated understanding of the chairman (2: *present, absent*), and norm status (2: *conform, violate*) as between-subjects factors revealed the predicted main effect of norm status, $F(1, 374) = 19.973$, $p < .001$, $\eta_p^2 = .051$, with significantly lower ratings in the *conform* condition (N = 189, M =5.21, SD = 1.639) than the *violate* condition (N = 193, M = 5.88, SD = 1.289), and with a small-to-medium effect size. There was also a small interaction between norm status and VP understanding $F(1, 374) = 4.659$, $p = .032$, $\eta_p^2 = .012$. This finding is neither predicted by nor in any way in opposition to our argument, though it does prefigure an important later finding that people put a great deal of stock in the presence or absence of understanding on the part of the VP.

Having established the efficacy of the experimental manipulation and having replicated the standard SEE, we next analyzed our dependent variable of primary interest: explanation ratings. Ratings were analyzed as the dependent variable in an ANOVA with the stipulated understanding of the VP (2: *present, absent*), the stipulated understanding of the chairman (2: *present, absent*), and norm status (2: *conform, violate*) as between-subjects factors (see Figure 3 and Figure 4).

If explanation judgments are a function of understanding, we would expect to find main effects of the VPs stipulated understanding and/or of the chairman's stipulated understanding. Both effects were observed: participants were more likely to indicate that the vice-president offered an explanation when he

(by stipulation) understood himself (N = 188, M = 5.26 SD = 1.649) than when he did not (N = 194, M = 3.50, SD = 2.089), $F(1, 374) = 85.230$, $p < .001$, $\eta_p^2 = .186$ (see Figure 3a), and participants were more likely to say that the VP offered an explanation when the chairman (by stipulation) understood (N = 195, M = 4.83, SD = 1.902) than when he did not (N = 187, M = 3.89, SD = 2.151), $F(1, 374) = 21.909$, $p < .001$, $\eta_p^2 = .055$ (see Figure 3b). These factors did not interact with each other.

If attributions of understanding are sufficient to screen off attributions of other mental states when attributing explanation, then we would expect the effect of norm conformity to vanish when understanding is fixed. This prediction was also confirmed: there was not a significant main effect of norm conformity on explanation ratings (*violate*: N = 193, M = 4.40, SD = 2.115, *conform*: N = 189, M = 4.33, SD = 2.047), $F(1, 374) = .189$, $p = .664$, $\eta_p^2 = .001$ (see Figure 4). Notably, though, there was one significant interaction: between norm conformity and chairman understanding. When chairman understanding was present, the ratings were numerically higher for violate than for conform. When chairman understanding was absent, the ratings were numerically higher for conform than for violate. This accounts for the significant interaction, but in no case was the difference between *violate* and *conform* itself significant.

Table 3

Explanation Ratings in Experiment 3 by Norm Status and Stipulated Understanding

Norm Status	Stipulated VP Understanding	Stipulated Chairman Understanding	Mean Explanation Attribution	SD	N
Conform	Yes	Yes	5.53	1.356	49
		No	4.96	1.609	45
		Total	5.26	1.502	94
	No	Yes	3.55	2.052	47
		No	3.27	2.171	48
		Total	3.41	2.106	95
	Total	Yes	4.56	1.988	96
		No	4.09	2.089	93
		Total	4.33	2.047	189
Violate	Yes	Yes	5.86	1.262	50
		No	4.59	2.061	44
		Total	5.27	1.791	94
	No	Yes	4.29	1.904	49
		No	2.90	2.033	50
		Total	3.59	2.080	99
	Total	Yes	5.08	1.788	99
		No	3.69	2.205	94
		Total	4.40	2.115	193
Total	Yes	Yes	5.70	1.313	99
		No	4.78	1.845	89
		Total	5.26	1.649	188
	No	Yes	3.93	2.001	96
		No	3.08	2.099	98
		Total	3.50	2.089	194
	Total	Yes	4.83	1.902	195
		No	3.89	2.151	187
		Total	4.37	2.079	382

Finally, it's worth noting the effect sizes associated with the vice president's understanding (η_p^2 = .186) and the chairman's understanding (η_p^2 = .055). While both effects were significant, the former had a large effect size, while the latter had a medium effect size.

