# What is a Philosophical effect?
# Models of data in experimental philosophy

In 1989 Pons and Fleischmann reported an anomalous increase in heat and the presence of neutron radiation in their small calorimeter. They claimed to have discovered 'cold fusion'. But their claim was unsupported. To treat the increase in heat as anomalous, they needed to assume that the efficiency of electrolysis was near 100% in the calorimeter, and that there were no relevant differences in heat loss between experimental and calibration conditions. But their data were consistent with reactions that would violate each of these assumptions, and they could not rule out either of these possibilities (cf., Shanahan 2002). Furthermore, while their understanding of cold fusion made predictions about the rate of neutron production, it made no direct predictions about the meter readings on the neutron detector, and there were many ways in the relevant meter readings could be produced in the absence of an increase in neutron radiation. In short, Pons and Fleischmann did not have a good model of how their calorimeter or neutron detector worked, and they didn't have a good theory of the data they were collecting. Consequently, their exciting report told us little about the world.

We know that Pons and Fleischmann were *wrong* about the existence of cold fusion. But from a philosophical perspective, the problems with their inference run much deeper. Even if they had discovered cold fusion, they would have been in no position to justify the claim that they had. They could neither articulate nor defend a theory that would explain why their meter readings were authoritative with respect to the existence of cold fusion. And when they were pushed, they could not explain why their data provided confirmation for a hypothesis about the existence of cold fusion. They could not rule out other interpretations of their data. So their claims, even if they had been true, would have been unjustified.

My primary aim in this paper is to show that research carried on under the banner of experimental philosophy often founders on a similar type of worry. My argument builds on insights first advanced by Patrick Suppes (1962), though it does not depend on the precise details of his claims about models of data. Like Suppes, I hold that theories should not simply be seen as sets of sentences, which yield concrete predictions that can be compared directly to the world using point-measurements or point-observations. Theories predict global phenomena, and this means that they can only be evaluated by reference to 'canonical data' that have been transformed, corrected, and analyzed in accordance with a hierarchy of models that justifies their interpretation (e.g., models of theories, models of experimental design, and models of data). Suppes argued that any attempt to justify the inference from data to hypothesis requires developing, articulating, and defending plausible models of the data that are collected, the scales that are used, and the statistical analyses that are carried out. I maintain that such models are necessary to justify our theoretical claims, and without such models, it is impossible to rule out a variety of potential distorting factors that may be buried deep in an analysis. This is where experimental philosophers run into trouble.

The models that are necessary to demonstrate that behavioral data reveal a *philosophically meaningful effect* are rarely present in experimental philosophy. Consequently, many experimental philosophers draw illicit inferences from the statistically significant data they have collected. In a sense, this is a familiar charge, but I hope to develop it in a way that points toward better explanatory practices. I begin by outlining a difficulty that arises in interpreting any Likert data, and I show that one way to obviate this worry is to articulate and defend a plausible theory of what a scale is measuring (unfortunately, I also argue that developing such a theory is likely to be a difficult task). I then examine three case studies, which reveal different ways in which the *evidential relationship* between data and hypothesis can be undermined, depending on the kind of analyses that are carried out. While none of these difficulties is insuperable, overcoming them requires a great deal of conceptual and theoretical work, and to the best of my knowledge, this work has not yet been carried out. Finally, I close by addressing two plausible responses to my claim that experimental philosophers must articulate and defend plausible models of their data, scales, and analyses.

**1. Measuring judgments**:

Experimental philosophers often ask participants to read thought-experimental probes, and to respond to them by marking numbers along Likert scales. This is not the only method they employ, but it is a familiar and common method.[1] To derive a philosophically meaningful conclusion from such data, experimental philosophers must assume—if only tacitly—that there is a high-degree of correlation between the responses they collect and the philosophically meaningful judgments that they are attempting to uncover. But it is hard to see why we should be willing to assume that this is true.

To begin with, I assume that most people do not think in Likert scales.[2] So such tasks require converting thoughts, which may take the form of diffuse *inklings,* into a novel representational format. Specifically, participants must find some way to treat their thoughts as determinate and digital representations, which can be recorded as numerical values. Most people succeed in offering responses in these experimental tasks; but it is less clear if they thereby succeed in offering an accurate representation of their mental state, or of a disposition that will extend beyond the context of the current experiment. After all, such conversions are not simple translations, and it is unlikely that they are lossless in terms of the information they carry (cf., Haugeland 1991; Cummins, Roth, & Harman 2013). To begin with, converting a thought into a discrete representation on a Likert-scale makes it impossible to record the kinds of yes-buts, well-

[1] Some studies examine reaction times in categorization tasks, acceptability judgments, or use paraphrase tasks to uncover tacit or implicit knowledge. Unique problems arise in evaluating each type of data, and the nature of those problems is unique to each methodology. I hope my worries are clear enough that they can be generalized to these other cases where doing so is appropriate.

[2] Thanks are due to an anonymous referee at Philosophical Studies, who asked me to clarify the arguments in the following two paragraphs.

maybes, and almost-but-not-quites that typically arise in thinking about a new, philosophically interesting phenomena. Yet participants must make a decision—deliberatively or reflexively—about how they are going to record their thoughts as numerical values on a clearly specified scale. Such decisions are unlikely to be constrained by either innate strategies or learned rules. And it is not at all obvious that we should expect everyone to carry out these conversions in precisely the same way. They might. But providing evidence that they do would be a herculean effort, and I am not quite sure how it could be achieved.

Furthermore, the lack of innate strategies and learned rules for responding to Likert-scales makes it unclear what it would mean to be 'successful' in converting a diffuse thought into a discrete representation on a Likert-scale. If there were biologically or socially entrenched *norms* for carrying out these conversions, we would be able to explain where someone had made a mistake, or where they had done a better job of recording their thoughts. But there is no obvious way to determine where people are more or less successful in carrying out these conversions, as there is no obvious standard against which to evaluate these conversions. And finally, since participants in these experiments are likely to have diverse backgrounds and learning histories, there will often be far more variation in their thoughts about a philosophical situation than can be adequately captured by the limited number of options that are available using a Likert-scale. Consequently, even if every participant were to follow the same rule or strategy for carrying out these conversions, we would still need some reason to believe that the differences that are collapsed on the Likert-scale are not philosophically interesting differences.

