Telling More Than We Can Know About Intentional Action

CHANDRA SEKHAR SRIPADA AND SARA KONRATH

Abstract: Recently, a number of philosophers have advanced a surprising conclusion: people's judgments about whether an agent brought about an outcome *intentionally* are pervasively influenced by normative considerations. In this paper, we investigate the 'Chairman case', an influential case from this literature and disagree with this conclusion. Using a statistical method called s*tructural path modeling*, we show that people's attributions of intentional action to an agent are driven *not* by normative assessments, but rather by attributions of underlying values and characterological dispositions to the agent. In a second study, we examined people's judgments about what they *think* drives asymmetric intuitions in the Chairman case and found that people are highly inaccurate in identifying which features of the case their intuitions track. In the final part of the paper, we discuss how the statistical methods used in this study can help philosophers with the *critical features problem*, the problem of figuring out which among the myriad features present in hypothetical cases are the critical ones that our intuitions are responsive to. We show how the methods used in this study have some advantages over *both* armchair methods used by traditional philosophers and survey methods used by experimental philosophers.

Introduction

A chairman is approached by his assistant and told about a new program they are thinking of starting that will help profits and harm the environment. The Chairman replies, 'I don't care at all about harming the environment, I just want to make as much profit as I can'. So the Chairman starts the program, and sure enough, the environment was harmed. Did the Chairman intentionally harm the environment? Most people say yes. However, when the same case is presented with the word 'harm' replaced by the word 'help', most people say the Chairman did *not* intentionally help the environment. Notice the two cases appear similar in all respects but for the different moral and evaluative responses they elicit (for example, typically, harming the environment is viewed as *bad* while helping the environment is viewed as *good*). Based on the Chairman case, and a number of cases in the literature much like it, many philosophers and psychologists have reached the somewhat surprising conclusion that normative factors pervasively

C.S.S. would like to thank audiences at Princeton University and the University of Michigan Ethics Discussion Group for valuable comments on earlier versions of this work. S.K.'s research was supported by a fellowship from the American Association of University Women.

Address for correspondence: Department of Philosophy, University of Michigan, 2215 Angell Hall 435 South State Street, Ann Arbor, MI 48109-1003, USA.
Email: sripada@umich.edu

Mind & Language, Vol. 26, No. 3 June 2011, pp. 353–380.

influence judgments of intentionality (Doris *et al*., 2007; Knobe, 2003, 2005, 2006; Mele, 2006; Nadelhoffer, 2004a and b, 2006; Alicke, 2008; Wright and Bengson, 2009; Holton, 2010). While substantial disagreement remains about which *specific* normative factors are at work, the minimal thesis that some normative factor or other exerts effects on intentionality judgments is now not much in dispute. Reporting on the widespread consensus, Joshua Knobe, the philosopher who originated the Chairman case, writes, 'At this point, there can be little doubt that moral considerations have an impact on people's use of the word ''intentionally''. The key remaining questions are about how this effect is to be understood' (Knobe, 2006).

In this paper, we aim to investigate the factors that drive judgment in the Chairman case using methods quite different than those used in prior studies. In Part 1 of the paper, we report on a novel empirical study that uses a statistical method called *structural path analysis*. A key advantage of this method is that it allows multiple factors associated with different models of intentionality judgments to be quantified and directly compared against each other (see Sinnott-Armstrong *et al*., 2008 for an application of structural modeling methods to understanding factors influencing judgement in Trolley Problems). Using structural path modeling, we show that normative factors are in fact unlikely to play a substantial role in driving people's divergent judgments in the Chairman case. We show the evidence instead supports an alternative view, based on an existing model called the *Deep Self Concordance Model* (Sripada, 2010), which holds that asymmetric judgments are significantly driven by assessments of the Chairman's underlying values, attitudes, and stable behavioral dispositions. But over and above our partisan defense of our favored Deep Self Concordance Model of intentionality judgments, a key theme in this part of the paper is *methodological*: we aim to illustrate the method of structural path modeling and point out ways that it can supplement existing methods used by philosophers, both of the traditional and the experimental variety.

In Part 2, we report on the results of a second study in which participants were presented with both versions of the Chairman case and asked what explains asymmetric intentionality judgments. Results showed that people's 'tracking judgments', i.e. their intuitive judgments about what features of cases their intuitions are responding to, are highly inaccurate. By a very large margin, the very normative factors that we showed in our first study are *not* driving asymmetric judgments in the Chairman case are overwhelmingly cited by people as explaining why people make asymmetric judgments.