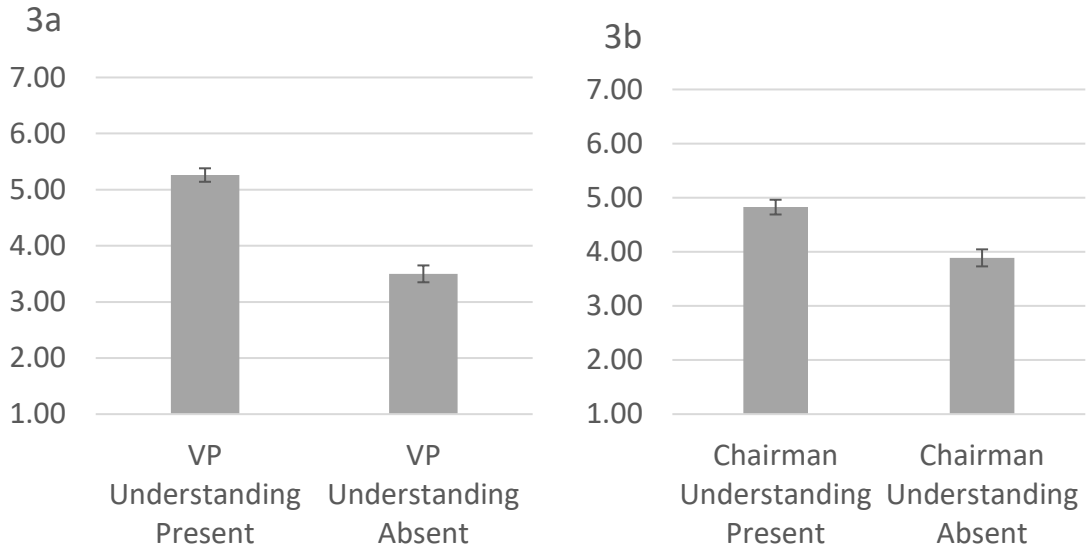

Figure 3: Mean explanation ratings as a function of the stipulated understanding of the VP (3a) and as a function of the stipulated understanding of the chairman (3b). Error Bars ± 1 SEM