Consider a similar argument advanced by James Bogen and James Woodward (1998). They begin by distinguishing data from phenomena, and they argue that phenomena must be inferred or estimated from patterns that arise in observable data. For example, consider the claim that lead melts at 327.5° ± 0.1° C (Bogen & Woodward 1998, 308ff). This claim is supported by numerous data points, observed over the course of multiple readings taken from a thermometer. To measure the melting point, observers must take temperature readings 'exactly' as a sample begins to melt; but minor variations in observations and variations in the mechanisms used to collect such observations yield a distribution of different responses (indeed, it is conceivable that the reading "327.5°" never occurs in the data). Standard assumptions about the character of these variations, as well as their magnitude, make it seem reasonable that the data will be normally distributed, and that the mean response and the associated standard error provides an adequate estimate of the temperature at which lead melts. In fact, these data do provide a good measure of probability that an interval, centered on the mean, contains the actual temperature at which lead melts (Bogen & Woodward 1998, 309). But the point on a thermometer at which any particular sample of lead will melt "depends not only on the melting point of lead, but also on its purity, on the workings of the thermometer and that of the sample, and a variety of other background conditions" (Bogen & Woodward 1998, 309).

Things become more complicated in the case of Likert data. In this case, we don't have a simple substance like lead, but a person who can judge, revise, and inhibit their immediate responses. The problem that I intend to call attention to turns on the fact that experimental philosophers do not tend to have a clear enough idea regarding what a philosophical effect would be, what that some pattern in their data reveal the presence of phenomena that some theory or other might predict or explain. We must always ask: "Why should we think that Likert data are going to reveal philosophically interesting phenomena?" When someone examines a statistical mean, or looks as standard deviations, they are modeling their data to make them intelligible; but they are also making assumptions about the distribution of data, and the ways in which that distribution relates to the underlying phenomena. The problem with Likert data, then, is that there is a fundamental measurement issue, and the only way to avoid having this issue cause serious problems is to have an account of why the facts about the world that we are trying to understand can be measured using an ordinal scale.

More traditionally minded philosophers have sometimes voiced similar concerns in making their arguments against the explanatory relevance of experimental philosophy. But things are not nearly as dire as these philosophers seem to suppose. There may be cases where it doesn't matter whether there is a bit of variation in the judgments people offer, at least so long as they agree with one another in broad brush strokes. Indeed, whether the differences that might emerge in such judgments matters at all will depend on what sort of question is being addressed. More specifically, it will depend on how strict the standards are for confirming or rejecting a hypothesis in experimental philosophy. Some hypotheses demand a higher degree of support than others. As Richard Rudner (1953, 2) famously notes, we ought to be more concerned about the presence of false negatives when we are considering the possibility that a lethal toxin is present in a batch of a drug, than when we are considering the possibility that a batch of belt buckles is defective. Mistakes in philosophy tend to be less serious than medical mistakes. But it is not clear (at least to me) how sensitive measures of philosophical intuition must be if they are to provide evidence in favor of a philosophical claim. Regardless of how we answer these questions, however, it should be clear that by carrying out their investigations experimental philosophers incur a burden to take a clear and decisive stand on how their data relate to the hypotheses under consideration. This is one of the primary reasons why the responses people offer on Likert scales should not be treated as point-measurements that can provide direct confirmation of a philosophical hypothesis. That said, it is important to remember that such data may provide confirmation for a hypothesis with a bit of further argumentation.

The sort of worry that I have just developed should help to make it clear why experimental philosophers need to articulate and defended a plausible *theory of their scales.* But it might not be clear yet just how this should be done. In general, experimental philosophers simply assume that '2's and '3's—as well as '6's and '7's—should be seen as different responses, as opposed to different performances of the same response. But just looking at the data that have been

collected provides no warrant for this assumption. There is nothing to rule out the possibility that one person's '2' is indistinguishable in every philosophically and psychologically significant respect from another person's '3'. Establishing that they are different responses will always requires a further argument. Likewise, it is not obvious why we should assume that every participant who choses the mid-point of a scale is providing an answer, as opposed to implicitly refusing to answer, refusing to answer but trying to please the experimenter, or simply expressing puzzlement at the anchors of the scale. Here too, establishing that the data should be interpreted in one of these ways requires an argument.

In light of these worries, I contend that the evidential support for the hypotheses advanced by experimental philosophers can be clarified by taking a stand on whether, and to what extent, it matters if people differ in their subjective thresholds for offering a particular response. As a first cut, we should note that how precise the answers to these questions must be depends on the *size* and *shape* of the effect that would have to be revealed in order to confirm a hypothesis. Some types of experiments should be expected to reveal relatively large differences between groups and small differences within groups. These experiments allow for more noise in the data, and more variation between participants, than experiments where we should expect to find relatively small differences between groups with large standard deviations. Put differently, where there are large between group differences, the scale can be a bit less sensitive, and it can allow for more variation within a group. But where the differences between groups are expected to be smaller, a more sensitive measure will be necessary. Unfortunately, most of the data collected in experimental philosophy to date suggest that philosophically interesting hypotheses tend to fall into the latter category; in general, these sorts of studies tend to yield small between-group differences, along with a great deal of within-group variance as revealed by wide standard deviations and substantial minority populations. Thus, drawing a philosophically interesting conclusion from these data requires demonstrating that the scale in an accurate mapping of the intuitions that people have. To the best of my knowledge, this is not an issue that experimental philosophers have tended to address—and it is an issue that is likely to undermine the justification for many of the claims that are made by experimental philosophers.

Alas, even once experimental philosophers address these sorts of problems, they will not be out of the woods. There are other sorts of difficulties that arise in making a choice about which statistical analysis should be used in investigating the patterns that are present in Likert data. Experimental philosophers tend to make predictions about *where* they will find statistically significant effects when they run a t-test or an ANOVA. But they rarely defend their assumption that this is the right analysis to carry out. These analyses are used to examine the difference between the mean responses of two groups. But such differences are not always the most relevant difference, relative to a philosophical hypothesis. Sometimes, what matters are shifts the in the proportion of people who accept a particular response; sometimes, what matters is the way that responses tend to 'clump' together. While survey responses are sometimes distributed normally,

they are often skewed toward one response, or distributed bi-modally, or spread out across an entire scale. Some philosophical hypotheses predict convergence on a particular judgment, while others predict a dimorphism in a population, and still others predict uncertainty about what the right response is. Each type of prediction yields a different set of assumptions about the shape and size of an effect that should be observed in a population. This is why experimental philosophers need to clarify their predictions about how their data must be shaped to confirm a hypothesis. By clarifying their commitments on these issues, experimental philosophers will find themselves in a better position to provide arguments demonstrating that there is an interesting evidential relationship between the data they collect and the hypotheses they are investigating.