Why are people's tracking judgments so inaccurate? Our explanation, in abbreviated form, goes as follows: Differences in normative features between the two versions of the Chairman case are highly *salient* and readily *accessible* to awareness. But the actual processes that underpin judgments of intentional action are not so readily accessible. These processes operate mostly outside conscious awareness and track features of cases that cannot easily be introspected. So when asked to furnish an explanation for divergent judgments of intentional action in the Chairman case, it is not surprising that people focus on normative features that are salient and consciously apparent, even if these features are not what actually drive their

judgments. We call the preceding explanation of why people focus on normative factors in explaining the Chairman asymmetry the 'Telling more than we know' explanation, in homage to Richard Nisbett and Timothy Wilson's famous 1977 paper in psychology that drew early attention to a broadly similar set of phenomena.

In the final part of the paper, we explore how the experimental and statistical methods used in this paper may be more generally applied by philosophers. Here we argue that standard methods of philosophical inquiry demand that philosophers solve the '*critical features problem*', the problem of figuring out which among the myriad features present in hypothetical cases are the critical ones that drive our intuitive responses. Given inaccuracies in people's tracking judgments, we argue that structural path modeling and related statistical techniques have an especially useful role to play in helping philosophers solve the critical features problem.

Because we seek to reach an audience mainly of philosophers who may be unfamiliar with structural path modeling, we present our main results in informal language and place more technical details in footnotes. In addition, we offer extra explanatory material throughout to aid understanding of our methods. Our hope is that by introducing a wide philosophical audience to structural path modeling and related methods, others may perhaps learn more about these methods and apply these methods to other philosophical questions.

1. What Drives Asymmetric Intuitions in the Chairman Case? A Structural Path Analysis

1.1 Study Vignettes and Questions

In order to investigate the basis of asymmetric intuitions in the Chairman case, we conducted a study in which 240 University of Michigan undergraduates completed a brief survey. Participants (n = 120) in the 'harm' condition read the following paragraph.

> The vice-president of a company 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 also harm the environment.'
>
> The chairman of the board 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 new program. Sure enough, the environment was harmed.

Participants in the 'help' condition (n = 120) read an identical paragraph, but the word 'harm' was replaced with the word 'help'.

Most studies of the Chairman case ask participants the question: Did the Chairman harm [help] the environment intentionally? They then go on to compare

Question #	Question Wording	Anchors for 7-point scale	Abbreviated Name
1	How much do you agree with the statement 'The Chairman intentionally harmed [helped] the environment'?	Strongly Agree, Strongly Disagree	Intentionality Judgments
2	In your view, how good or bad is the outcome that the environment is harmed [helped]?	Very Good, Very Bad	Goodness/Badness Judgments
3	In your view, what is the Chairman's moral status?	Very Moral, Very Immoral	Moral Status Judgments
4	What are the Chairman's values and attitudes towards the environment?	Very Pro-environment, Very Anti-environment	Chairman Values/Attitudes Judgments
5	In the vignette above, the Chairman's action brings about an outcome in which the environment is harmed [helped]. In your view, to what extent is the Chairman the kind of person who will, in other contexts and situations, bring about outcomes similar to this one?	Very Likely, Very unlikely	Generalizability Judgments
6	What are your own values and attitudes towards the environment?	Very Pro-environment, Very Anti-environment	Self Values/ Attitudes Judgments

Table 1 *Questions used in the study. Participants answered by placing an 'X' on a seven point scale below each question.*

the proportion of participants who say 'yes' to this question in the harm condition against the proportion saying 'yes' in the help condition. In the present study, in addition to asking a similar question about intentionality (the precise wording is found in Table 1), we asked participants five additional questions designed to measure key candidate explanatory variables that might be driving people's judgments (Table 1). The order of presentation of all questions was systematically

varied to mitigate potential order effects (120 counterbalanced orders were used in the harm condition as well as the help condition).[1]