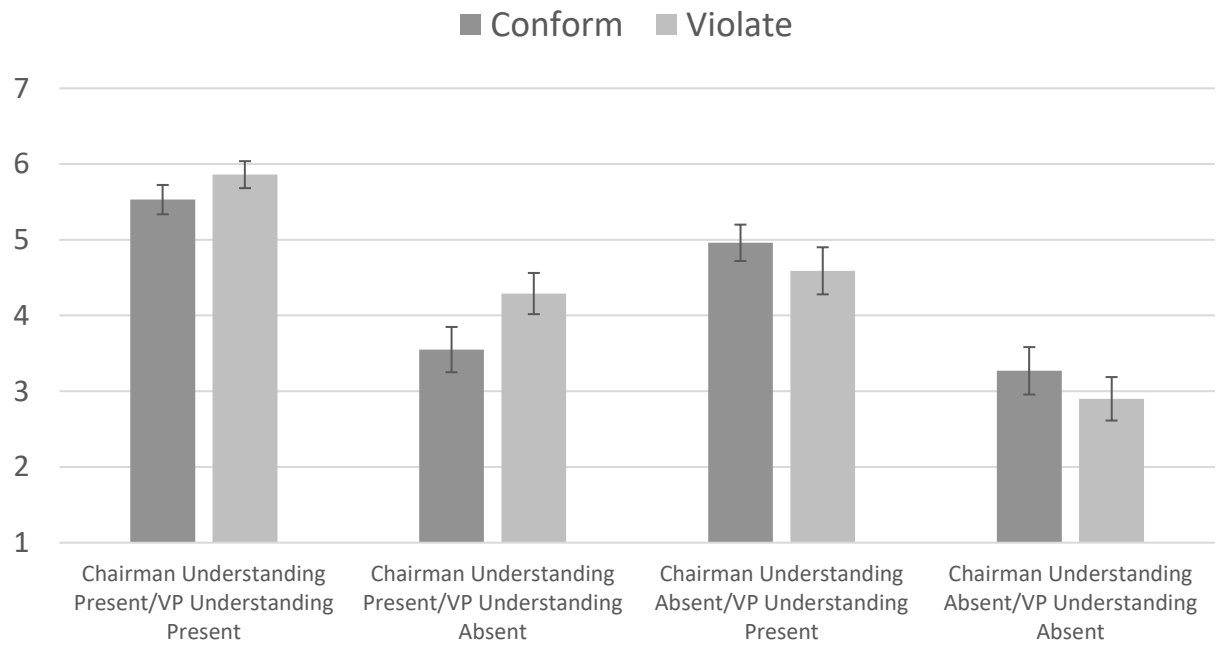

Figure 4: Mean explanation rating as a function of both stipulated chairman understanding and stipulated VP understanding. No differences based on conformity are significant. Error Bars ± 1 SEM

4.3 Discussion

Experiment 3 confirmed our interpretation of Experiment 2, suggesting that attributions of explanation are in fact affected by attributions of understanding. Specifically, Experiment 3 went beyond the mediation analysis in Experiment 2 by demonstrating a causal effect of stipulated understanding on explanation judgments. The results also went beyond the initial experiments in showing that the vice-president and the chairman's understanding were *both* relevant, though to differing degrees. Perhaps most interestingly, given the robustness of the SEE as a measure of mental states exhibited in Experiments 1 and 2 as well as elsewhere, there was no significant SEE exhibited for explanation once understanding was stipulated. This suggests that stipulating understanding is at least *sufficient* to block the influence of other mental states. If, for example, the audience's *desires* were partially constitutive of explanation (and not captured somehow by attributions of understanding), then we would have expected an SEE even when understanding was held fixed.That the speaker's understanding is more influential than the listener's is a somewhat surprising result, even for psychologistic theories of explanation. Wilkenfeld (2014, p. 3371) is explicit that the relevant understanding when determining whether an explaining act has taken place (and hence, derivatively, whether one has an explanation) is the understanding of the audience. Similarly, while Achinstein (1983) argues that the intent of the speaker is relevant, the relevant intent to count as explaining is that which takes into account the potential understanding of the audience. Waskan et al. (2014) remain neutral on whose understanding is relevant, and so are consistent with the paramount importance of the speaker's understanding, but do not predict these results.

4.4 General Discussion

Across three experiments, we find support for psychologism. First, we demonstrate by using the SEE that attributions of understanding are modulated by the same kinds of mental state inferences that affect attributions of intentional action, knowledge, and a variety of other mentalistic terms. We then demonstrate that attributions of explanation are similarly susceptible to an SEE. The dependence of

explanation on understanding is tested directly in Experiment 3: explanation judgments closely track stipulated understanding. Since explanation judgments do not exhibit an SEE when understanding is held fixed, there is reason to believe that understanding attribution is such a powerful driver of explanation attribution that it swamps other mental state attributions that might have been expected to produce an independent SEE.[12]

Taken together, these findings support accounts of explanation that situate explanation with regard to understanding and understanders at the expense of accounts that favor exclusively formal or objective criteria. This is a potentially surprising finding, as the majority of historical accounts have identified explanations with some privileged structure (e.g., Friedman 1974, Hempel 1965, Kitcher 1989) or content (e.g., Bechtel 2008, Machamer, Darden, & Craver 2000, Salmon 1971, Salmon 1984, Woodward 2003). Wilkenfeld and Lombrozo (2015) argue that even some putatively pragmatic accounts (Bromberger 1966, Garfinkel 1981, Van Fraassen 1980) ultimately ground explanation as a product of structure and content. By contrast, a minority view (e.g., Achinstein 1983, Scriven 1962, Wilkenfeld 2014) identifies explanations with some cognitive upshot (i.e., generating understanding), and so accord far better with the data collected here.[13]