Finally, and relatedly, it is important to remember that the presence of a stable minority, or the existence of a bi-modal distribution, can be obscured by analyses that target differences in mean-responses and standard deviations. In many cases, the analyses employed by experimental philosophers make it difficult to see whether an effect is driven by heavily skewed responses in one group, or whether it reveals a shift from a bi-modal to a normally distributed pattern of responses, or whether it reveals a shift from a 'highly peaked' to a flattened distribution of responses. Again, it is no easy task to determine whether these issues matter, relative to a particular hypothesis. And there is no mechanical procedure for determining how a dataset would have to be shaped if it were to confirm a philosophical hypothesis. But without a clear *theory of the data* that addresses these issues, the fact that some analysis has revealed the presence of a statistically significant result cannot *on its own* support any philosophically interesting conclusion. To see why, it will help to examine three case studies more closely, and to see how kinds of different analyses, based on different kinds of assumptions, can founder on different kinds of conceptual and theoretical difficulties, as well as what might be done to address these difficulties.

**2. Measuring judgments (redux)**:

I have selected three case studies to highlight three distinct types of problems. Each study is drawn from a different domain that is commonly targeted by experimental philosophers (consciousness; free will; moral cognition). But more importantly, each study employs a different strategy for measuring the presence of significant effects in Likert data: (§*2.1*) Wesley Buckwalter and Mark Phelan (2013) use an ANOVA to measure difference in means between different groups; (§*2.2*) Eric Schulz and his colleagues (2011) use a regression analysis to extract a $R^2$ value; and (§2.3) Elinor Amit and Joshua Greene (2012) measure the correlation between two variables. Each of these cases illustrates, in a concrete way, a different kind of issues that can arise in the analysis of Likert data. Specifically, they reveal: (§2.1) the problems caused by high within-condition variance; (§2.2) the problems caused by small effect sizes; and (§2.3) the ways

that transforming data can make them more 'well-behaved', while also making them more difficult to interpret.[3]

*2.1 Measuring differences in means*

Many studies in experimental philosophy investigate the impact of a simple change in a situation or a scenario on explicit judgments, so it makes sense to begin here. In a recent experiment that typifies this approach, Buckwalter and Phelan (2013) asked participants to read a short narrative about a robot, and to respond to questions about its ability to smell a particular object—like many experimental philosophers, they used Likert-scales anchored at (1) Clearly no, (4) Not sure, and (7) Clearly yes. Their primary hypothesis was that variation in function—but not variation in complexity—would significantly impact judgments about a robot's ability to smell an object; Buckwalter and Phelan also predicted that this effect would not be moderated by the valence of an object's smell (bananas; a chemical; vomit). Using an analysis of variance, they found that a robot's specified function had a significant effect on people's responses, while an object's valence did not significantly affect these judgments; they also found that participants were more likely to say that a simple robot designed to handle biowaste could smell vomit than to say that a simple robot designed to make smoothies could.[4] These results seem to confirm their hypotheses, but a closer look at their data suggests that things were more complicated (see Buckwalter & Phelan 2013, *Online supplementary material*).

Variation in the function of complex robots never yielded a difference larger than a quarter-point on the designated 7-point scale. For such a small difference between two groups to be theoretically meaningful, two conditions would have to be met: first, the scale would have to be incredibly sensitive to differences between populations; second, there would have to be good reason to suppose that every participant interpreted the scale in *precisely* the same way. As I noted above, such claims are difficult to defend, and Buckwalter and Phelan wisely do not attempt to defend them. But I would suggest that the rather large standard deviations (range = 1.25-2 points), relative to the differences between the two groups, reveal that there was a high degree of variation *within* groups—though they do not establish what kind of variation this is. These standard deviations also suggests that the measure was not particularly sensitive, and that at least

[3] Thanks are due to J. Brendan Richie for help with the framing of these issues.

[4] *Function, Object*: $F(2, 241)=5.02$, $p<.01$; there was no significant correlation between affective valence and the ability to smell an object, $r(251)=0.066$, $p=.30$. Biowaste vs Smoothie, $U(40) = 94.0$, $Z=3.342$, $p<.001$. They also found a significant interaction between function, complexity, and object, which they take to be irrelevant to the hypotheses under consideration: *Function*, *Complexity*, *Object*, $F(2, 241)=4.67$, $p<.01$. Buckwalter and Phelan do not explain why they ignore this interaction, but it is worth noting that ANOVAs can reveal significant main-effects that are meaningful as higher-level expressions of the interaction between multiple variables; where this happens, interpretations based solely on the main-effect are likely to be false or misleading. Buckwalter and Phelan should have investigated the structure of thee three-way interaction (using post-hoc tests and adjusting their significance-level to avoid false-positives).

*some people's* responses fell far away from the mean response—though these data do not reveal how the data were shaped.

To make their case, Buckwalter and Phelan would need to develop a sophisticated account of why the difference that is revealed in this task is relevant to their target hypotheses. It is unclear precisely what sort of account this would be, but it would need to make it clear why a small variation within a population would have an important effect on judgments about mindedness or consciousness more broadly. Without such an account, the data show that there was a significant difference between *responses* that were provided by the target populations, but more needs to be done to show that this particular difference is sufficient to establish a difference in people's *thoughts* about robots. But perhaps they have more recourse for supporting their claims.

Turning to the case of simple robots, variation in function had a much larger impact on people's judgments (vomit = Δ1.81 points; chemical = Δ1.02 points). Here too, there were large standard deviations that make it difficult to interpret the data. This difference may reveal that changes in function shifted most people's responses in the predicted direction; it may reveal that there was a shift in the proportion of participants who offered a particular response; or it may reveal the presence of a stable minority in one or both cases. The data presented by Buckwalter and Phelan cannot, on their own, discriminate between these interesting hypotheses about the cause of the difference in responses. But importantly, the data that they collected could do so. It would be a trivial task to carry out the analyses that show how these participants were affected by the experimental treatment. And with these analyses in hand, they could attempt to develop an account of how these data are relevant to claims about people's thoughts about robots (though they would still need to provide a clear sense of which hypothesis is relevant to the philosophically meaningful issues they are interested in). But unfortunately, there would still be deep problems with the interpretation of these data, at least relative to the hypothesis that is under consideration in the paper.