1.2 The Candidate Models

The questions in Table 1 were used in order to test a number of candidate models for explaining asymmetric judgments of intentionality in the Chairman case (Table 2). The first model is the 'Good/Bad Model', which was initially proposed by Joshua Knobe (Knobe, 2005, 2006). Though Knobe no longer subscribes to the model (Knobe's current model is discussed and tested in Section 1.4), the Good/Bad Model has been influential in the literature in recent years and continues to be explicitly endorsed (e.g. Beebe and Buckwalter, 2010), or implicitly assumed by a number of writers, and thus warrants continued scrutiny and testing. According to the Good/Bad Model, people's intentionality judgments are strongly influenced by the evaluative valence of the outcome brought about by the agent. That is, people are much more willing to say an agent brought about an outcome intentionally if they view the outcome as bad compared to when they view the outcome as good. Applying this model to the Chairman case, the outcome of the environment's being harmed is viewed as bad in the harm condition, but the outcome of the environment's being helped is viewed as good in the help condition—thus resulting in differing judgments of intentionality. Another influential model, developed by Mark Alicke (2008, 1992), focuses not on the goodness or badness of the *outcome*, but rather on the moral status of the *agent*. According to this 'Moral Status Model', seeing an agent violate basic moral standards causes people to experience attitudes of moral blame and moral disapproval, which in turn makes them more likely to say the agent intentionally brought about the outcome.[2] Applying this model to the Chairman case, in the harm condition the Chairman is judged as being an immoral person because his harming the environment violates basic moral standards, so people are more likely to say he intentionally brought about the outcome. In the help condition, the Chairman is judged to be an immoral person to a much lesser extent because he does not harm the environment, so people will be correspondingly less likely to say he intentionally brought about the outcome—resulting in an overall asymmetry in intentionality judgments between

[1] Question 4 and 6 were grouped into a single block. The block was allowed to vary in order of presentation with respect to the other questions, but within this block, order did not vary (Questions 4 was always asked prior to Question 6). Grouping these two questions into a block was done for two reasons: 1) to reduce the number of counterbalanced orders to a more manageable level. A total of five elements (four individual questions and one question block) were counterbalanced, so there were 120 orders (representing *all* possible orders of these five elements) in the harm condition as well as the help condition; and 2) Questions 4 and 6 had very similar wording, and it was thought that presenting them next to each other would enhance participants' ability to rapidly see how the two questions differed.

[2] Thomas Nadelhoffer has put forward an interesting model that in some ways is closely related to the Moral Status Model. We discuss his model in detail in Appendix A.

Model	Author	Explanatory Variable(s) (measured in this study) associated with each model
Good/Bad Model	Joshua Knobe (no longer endorses this model)	Goodness/Badness Judgments
Moral Status Model	Mark Alicke	Moral Status Judgments
Deep Self Concordance Model	Chandra Sekhar Sripada	Chairman Values/Attitudes Judgments, Generalizability Judgments
Indirect Influence Model	Joshua Knobe	Potentially all the variables
Interaction Model	Thomas Nadelhoffer	Relationship between Goodness/Badness Judgments and Moral Status Judgments

Table 2 *Summary of the models of intentionality judgments tested in this study.*

the two conditions.[3] We measured the predictions of the Good/Bad Model and Moral Status Model (which we often refer to together as 'normative factor models') with questions 2 and 3, respectively (Table 1).

Another model we sought to test is the Deep Self Concordance Model. This model is more complex and is explained in depth in another paper (Sripada, 2010), so here we keep our explanation brief. According to the Deep Self Concordance Model, many factors play a role in influencing intentionality judgments, but an underappreciated factor is whether the outcome the agent brings about concords with the agent's underlying core values, attitudes, and behavioral dispositions, which constitute the agent's 'Deep Self'. The Deep Self Concordance Model differs from normative factor models in that it focuses *not* on people's evaluative judgments of the agent, but rather on people's *ascriptions* of evaluative attitudes to the agent. Evaluative judgments of an agent and ascriptions of evaluative attitudes to an agent have a similar 'sound' and they often co-occur (that is, people routinely make *both* kinds of assessments simultaneously), so it is sometimes easy to get these two kinds of assessments confused. But to understand the Deep Self Concordance Model, it is essential to see that these two assessments are fundamentally different. If George Bush makes a speech about his views on abortion, people will form assessments about the descriptive question of what are Bush's core values and attitudes towards abortion (e.g. pro- versus anti-abortion attitudes), as well as the evaluative question of whether Bush is moral or immoral for holding the values and attitudes that he

[3] The Moral Status Model does not require that the Chairman is judged to be moral (rather than immoral) in the help condition. It only requires that that there is a significant difference in the degree to which he is judged to be immoral in the harm and help conditions.