Of course, one could dismiss the present data as irrelevant to the philosophical project of characterizing explanation. Our data certainly fall short of supporting normative conclusions about how explanations ought to be classified, and our stimuli were designed to reflect everyday explanations rather than explanations within science. That said, we think these findings should not be so easily dismissed, for three reasons. First, it is methodologically valuable to at least begin with the principle that the lay concept

[12] Our framing of the discussion in terms of understanding attributions sidesteps the question of whether these patterns of behavior speak to the nature of the concept or to how people deploy it. This is arguably a feature, as Machery (2008) persuasively argues that such questions might be unanswerable given the present state of the philosophy of concepts.

[13] Of these views, Achinstein (1983) grounds explanation in the intent to produce understanding rather than in the success of such production; Scriven (1962) and Wilkenfeld (2014) are not committed to this role for intentions.

is continuous with the technical notion of explanation, as, if true, that assumption would yield a more unified view.[14] We are inclined to follow a maxim of David Lewis (1980, 217): "We might settle for less, but let's start by asking for all we want." Second, at some level, projects of conceptual clarification must *eventually* answer to the tribunal of usage, or risk clarifying a totally different concept than the one of initial interest. Finally, even on the most dismissive view that the patterns exhibited here teach us nothing about people's actual concepts of explanation or understanding, our findings still reveal potential biases in how people employ these concepts. As such, they tell us something about the starting point from which we theorize, which is valuable if for no other reason than telling us what biases to watch for in ourselves.[15]

5. Alternative Hypotheses

While we have so far eschewed committing to any one account of the SEE, our reliance on it as a marker of the mental might itself seem controversial. In this section, we explore what conclusions can be drawn on the basis of other accounts of the SEE. We will maintain that, whatever one's account of the SEE, our data support psychologism about explanation. In any event, in order to be extensionally adequate, any account must accommodate our data.

First, we argue that on any account of the SEE, our results show that understanding attributions mediate explanation attributions. This follows from the mediation analyses in Experiment 2, and from the experimental manipulation of Experiment 3. The more surprising result, however, is the juxtaposition of Experiments 2 and 3. While Experiment 2 suggests that attributions of explanation are susceptible to a SEE (and therefore suggests a broad scope for SEE susceptibility), Experiment 3 shows that fixing understanding is sufficient to eliminate this effect on explanation (and this suggests a quite targeted set for the mental state inferences in operation). This has the interesting consequence of making our results of interest

[14] In addition to Waskan et al (2014), Woodward (2003, Chapter 1) advocates a similar ideal.
[15] Alexander and Weinberg (2007) make a similar point.

regardless of how broad the set of attributions ultimately found to exhibit an SEE turns out to be. The narrower the set, the more surprising the finding that explanation (and understanding) fall within it. For example, if only beliefs-conjoined-with-an-intention are subject to an SEE, then it would be very surprising that judgments of explanation are so affected. Conversely, the broader the set, the more exciting the result that Experiment 3 does not detect any SEE once we control for understanding.[16] For example, suppose that attributions of any state to any kind of a person—epistemic, mental, physical, etc.—are ultimately found to be subject to an SEE. In that case Experiment 3 not only demonstrates psychologism about explanation, but a particularly well-behaved psychologism about explanation—since understanding is sufficient to block the influence of other SEE-susceptible states. This result would go far beyond what is shown by Experiment 3 alone, or in concert with the Waskan et al. results.