Buckwalter and Phelan report a three-way interaction between a robot's Function and Complexity, and the Object it must smell.[5] In light of the descriptive data I have just addressed, this interaction should seem unsurprising. Function seems to have played a relatively substantial role in judgments about simple robots, but not in judgments about complex robots. So these data may provide interesting support for the hypothesis that function only matters for simple robots, because functional considerations act as a stand-in for tacit views about a robot's complexity (Rosenthal 2011). Buckwalter and Phelan contend that this higher-level interaction is irrelevant to the evaluation of their hypotheses—but they never defend this claim, and it is unclear why they make it given the structure of their data. This is troubling given that they never discuss the extent to which their data may provide support for a hypothesis about complexity that they intended to rule out. Of course, three-way interactions are difficult to interpret, but addressing

[5] *Function*, *Complexity*, *Object*, $F(2, 241)=4.67$, $p<.01$.

these data would require delving into conceptual and empirical issues that would pay big dividends in the attempt to link data and hypothesis. I cannot address these issues in full, but here is the upshot.

The primary hypothesis that Buckwalter and Phelan set out to examine was that variation in function—but not variation in complexity—would gave a significant impact on people's judgments about a robot's ability to smell. To determine whether their data *support* that hypothesis (or any hypothesis at all), they would need to answer several interrelated questions: How much within-group variation should be expected in a task like this, and how much is allowed for by the hypothesis under consideration? How large must the difference between two groups be to establish this type of hypothesis? And, does the hypothesis require a shift in the proportion of participants who rely on functionalist considerations, or does it instead require a shift in the overall tendency to rely on functionalist considerations? I can imagine numerous different answers to these questions, but each would require taking a stand on important theoretical issues about how to interpret data once they are collected. Buckwalter and Phelan do not address these questions, but few experimental philosophers have taken the time to address these kinds of considerations. This makes it hard to justify any philosophically meaningful inferences about the data.

### *2.2 Regression analyses*

I contend that similar worries affect many studies in experimental philosophy. But they are not always easy to see, as they often emerge in subtle ways. Recall the concern I advanced in Section 1: attempt to derive philosophically interesting conclusions from experimental data confront the fact that the theoretical notions at play in thinking about philosophical cases do not have directly observable analogs in the experimental data (cf., Suppes 1962, 253). There is no doubt that differences and similarities often show up in datasets. But what are we to make of them? My claim is that it is a difficult task to show that these differences track philosophically relevant differences and similarities in thought. The responses to experimental probes require participants to convert their thoughts into discrete values, and drawing conclusions from them thus requires explaining why the statistical tests that are employed are able to uncover *relevant* differences in a way that distinguishes them from *irrelevant* differences. This requires providing an account of what the values on a scale mean, as well as an account of how much and what kinds of variation is necessary to reveal a difference in judgments. My aim in this section is to show that more sophisticated methods of statistical analysis, which are used to uncover the cause of variations in responses face similar sorts of worried.

In making this case, it will help to turn to another experiment, which targeted the impact of personality and philosophical expertise on judgments about free will. Schulz and his colleagues (2011) collected data about personality traits using a well-established measure; they also developed a measure to collect information about participants' knowledge of philosophical debates about free will; and they

collected responses to a series of questions about the compatibility of neural-determinism, freedom, and responsibility, using a Likert scale anchored at (1) absolutely disagree and (7) absolutely agree. The result of a linear regression on these factors revealed that one aspect of extroversion ('warmth') predicted compatibilist intuitions, and that performance on their philosophical expertise task predicted incompatibilist intuitions. These are interesting data, but Schulz and his colleagues face a variety of unacknowledged difficulties in drawing inferences from these results.

When everything goes well, a linear regression will provide a measure of the percentage of variance that is explained by the independent variables.[6] Schulz and his colleagues were most interested in whether philosophical expertise would diminish the extraneous impact of personality traits on people's judgments about free will; and they report that there were no significant differences in the effect of warmth as a result of philosophical expertise. Specifically, they found that 5% of the variance in the sample could be accounted for by 'warmth', while an additional 9% of the variance in the sample could be accounted for by philosophical expertise.[7] But this leaves 86% of the variance in the model to be explained by other factors, including random variation and individual differences in the interpretation of scales, as well as other unknown types of differences within the population. This is not to deny that there is a significant linear relationship between the dependent and independent variables. These data reveal such an effect, but they also show that the independent variables were *relatively poor* predictors of the responses that people gave. But is this fact relevant to the hypothesis under consideration?

Of course, even a small effect can be theoretically meaningful, provided there is reason to expect a small effect, and provided there is reason to believe that such an effect is important *relative to the hypothesis under consideration*. To show that this effect is meaningful, Schulz and his colleagues would need to explain why minor variations in compatibilist intuitions were theoretically interesting and relevant to the philosophical issues at hand. They do not defend such claims, and I am not sure what it would take to demonstrate that such a small difference was philosophically interesting. Thus, it remains unclear whether Schulz and his colleagues uncovered a theoretically interesting relationship between personality traits, philosophical training, and philosophical judgments.

It is worth noting, moreover, that the difficulties in this case are exacerbated by the fact that some extroverted people may have lower thresholds for offering

---

[6] A linear regression provides a conditional probability distribution for values of a dependent variable, given the relevant independent variables; a regression line expresses the predicted values of a dependent variable, given these independent variables. Since the world is a messy place, data are rarely fully predicted by the independent variables. The deviation of a response from the predicted value is called the 'residual', and $R^2$ is calculated by subtracting the residual variability of a model from 1 to yield a measure of the variance that is explained by each independent variable. Ideally, $R^2$ measures the correlation between the value of an independent variable and the value of a dependent variable. Where they are perfectly correlated, $R^2 = 1$; where there is no relationship between them, $R^2 = 0$.