does. These two judgments regarding George Bush are fundamentally distinct, as evidenced by the fact that most people probably agree that Bush has values and attitudes opposed to abortion, but nevertheless there is substantial disagreement about whether Bush is moral or immoral for holding these values and attitudes. With these distinctions in mind, we can now see the key difference between the Deep Self Concordance Model and normative factor models. The Deep Self Concordance Model claims that it is *descriptive* judgments about the core values and attitudes possessed by an agent (i.e. ascriptions of values and attitudes to the agent's Deep Self) that influence intentionality judgments. People also typically make e*valuative* judgments of the agent (or the outcome that the agent brings about), but the Deep Self Concordance Model claims that these evaluative judgments do *not* influence intentionality judgments.[4]

Applying the Deep Self Concordance Model to the Chairman case, the model first predicts that in both the harm and help condition, people ascribe to the Chairman core underlying anti-environment values and attitudes. This is because the Chairman says 'I don't care at all about harming [helping] the environment', which is taken to express contempt or hostility towards the environment. In the harm condition, the outcome of harming the environment is concordant with the Chairman's underlying anti-environment attitudes, so people say the Chairman brought about the outcome intentionally. In the help condition, the outcome of helping the environment is discordant with the Chairman's underlying anti-environment attitudes so people say the Chairman did not bring about the outcome intentionally. We measured the predictions of the Deep Concordance Self Model in two ways. With question 4 (Table 1), we probed participants' views about the Chairman's values and attitudes towards the environment. With question 6, we probed people's views about the *generalizability* of the kind of outcome the Chairman brought about to other contexts and situations. According to the Deep Self Concordance Model, generalizability of this sort is a characteristic feature of the attitudes contained in an agent's Deep Self. That is, the values and attitudes of the Deep Self are a core and stable part of the person (i.e. they are *trait*-like) and thus they dispose the person to bring about outcomes concordant with these values and attitudes across a range of situations and contexts. For example, suppose one person deeply values education and academic success, while another person deeply disdains education and views it with contempt, and both do very poorly on a particular examination. People will say that the second person, but not the first person, is the kind of person who will, in other contexts and situations, bring about similar outcomes, because the outcome of doing poorly on an exam concords with the anti-education attitudes of the second person's Deep Self, but not the pro-education attitudes of the first person's Deep Self. In a similar way, asking whether the Chairman will bring about similar outcomes in other contexts and

[4] The difference between the Deep Self Concordance Model and normative factor models is also discussed at length in Sripada (2010), especially Section 3.2.

situations provides another way to probe whether participants see the outcome associated with the Chairman's action as springing from the values, attitudes, and behavioral dispositions of his Deep Self.[5]

We asked an additional question to probe subjects' own values and attitudes towards the environment (Question 6, Table 1). This question is relevant to testing the Indirect Influence Model, which we discuss in section 1.4. Finally, questions 2 and 3, which probed people's normative judgments of the outcome and the agent, respectively, allowed us to perform a test of an important model of intentionality judgments put forward by Thomas Nadelhoffer. Because of certain statistical limitations, we could not test Nadelhoffer's model at the same time as we tested the other normative factor models, and thus we discuss Nadelhoffer's model separately in Appendix A. Overall, five candidate explanatory variables were measured, three variables associated with various normative factor models and two associated with the Deep Self Concordance Model. Our main question was: What are the relative contributions of each of these candidate variables to explaining the asymmetry in intentionality judgments in the Chairman case? *Structural path analysis* provides a powerful quantitative method by which this question can be answered.

1.3 A Structural Path Analysis of Responses to the Chairman Case

In Figure 1, a structural path model[6] is shown depicting relationships between the key variables associated with our study—the 'Case' variable, which represents our manipulation (presenting participants with either the harm or the help version of the survey), the four candidate explanatory variables, and the outcome variable (levels of agreement with the statement that the Chairman intentionally harmed [helped] the environment). Each variable is hereafter referred to using the shorthand notations in Table 1. *Paths*, shown in Figure 1 as black unidirectional arrows, represent links between these variables, and their positions correspond to our antecedent hypothesis that Case influences the four candidate explanatory variables which in turn influence Intentionality Judgments. An additional path links Case *directly* with Intentionality Judgments. This path represents the influence of Case on Intentionality Judgments that is not mediated through influences on the candidate explanatory variables.