Thus, while we settled on one account of the SEE to motivate our use of the SEE to investigate psychologism, one who doubts the correctness of this account can see our conclusion as disjunctive: either psychologism is true in a surprisingly interesting way, or a surprisingly understanding-specific psychologism (where understanding attributions block other mental effects with respect to explanation attributions) is true. (Given the ever-widening scope of the SEE, we strongly suspect the latter.) In either case we have made a valuable conceptual advance. Our initial conception of the SEE was thus a ladder which can be kicked away once our ultimate conclusion is reached (somewhat akin to Wittgenstein's 2013/1921 Proposition 6.54).

While other hypotheses could be proposed to explain our data, all of the alternatives of which we are aware face difficulty. For instance, one possibility is that understanding attributions depend on explanation attributions, rather than the reverse. This possibility is challenged by the mediation analyses

[16] This point about mitigating the SEE is even more poignant if the SEE is not restricted to mental states at all, as perhaps suggested by the example of 'caused' in Knobe and Fraser (2008). That being said, also note that the obvious way to connect causal judgments to mental-state judgments in terms of responsibility (see n. 7) does not apply to this case, as offering the explanation of why the decision will yield the result does not confer responsibility for that result.

from Experiment 2: the effect of norm conformity on understanding attributions was not mediated by explanation. Moreover, this possibility would require additional assumptions to make sense of the experimental results from Experiment 3: why would stipulating the presence or absence of understanding eliminate the SEE for explanation? A related possibility is that the relationship between understanding and explanation is evidential, rather than constitutive. That is, the presence or absence of understanding might affect attributions of explanation not because psychologism is true, but because the presence of understanding is good evidence that an explanation was offered, and the absence of understanding is good evidence than an explanation was withheld. This possibility fits squarely within traditional accounts of understanding, which identify it very closely with explanatory knowledge (e.g., Hempel 1965, Khalifa 2017). Moreover, this idea is consistent with the result from Experiment 3 that stipulating understanding affected attributions of explanation. However, we would again have expected attributions of explanation to mediate the effect of norm-conformity on understanding attributions in Experiment 2 (if explanations are thought to reliably affect understanding). Moreover, even granting that the presence or absence of understanding plays an evidential role in explanation attribution, the complete absence of an SEE in Experiment 3 suggests something stronger. If the evidence provided by understanding were independent of the pathway by which explanation judgments exhibit an SEE (for example, if part of the explanation SEE were the result of people attributing desires to the VP or the chairman), it would be puzzling why Experiment 3 would see the SEE vanish completely.

Another possibility is that norm conformity influenced attributions of explanation not because of its impact on mental state inferences, but because of its influence on participants' causal representations.[17] Specifically, prior work has found that people disproportionately attribute causal force to an agent who violated a norm relative to an agent who engaged in the same behavior while conforming to a norm (Knobe

[17] We are grateful to a reviewer and to Joshua Knobe (in conversation) for articulating interesting versions of this proposal.

& Fraser 2008). Perhaps in the violate case, participants were more inclined to attribute to the chairman (or the VP) causal force in producing the side effect. As a result, they could have been more compelled to judge that the side effect was well explained, and retrospectively that the vice-president offered a better explanation. This hypothesis accounts for our data without any assumption of psychologism. However, we find three main problems with this causal-attribution hypothesis. First, and most importantly, what was being explained was *why the process would cause the side effect*, not why the chairman's decision would cause the side effect, nor why the VP's recommendation would do so. It's not immediately obvious why the perceived causal role of an agent would affect the status of an explanation for why a subsequent process generated a particular effect. Research on causal chains has indeed found that in some cases, attributing greater causal responsibility to an earlier cause decreases the perceived causal role of a later cause (e.g., Murray & Lombrozo, 2017); such an effect in this case would, if anything, predict a pattern opposite to that observed. Second, this hypothesis cannot readily explain the mediation results from Experiment 2. If norm conformity affected causal attributions, which in turn affected explanatory judgments, we might have expected explanation attributions to mediate the effect of norm conformity on understanding attributions, or for the effect to be symmetrical. But this is not what we found: there was instead an asymmetrical effect whereby understanding attributions mediated the effect of norm conformity on explanation attributions. Finally, this hypothesis would face the puzzle of explaining why an explanation SEE was not found in Experiment 3. If the effect was due to causal attributions, why would stipulating understanding block this effect?