[7] Warmth ($R^2$=.05, F=6.6, p=.011); philosophical expertise ($R^2$=.14, F=12.4, p=.001).

each response on the scale, and similarly by the fact that some philosophers may have been more reticent about offering responses that skewed toward the 'absolutely' end of the scale, even if they agreed to some extent with the claim that they were presented with. People who were 'warmer' may have more compatibilist intuitions, and philosophers may have been more likely to be incompatibilists. But establishing that this was the case would require providing a further justification for the claim that the effects revealed in this study were not simply the result of statistical outliers, scale-driven oddities, or other types of abnormalities in the dataset. Schulz and his colleagues do not rule out the possibility that they solicited judgments from a few extroverted compatibilists and a few philosophically trained incompatibilists. This possibility could be examined using a scatterplot and a trend-line. But the dataset is likely to be quite noisy, and there are likely to be many data points that fall very far away from the trend-line, given the enormous amount of variance in the model that was not accounted for by the target variables—remember, 85% of the variance arises as a result of unknown factors and unexplained variability in responses. So, while this analysis does reveal a significant effect in the dataset, the presence of this effect can tell us little about how the variables under consideration are related to one another. Without a clear and plausible model of the scale that was used, and without a clear and plausible model of the analyses that were employed, it is hard to tell whether these data support any philosophically meaningful claim whatsoever.

*2.3 Transformations and Correlations*

At this point, it would be easy to assume that the worries I have been discussing are the result of inadequate training in statistical methodology. But I think nothing could be further from the truth. The problem concerns deep and difficult issues regarding the relationship between data and hypotheses. To make this point clear, I want to show that similar worries arise even in the most carefully conducted, sophisticated, and cautiously analyzed psychological studies. Consider a recent study examining the role of working memory style in the production of moral judgments. Amit and Greene (2012) used a well-established memory task to determine whether their participants relied more heavily on visual or verbal working memory; they then asked their participants to respond to a series of high-conflict moral dilemmas, using 7-point scales anchored at (1) completely not appropriate and (7) completely appropriate. Using a correlational analysis, they found that people who relied more heavily on visual working memory were, on average, less likely to provide 'utilitarian judgments' about high-conflict moral dilemmas.[8]

There are a couple of things that are worth noting about their analysis. First, they note that their data was highly skewed. So, before they carried out their

[8] They found a moderate negative relationship between visual cognitive style and utilitarian judgment, $r(49)=-.37$, $p=.007$. This effect was stable across plausible types of demographic variation (education; politics; gender; religion).

analysis, they used a logarithmic transformation to normalize the data. This is important because parametric analyses assume that the data under consideration are normally distributed (or close to it). In situations where data are highly skewed, log transformations are often used to rescale the data; this yields a more normally distributed dataset, while preserving the mathematically relevant properties of the original data. This makes it easier to detect real differences that would be hard to see in data that are clumped at one end of a scale. Amit and Greene were interested in the correlation between moral judgments and memory style, and this transformation allowed them to examine the presence of this correlation in their dataset. Second, they offer a scatterplot that reveals fairly noisy data, with some people taking a 'deontological' stance independently of their cognitive style. Nonetheless, it is clear from the trend line running through this dataset that there was a linear relation between memory style and moral judgments. Put differently, the scatter plot reveals that Amit and Greene have found solid evidence that a more visual cognitive style is a *moderate positive predictor* of the likelihood of offering 'more deontological' responses to moral dilemmas (see Amit & Greene 2012, Figure 2). So far, so good. But as philosophers, we should now want to ask: What follows from this result?

As I argued above, it is difficult to infer philosophical relevance from statistical significance when we must work out how a 7-point scale relates to more diffuse philosophical judgments. But these sorts of worries are exacerbated when we must interpret a log-transformed scale. An experimental philosopher may be able to explain *why* the difference between a mean response of '6' and a mean response of '4' on a 10-point scale is theoretically meaningful, at least where they have some reason to believe that everyone interprets the scale in the same way, and where there are clear anchors at '1', '5' and '10'. But it is much less clear how we should interpret a difference between 0.6 to 0.4 on a log-transformed scale. To interpret these data, we would first need to know where the they are centered, and we would then need to produce and justify a translation schema that could explains what this difference amounts to in light of the original scale. Let me be clear about my worry, here. I have absolutely no doubt that there is a real, and psychologically tractable trend in the judgments that people offer; and this trend is likely to track working memory style in precisely the way that Amit and Greene suggest. But without a theory of the scale, and a theory of the analyses and transformations that have been carried out, these data cannot tell us whether people with a more visual cognitive style are 'more deontological' in any sense that will be of interest to even empirically minded philosophers. Since these values appear to clump around the center of this log-transformed scale, an argument must be given to explain why a difference of 0.2 on *this scale* is philosophically meaningful, relative to a claim about the moral judgments that people make—and this is no mean task. A plausible argument for this claim will have to explain the relation between the log-transformed scale and the 7-point scale, and then explain why the corresponding difference on the original scale is large enough to warrant drawing an interesting philosophical conclusion about the judgments that people make independently of the scale on which they are

recording their judgments. Again, I'm at a loss about how to proceed, so where does this leave us?

**3. Where things stand:**

I began by suggesting that Pons and Fleishmann could not provide support for their hypothesis because they lacked a model of their instrument, and because they smuggled illicit assumptions into their analyses. Over the course of this paper, I have endeavored to reveal that similar worries threaten to block the inference from statistical significance to philosophical relevance in experimental philosophy. Using surveys to support philosophically meaningful conclusions is a difficult business. It requires developing, articulating, and defending a hierarchy of models, that includes:

1. **A model of the philosophical theory** under consideration, which includes and an account of how its core commitments hang together, and how they can be investigated empirically;

2. **Models of the experiment**, which explain what kinds of statistical regularities would need to be present in order to establish the existence of meaningful differences between populations, an account of how big those differences would have to be in order for them to be philosophically meaningful, and an account of how the dataset must be shaped to reveal something interesting about the philosophical theory;

3. **Models of data**, which explain what the relationship is between the responses that participants offer and their philosophically meaningful judgment, including an account of what the scale is measuring, and why it is sensitive enough (and why it does not over generalize); and

4. **Models of study design and data collection,** which are sensitive to considerations about the homogeneity of populations, the assignment of participants to various treatments, and the fit of various experimental parameters with the philosophical theory that is being addressed.