Paths are associated with coefficients called *path coefficients*. Each coefficient represents the magnitude of the influence of the variable at the start of the path on the variable at the end of the path, controlling for the influence of other

[5] Generalizability is a *characteristic* feature of attitudes of the Deep Self, but not a *necessary* feature. For example, the death of Moby Dick is an outcome that is strongly concordant with Captain Ahab's Deep Self. But it is not easy to see how Captain Ahab could bring about outcomes like this one in other contexts and situations because Moby Dick is a single individual and his death is an event that can only occur once.

[6] All structural path analyses were performed with *EQS* (Multivariate Software Inc, Encino CA), a commercially available software package for implementing structural equation modeling.

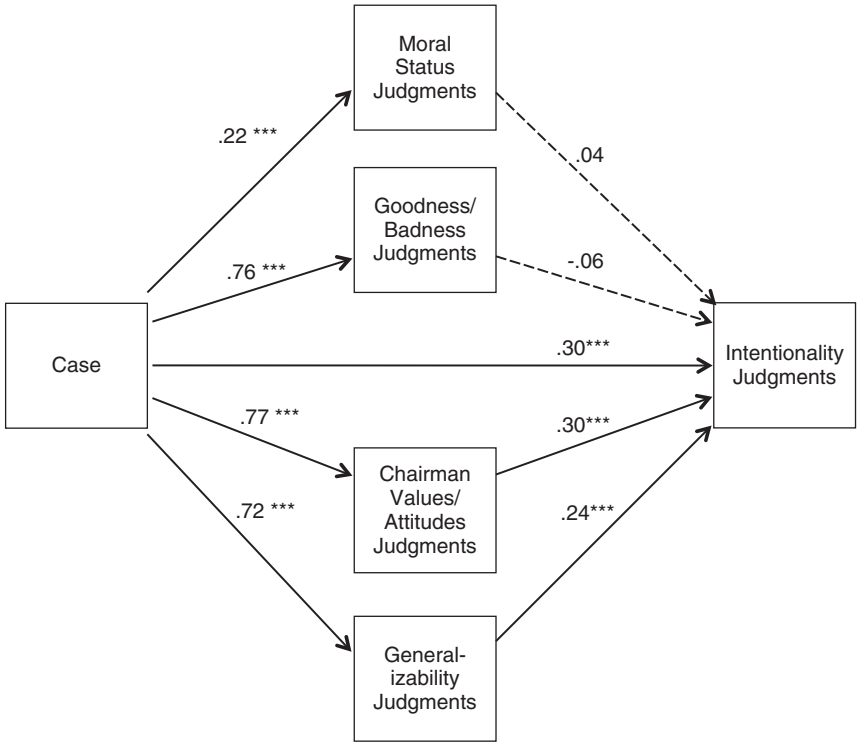

Figure 1 *Initial structural path model of four candidate variables that potentially explain the relationship between the case manipulation and judgments of intentional action in the Chairman case. Statistically significant paths are shown as solid arrows, while non-significant paths are shown as dashed arrows.* $^{*} = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$

predictor variables. More specifically, suppose the value of the path connecting some variable '*A*' (at the start of a path arrow) to another variable '*B*' (at the end of this path arrow) is p. Then the interpretation of this path coefficient is: If *A* is increased by 1 unit, then *B* changes by p units, holding fixed the contribution of all other variables that have paths directly connecting them to *B*.[7] The sign of a path coefficient reflects the direction of the influence between *A* and *B*. To ease interpretation of Figure 1, we anchor the sign of path coefficients so that positive numbers mean that the direction of influence along the path is that predicted by the model associated with that path. For example, the *Case* → *Moral Status Judgments* path is positive indicating that the prediction regarding this path by the Moral Status Model (that people will rate the Chairman as being more immoral in the harm

[7] As is common practice, path coefficients in Figures 1 and 3 are expressed in *standardized units* to make comparisons between the different coefficients in the models meaningful.

condition than in the help condition) is borne out by the data, and similarly for the other paths in the model.[8]

Path coefficients are calculated by the method of *maximum likelihood*, an estimation procedure with many desirable properties, and that is widely used in structural path analysis (Bentler, 1995).[9] Once path coefficients are calculated, statistical tests can be employed to assess 1) the goodness of fit of the overall structural path model; and 2) the statistical significance of individual paths. In Figure 1, statistically significant paths are shown with asterisks next to the path coefficient. If a path is not statistically significant (that is, the null hypothesis that the path's coefficient is *not* different than zero *cannot* be rejected), then no asterisk is displayed and the path is shown as a dotted arrow. In order to interpret path coefficients in a model, it must first be confirmed that the model has good overall fit. Interpreting path coefficients in a model with poor fit is akin to interpreting distances in a map that is known to be inaccurate. The model in Figure 1 has '*good*' fit with the data. Later, we will introduce another model (Figure 3) with significantly *better* fit with the data. However, most path coefficients in the two models are nearly identical (the difference between the two models is that some paths have been added and other paths removed). Since it will significantly ease exposition to first discuss the model in Figure 1, we will defer discussion of the better fitting model (and the methods by which it was derived) for later.