Of course, there are no doubt other candidate explanations for our data. However, at this point we take the burden to have been shifted to the anti-psychologist to produce an alternative.

Finally, our results not only speak to psychologism about explanation, but also serve as a constraint on the correct theory of the SEE. Whatever theory one has for the SEE, one must account for why it applies to understanding directly, but applies to explanation derivatively. This constraint causes a *prima facie*

difficulty to, for example, Hindricks' account, according to which what people exhibiting an SEE track is agents' insensitivity to norms. This account does not really make sense of our case, as there is no operative norm to which the *vice-president* (who is the subject of the effect) is being insensitive. As another example, the present finding is *prima facie* evidence against Machery's (2008) 'trade-off hypothesis' that people think those who accrue costs do so intentionally, as the vice-president did not incur any costs of note.

6. Relation to 'Explanatory anti-psychologism overturned by lay and scientific case classifications'

Reassuringly, our results are quite consistent with those found by Waskan et al. in their 2014 paper, 'Explanatory anti-psychologism overturned by lay and scientific case classifications.' Much of what we have done amounts to a conceptual replication of their results. Moreover, we endorse their arguments for the value of investigating laypeople's and scientists' explanatory judgments: evidence of massive discordance between common intuitions and the presumptions of philosophers speaks against the general methodology of ignoring the psychological aspects of explanation. For the very reason that we share so many similarities, however, it is worth being explicit about the ways in which our results differ from theirs.

First, and perhaps most importantly, we show that there is strong reason to believe not only that explanations are judged by the extent to which they produce some mental state (namely understanding), but that understanding is sufficient to play this mental state role. Moreover, by providing simple candidate explanations – rather than stipulating the properties of underspecified explanatory models in science – we can be more confident that it was understanding itself that played this role, and not an inference to some other feature.

A second major difference between our results and those of Waskan et al. is our attempt to isolate exactly whose understanding is relevant. We find that explanation is particular-audience- and speaker-

understanding-relative. The stimuli used by Waskan et al. varied in whether it was the explainer whose understanding was required, or the scientific community generally (though even in the former case, the text specified that the information was made publicly accessible and was easy to understand). Moreover, the relatively stronger impact of the understanding of the speaker is an unexpected finding.

A third difference is that we extend Waskan et al.'s results from scientific models to more everyday scenarios in which explanations are offered. Without this demonstration it would be risky to generalize about the nature of explanation from the scientific case alone; it would always be possible that scientific explanation is constrained by norms specific to science.

Fourth, our studies used very different methods from those of Waskan et al., and so it is revealing that they support similar conclusions. Experiment 2 used a new, indirect measure of the effect of mental state attributions on explanation, and Experiment 3 involved direct experimental manipulations of which particular individuals achieved understanding.

In sum, our findings both support and complement those of Waskan et al. using very different methods; they not only offer convergent evidence, but also help sharpen the claim that *understanding* is the critical mental state behind folk psychologism about explanation.

7. Conclusion

Across three experiments, we find support for a psychologism about explanation of a surprisingly robust variety. In so doing, we not only shed light on the folk conception of explanation and its relationship to understanding, but also uncover new data that must be accommodated by any theory that claims to make sense of the SEE.

Acknowledgements: We would like to thank the University of California, Berkeley, the University of

Pittsburgh (including the Center for Philosophy of Science and the department of History and Philosophy of Science), and grants from the John Templeton Foundation and James S. McDonnell Foundation for their generous support. We would also like to thank Joshua Knobe and James Beebe for helpful conversation.

References

Achinstein, P. (1983). *The Nature of Explanation*. Oxford University Press.

Alexander, J., & Weinberg, J. M. (2007). Analytic epistemology and experimental philosophy. *Philosophy Compass*, *2*(1), 56–80.