Where each of these models is clear and plausible, it will be possible to explain why the patterns that emerge in a dataset actually support a particular hypothesis; and where they are deficient, this fact can help to explain *why* the data cannot support the hypothesis under consideration.

Importantly, I am not claiming that experimental philosophers need to provide a theory of the instruments they use to collect data, and the statistical methods they employ, so long as they work reliably in the collection of data (cf., Bogen & Woodward 1998). There is reason to believe that the data collected using these methods are reproducible, and that they do tell us something important about how people tend to respond in the context of these experiments. The problem

arises in attempts to show that these data reveal a philosophically meaningful *effect*. Here, it must be shown that the instruments and analyses that are being used are good at collecting the data that we want to collect; and this requires explaining why the data provide a window onto the way in that people are likely to behave outside of these contexts (whether in their attempts at philosophizing, or in their interactions with the world more broadly). To address these issues, we need to ask *why* the data support a particular hypothesis.

Unfortunately, experimental philosophers do not typically present arguments to explain what their scales measure, nor do they typically address the most difficult kinds of questions about why the statistically significant effects they uncover are theoretically meaningful. Put bluntly, their analyses frequently rely on undefended assumptions about the meaning of the numerical values that show up on their scales, as well as undefended assumptions about the meaning of the differences between these numbers; they also tend to assume, without argument, that any significant difference in their dataset is meaningful, regardless of how noisy the dataset is, regardless of how individual judgments are related to the mean response, and regardless of how heavily skewed responses are. But they can do better. They can provide more explicit accounts of the tools they have employed in attempting to confirm or reject philosophically interesting hypotheses. But constructing such models is likely to be complex task.

Sometimes the shape and distribution of data matter to the evaluation of a particular hypothesis; and whether data are skewed or highly peaked may be relevant to answering some philosophical questions. Sometimes data must be transformed before they can be analyzed, but whether they do always requires addressing hard questions about whether the variance between populations is *homogenous enough* to use parametric analyses—especially given that nonparametric statistics are less reliable with small samples; where data must be transformed, this introduces a variety of further complications into the interpretation of the resulting data—as we saw in the case of the analyses carried out by Amit and Greene. So, experimental philosophers are in a tough position, to say the least. They need to develop and advance plausible models of the scales they use and their relation to the transformations they employ, and they need to explain why their analyses and their interpretations of results are justified by a hierarchy of models that mediates between statistical significance and philosophical relevance. They also need to explain why differences in responses are philosophically meaningful, and they need to make it clear where and when differences in the distribution of responses can safely be ignored. Until experimental philosophers take up the defense of such models, they risk being insensitive to the wide variety of ways in which statistically significant differences can mask factors that militate against treating them as evidence of meaningful differences in judgments or populations. The hierarchies of models lying between raw observations and philosophical hypotheses will be difficult to construct, and there are few of them already on offer for the types of studies that experimental philosophers tend to use. But I maintain that such models are nonetheless necessary.

## 4. Two worries

I wish that I could stop here. But there two interesting worries that are commonly expressed when I present these sorts of arguments.[9] According to the first, I have proved too much, as my arguments can be applied to most of the work carried on in cognitive and social psychology. Given the undeniable success of this ongoing research, it may seem that my arguments are misguided. According to a second closely related worry, my arguments do not, and indeed cannot establish that experimental philosophers are likely to make false claims. To the extent that they employ methodologies that have been successful in cognitive and social psychology, experimental philosophers should tend to get the right answers—and this should be true even if they cannot articulate or defend plausible models of their data. In this concluding section, I address both of these worries to clarify the force of my argument.

To begin with, these issues are relatively standard issues that apply to any use of statistical analyses in empirical research; but they are especially important in the case of social psychology, where Likert scales are used in an attempt to uncover an effect. That said, I have no doubt that many cognitive and social psychologists have carried out important research without being able to articulate or defend plausible models of their data. Yet, over the past couple of years, this has started to change. There have been numerous failed attempts to replicate psychological results, and many people have acknowledged that they have file-drawers filled with null-results and failed replications. There is a rapidly growing sense that many exciting and noteworthy results may have emerged as a result of cherry-picking data or p-value hacking (Nuzzo 2014). But far more interestingly, social and cognitive psychologists have started to acknowledge that the collection, analysis, and interpretation of psychological data should be held to a higher standard. Put simply, cognitive and social psychologists recognize that the presence of a statistically significant results does not, on its own, confirm a psychologically interesting hypothesis. And they recognize that they need to do something to make their investigations more reliable, more reproducible, and more informative!

In light of these recognitions, a number of plausible interventions have been suggested; these include reporting effect sizes and confidence intervals, using Bayesian statistics, and carrying out multiple different kinds of analyses on a dataset (Nuzzo 2014). However, some cognitive and social psychologists have gone even further, suggesting that study pre-registration should be used to ameliorate these problems (Chambers et al 2013). They suggest a two-step strategy for publishing a paper, which runs roughly as follows: scientists should first submit their introduction and methods section, specifying strategies to deal with statistical outliers and data that are not normally distributed; they should then collect their data, and journals should agree to publish papers that have

[9] Thanks to Rik Hine and to an anonymous referee for pushing me to clarify both of these points.

articulated interesting tests of interesting hypotheses, regardless of whether the experiment yields statistically significant results, marginally significant results, or even statistically insignificant results. Intriguingly, Chris Chambers (2012) has even suggested that psychologists and cognitive scientists should make their raw data publically available, and that they should also standardize their statistical analyses to prevent the intrusion of illicit factors into their analyses. While these solutions are institutional in nature, and the recommendations I have offered are more directly focused on individual practices, I see both viewpoints as deriving from a common source: the recognition that we need better models of data and analyses.

This recognition that the confirmation of a psychological theory is not so straightforward has been a long time in coming; however, it is rapidly becoming clear that psychological theories predict global phenomena, and that they must be evaluated against the backdrop of a hierarchy of models that can justify their interpretation (Suppes 1962). For the cognitive or social psychologist, this amounts to a reconceptualization of the models of data and statistical analyses that are necessary to find a result. Where they have pre-registered their methods, and clearly articulated strategies for dealing with statistical outliers and other abnormalities in their data, psychologists will be able to defend a plausible model of their data if they are ever asked to do so. But more importantly, they will be on a much firmer foundation in making claims about effects that are real, robust, and interesting—while also making it clear which hypotheses are not supported by the existing data.

I have argued that experimental philosophers often find themselves in the same boat as cognitive and social psychologists. By borrowing scientific methods, they also inherit the problems with those methods. But this does not capture the problem I have raised in its full generality. Put simply, philosophy takes place at a higher level of abstraction than most forms of psychological investigation. And effects can be psychologically real, but largely irrelevant to concerns of philosophers. So, even if the problems with existing statistical methodologies are rectified in an institutional fashion, experimental philosophers will still have work to do. As philosophers, they will need to provide the intervening models that can justify their claim that a psychologically real effect is philosophically meaningful. And this is not an issue that can be addressed from outside of an experimental project. This is why experimental philosophers need to articulate and defends a theory of how the responses they collect relate to philosophically meaningful judgments; and this is why they need to be able to explain what kinds of statistical results are sufficient to establish a philosophically relevant difference between two populations.

Of course, providing these models will not silence philosophical debate about the standards that experimental philosophers set, and there will always be room for empirical challenges grounded in other sorts of data. But where such models are in place, it will be clear where such debates should take place. Perhaps more importantly, experimental philosopher will be well positioned to respond to the challenges that are advanced by traditionalists and experimentalists alike. They

will have a clear account of why their results should be seen as philosophically meaningful, and they will have explained why their data speak in favor of, or speak against a philosophically interesting hypothesis. Challenges to a particular study can thus be framed in a way that advances philosophical debates, and deepens empirical investigations. Rather than focusing on abstract suppositions about the relevance of folk-psychological data to philosophical investigations, we can get down to the dirty business of mapping folk-psychological intuitions and deciding whether to defend conservative or revisionary hypotheses. In this way, experimentalists and traditionalists can begin to work together to address philosophically interesting hypotheses. I take it that this was one of the main, motivating goals of experimental philosophy all along.

At this point, I hope that it will be clear that my concern is not with the truth or the falsity of experimental results per se. I agree wholeheartedly with the claim that my arguments do not, and cannot establish that experimental philosophers are likely to make false claims. But neither philosophical nor scientific investigations are about truth or falsity *as such*. My primary concern is the extent to which experimental philosophers can be accountable for the claims they make. As Eric Winsberg and his colleagues (in press) argue, there are important differences between questions of scientific accountability and questions about the extent to which a scientific research process yields true claims about the world:

> Many true justified claims are not contributions to science, and many real contributions do not involve true justified claims—to put it mildly. Only claims that can be backed up by someone who is accountable for how they are produced and presented can count as legitimate contributions to the scientific conversation. For someone to be accountable for a scientific claim, she must believe that there is coherent set of epistemic and methodological standards that govern its production, and she must take responsibility for defending those standards and explaining how they are met.

My aim in this paper has been to show that the methodology that is employed in experimental philosophy yields a burden, to be accountable for the claims that are made in precisely this sense. Perhaps it is possible to get away with something less than a fully explicit set of standards at some points during the research process; but it must be possible to explicate the standards that have been employed when doing so becomes necessary. And one thing we should keep in mind about philosophers is that they are always ready to pose a challenge.

## 5. Coda

Some behavioral research does avoid the worries I have raised in this paper. But I am apprehensive about discussing it for three related reasons:

1. Data always calls for interpretation, and there is always room to disagree about interpretation even where plausible models of data are provided.

2. Strategies for addressing my worries are likely to be case-specific; what works in one situation may not work in another, and it is hard to know *a priori* which insights gained from successful research provide a plausible strategy for addressing my worries in other cases.

3. I don't know how to address my worries about Likert-scales, and behavioral research that is more successful typically uses other measures.

Nevertheless, I include a case study to highlight two ideas that are likely to be broadly relevant for thinking about how to move forward: we should strive to calibrate response measures (instead of using un-calibrated Likert-scales); and we should use multiple approaches to reveal that patterns are detectable, measurable, and definable from different and independent perspectives (Wimsatt 1974; 2007). Supposing that it is possible for behavioral results to be triangulated against results derived using methods that depended on different commitments and different presuppositions, we will have good reason to believe that the data are tracking a real pattern. This provides a foundation for interpreting them in a way that is explanatorily productive. As I argued in the introduction, this is just what it would take to avoid the sorts of errors that emerged in the experiments on cold fusion; and I maintain that this is often what it will take to show that behavioral results are tracking philosophically significant patterns.

Molly Crockett and her colleagues (2014) examined people's willingness to inflict pain on themselves and on others for profit. Each participant made between 150-160 choices about whether to accept a sum of money and whether to inflict a specified number of electric shocks. These decisions varied along two dimensions: who would receive the shocks (self vs. other), and whether the choice decreased the number of shocks at a cost, or increased them for a benefit. Since people experience shocks differently, Corckett and her colleagues first determined the point at which each participant found a shock to be painful but not intolerable (Crockett 2014 reports that the threshold varied from 0.4 mA to more than 10 mA). They then informed participants that shocks would be delivered at each person's subjective pain threshold; so each shock had the same value for every participant. Finally to avoid the possibility that people would habituate to repeated shocks (making later choices depend on earlier ones), they informed participants that one choice would be selected at random, and that the quantity of shocks and rewards from that choice would be delivered at the end of the experiment. This yields a response measure that is better calibrated than a Likert scale, and it allows for direct comparisons of the responses that different people offer from different subjective perspectives.

Crocket and her colleagues found that people chose to shock others less frequently than they chose to shock themselves, and that they chose to inflict

fewer shocks on others than they chose to inflict on themselves.[10] They also found that people were willing to increase the number of shocks they received for a small reward, but that nearly twice as large of a reward was necessary to choose to shock someone else.[11] Follow-up tests revealed that these effects were similar when the number of shocks was increased to yield a profit, and when the number of shocks was decreased at a cost.[12]

To investigate the nature of these responses, Crockett and her colleagues (2014, 17321) compared these data to various computational models and "found that decider's choices were most parsimoniously explained by a model that allowed for distinct valuation of harm to self and other, together with a factor that accounted for loss aversion for both shocks and money".[13] Targeted analyses were then carried out to show that this model accurately predicted participant responses. Indeed, most participants did place a higher cost on harming another person than they placed on harming themselves, and variation in these responses was captured by a single loss-aversion parameter—demonstrating a tight correlation between aversion to harming oneself and another person, and a tight inverse correlation between the value of increasing pain and decreasing money.