Alfano, M., Beebe, J., & Robinson, B. (2012). The Centrality of Belief and Reflection in Knobe-Effect Cases. *The Monist*, *95*(2), 264–289.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. Psychology Press.

Beebe, J. (2013). A Knobe Effect for Belief Ascriptions. *Review of Philosophy and Psychology*, *4*(2), 235–258.

Beebe, J. R., & Buckwalter, W. (2010). The Epistemic Side-Effect Effect. *Mind and Language*, *25*(4), 474–498.

Bromberger, Sylvain. (1966). Why Questions. In *Mind and cosmos: Essays in contemporary science and philosophy* (pp. 86–110). Pittsburgh: Pittsburgh University Press.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Dalbauer, N., & Hergovich, A. (2013). Is What is Worse More Likely?—The Probabilistic Explanation of the Epistemic Side-Effect Effect. *Review of Philosophy and Psychology*, *4*(4), 639–657.

De Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*(1), 137–170.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2399–2402). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1753688

Friedman, M. (1974). Explanation and Scientific Understanding. *Journal of Philosophy, 71*(1), 5–19.

Garfinkel, A. (1981). *Forms of explanation*. Yale University Press New Haven.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science, 15*(2), 135–175.

Kelp, C. (2015). Understanding Phenomena. *Synthese, 192*(12), 3799–3816.

Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. New York: Cambridge University Press.

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher & W. Salmon (Eds.), *Scientific Explanation* (pp. 410–505). Minneapolis: University of Minnesota Press.

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis, 63*(3), 190–194.

Knobe, J. (2007). Reason explanation in folk psychology. *Midwest Studies in Philosophy*, 31(1), 90–106.

Knobe, J., & Fraser, B. (2008). Causal Judgment and Moral Judgment: Two Experiments. In W. Sinnott-Armstrong (Ed.), *Moral Psychology*. MIT Press.

Lewis, D. (1980). Mad pain and Martian pain. *Readings in the Philosophy of Psychology, 1*, 216–222.

Lombrozo, T., & Wilkenfeld, D. (2015). Inference to the Best Explanation Versus Explaining for the Best Inference. *Science and Education, 24*(9–10), 1059–1077.

Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking About Mechanisms. *Philosophy of Science, 67*(1), 1–25.

Machery, E. (2008). The Folk Concept of Intentional Action: Philosophical and Experimental Issues. *Mind and Language, 23*(2), 165–189.

Murray, D., & Lombrozo, T. (2017). Effects of manipulation on attributions of causation, free will, and moral responsibility. *Cognitive science, 41*(2), 447-481.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Pettit, D., & Knobe, J. (2009). The Pervasive Impact of Moral Judgment. *Mind and Language*, *24*(5), 586–604.

Salmon, W. C. (1971). *Statistical Explanation & Statistical Relevance*. Pittsburgh: University of Pittsburgh Press.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton University Press.

Scriven, M. (1962). Explanations, predictions, and laws. In *Minnesota Studies in the Philosophy of Science* (Vol. 3, pp. 170–229).

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, 44(3), 510–515

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87–100.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Waskan, J., Harmon, I., Horne, Z., Spino, J., & Clevenger, J. (2014). Explanatory anti-psychologism overturned by lay and scientific case classifications. *Synthese*, 191(5), 1013-1035 5Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, *190*(6), 997–1016.

Wilkenfeld, D. A. (2014). Functional Explaining: A New Approach to the Philosophy of Explanation. *Synthese*, *191*(14), 3367–3391.

Wilkenfeld, D. A., Plunkett, D., & Lombrozo, T. (2016). Depth and Deference: When and Why We Attribute Understanding. *Philosophical Studies*, *173*(2), 373–393.

Wittgenstein, L. (2013/1921). *Tractatus logico-philosophicus*. Routledge.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Ylikoski, P. (2009). The Illusion of Depth of Understanding in Science. In H. D. Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press.