Crocket and her colleagues also found that people were slower to respond when inflicting shocks on others than when they were inflicting shocks on themselves; and they found a negative correlation between trait-psychopathy and the aversion to causing pain (both to self and others), which moderated an apparent gender difference in pro-sociality. This allowed them to embed their data in existing accounts of the processing deficits that arise in sub-clinical psychopathy, and it allowed them to rule out a potential confound. Finally, they ruled out the possibility that people assumed that they could tolerate more pain than others could, and showed that subjective reports of this sort were not predictive of the responses people actually gave.

This left two plausible explanations of the data (Crockett et al 2014, 17323). According to the first, people may be averse to causing bad outcomes, yielding slower response times because pro-social responses require being more thoughtful. According to the second, people may respond more slowly because the outcomes are less certain and this calls for greater deliberation. Crockett and her colleagues argue that some participants saw themselves as adopting as risk-averse strategy; and an exploratory analysis revealed that decisions directed toward others were 'noisier', which is what would be predicted if this result were driven by uncertainty. Either way, the data suggest that people have a "disposition to value others' suffering more than one's own", and addressing the

[10] In Experiment 2: $F(1, 40)=7.033$, $p=.011$; $F(1, 40)=6.30$, $p=.016$

[11] Study 2: $t(40)=2.039$, $p=.048$; $t(40)=2.703$, $p=.01$

[12] Increasing the number of shocks for a profit: $t(40)=2.195$, $p=.034$; $t(40)=2.027$, $p=.049$. Decreasing shocks at a cost: $t(40)=2.696$, $p=.01$; $t(40)=2.6517$, $p=.011$.

[13] This model correctly predicted 90% of choices in Experiment 2; Bayesian model comparisons were used to show that this model was favored over a number of alternatives, including more standard models of economic choice behavior.

origin of this disposition would require examining other work, which would take us beyond the bounds of this coda.

I hope this brief summary helps to clarify what it takes to be sensitive to the some of the factors that I have discussed in this paper. Crocket and her colleagues calibrate responses in a way that accommodates predictable patterns of behavioral variation; so they can compare responses without worrying about the effects of intersubjective variation in response strategies. This provides data that can be fit to more general models of decision-making. And importantly, the models of decision-making they examine are derived from research in learning theory, and they gain support from results in computational neuroscience and machine learning. By embedding their behavior data in the context of these models, Crocket and her colleagues show that multiple empirical methods, which depend on different commitments and different assumptions, converge on the same patterns in the world. While there is still plenty of room for debate about the relationship between these models and the data, this puts them on firm ground in taking the data to support a theoretically meaningful outcome: in this context, people cared "more about an anonymous stranger's pain than their own pain, despite the fact that their decisions were completely anonymous, with no future possibility of being judged adversely or punished" (Crockett et al 2014, 17323). This is a surprising result, and it is worth pursuing both theoretically and empirically.

**Acknowledgments:** Eric Winsberg and Rebecca Kukla helped me see that the relationship between models of data and scientific explanation was relevant to experimental philosophy. I received helpful feedback on an early version of this paper from Rik Hine and an audience at the Southern Society for Philosophy and Psychology (Austin, 2013). Ruth Kramer, James Mattingly, and J. Brendan Ritchie read drafts of this paper, and offered comments that made the arguments stronger than they otherwise would have been. Finally, I would like to thank all of the anonymous reviewers of this paper; I appreciated the time they took to offer comments, even where I disagreed with them.

## Works cited:

Amit, E. & J. Greene (2012). You see, the ends don't justify the means. Psychological Science, 23 (8), 861-868.

Bogen, J. & J. Woodward (1998). Saving the phenomena. The Philosophical Review, 97, 3, 303-352.

Buckwalter, W. & M. Phelan (2013). Function and feeling machines. Philosophical Studies, 166, 2, 349-361. Online supplementary material available at http://goo.gl/D27JTg, accessed 1 March 2015.

Chambers, C. et al (2013). Trust in science would be improved by study pre-registration. The Guardian, 5 June 2013, http://goo.gl/L1Hzck, accessed 31 January 2014.

Chambers, C. (2012). The Dirty Dozen: A wish list for psychology and cognitive neuroscience, http://goo.gl/XluVRQ, Accessed 31 January 2014.
Crockett, M. (2014). “Behind the scenes of a 'shocking' new study on human altruism,” The Guardian, http://gu.com/p/43gpm/stw (accessed, 2 December 2014).
Crockett, M., Z. Kurth-Nelsona, J. Siegela, P. Dayan, & R. Dolan (2014). “Harm to others outweighs harm to self in moral decision making,” PNAS, 111, 48: 17320-17325.
Cummins, R., M. Roth, & I. Harmon (2014). Why it doesn’t matter to metaphysics what Mary learns. Philosophical Studies, 167, 3, 541-555.
Haugeland, J. (1991). Representational genera. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 61–89). Hillsdale, NJ: Lawrence Erlbaum Associates.
Nuzzo, R. (2014). Scientific method: Statistical Errors. *Nature* 506, 150–152.
Rosenthal, D. (2011). Mental quality, valence, and intuition: Comments on Edouard Machery. Available at: https://wfs.gc.cuny.edu/DRosenthal/www/DR-MERG.pdf
Shanahan, K. (2002). A systematic error in mass flow calorimetery demonstrated. Thermochimica Acta, 382 (2), 95-100.
Schulz, E., Cokely, E.T., & Feltz, A. (2011). Persistent bias in expert judgments about free will and moral responsibility. Consciousness and Cognition, 20, 4, 1722-1731.
Suppes, P. (1962) Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), Logic, Methodology and Philosophy of Science. Stanford: Stanford University Press, 252-261.
Wimsatt, W. (1974). “Complexity and organization,” in K. Schaffner & R. Cohen (eds) Boston Studies in the Philosophy of Science, 20, 67–86.
Wimsatt, W. (2007). Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality. Cambridge: Harvard University Press.
Winsberg, E., Huebner, B., & Kukla, R. (in press). Accountability, values, and social modeling in radically collaborative research. Studies in the History and Philosophy of Science, 46:16-23.