In interpreting the model in Figure 1, let us start with paths linking Case to the four candidate explanatory variables. These paths are all solid with three asterisks next to the path coefficients, indicating that with a very high degree of statistical confidence, *CASE* does in fact influence these candidate explanatory variables. The *CASE → Goodness/Badness Judgments, CASE → Chairman Values/Attitudes Judgments, and CASE → Generalizability Judgments* paths are particularly large, indicating that a large portion of the variation in these variables is explained by

[8] The predictions of each model are discussed in section 2.2. The predictions of the Deep Self Concordance Model require that the Chairman Values/Attitudes Judgments variable be 'reverse coded' in one of the two conditions. Reverse coding means that the variable is flipped around its midpoint. Thus on a 7-point scale, a 1 is scored as a 7, a two as a 6, and so on. Reverse coding of this variable is required for one of the two conditions because the Deep Self Concordance Model predicts correlations of *opposite directions* in the two conditions. That is, according to the Deep Self Concordance Model, in the harm condition, rating the Chairman as more anti-environment predicts *greater* agreement with the statement that the Chairman intentionality harmed the environment. But in the help condition, rating the Chairman as more anti-environment predicts *lesser* agreement with the statement that the Chairman intentionally helped the environment. The path between Case and Intentionality Judgments is positive reflecting the 'dummy coding' scheme we employed in which the harm condition was coded as '0' and the help condition was coded as '1'.

[9] All dependent variables were inspected for deviations from the normality assumptions of maximum likelihood estimation. For all dependent variables, skewness was less than 0.5 and kurtosis was less then 1.5. In addition, collinearity diagnostics were inspected and the presence of multicollinearity was disconfirmed by tolerance values >0.2 and variance inflation factors less than 4.

Case. If we were to stop the analysis here, we would conclude that all four variables contribute to explaining asymmetries in intentionality judgments.

However, the paths on the right side of Figure 1 linking the four candidate explanatory variables to Intentionality Judgments suggest a quite different interpretation. According to this portion of the model, Goodness/Badness Judgments and Moral Status Judgments do *not* make a statistically significant contribution to Intentionality Judgments. In contrast, the contributions of Chairman Values/Attitudes Judgments and Generalizability Judgments to Intentionality Judgments are highly statistically significant. Furthermore, Case also significantly contributes to Intentionality Judgments through a direct path that represents the influence of the case manipulation on Intentionality Judgments that is *not* mediated through the four candidate explanatory variables.

Notice the pattern observed in Figure 1 in which Case is a strong predictor of Goodness/Badness Judgments and Moral Status Judgments, but these latter two variables are relatively weak and highly insignificant predictors of Intentionality Judgments. This pattern is highly suggestive that Goodness/Badness Judgments, Moral Status Judgments, and Intentionality Judgments all have a common cause—Case. If two variables, say Variable 1 and Variable 2 (Figure 2), are heavily influenced by a common cause (and these variables do not otherwise causally influence each other), the influence arising from this common cause will result in Variable 1 and Variable 2 being strongly correlated. But this relationship is *spurious*. If one controls for the variance associated with the common cause, then Variable 1 and Variable 2 should be uncorrelated (i.e. the correlation between Variable 1 and Variable 2 should be non-significant). This is indeed the pattern we observe in our results. The correlation between Goodness/Badness Judgments and Intentionality Judgments is .51 ($p < 0.001$), and the correlation between Moral Status Judgments and Intentionality Judgments is .18 ($p = 0.005$), both highly significant. But the *partial correlation* between Goodness/Badness Judgments and Intentionality Judgments after controlling for Case is .01 ($p = 0.891$),[10] and the *partial correlation* between Moral Status Judgments and Intentionality Judgments after controlling for Case is .05 ($p = 0.535$), both non-significant. This pattern is consistent with the hypothesis that Goodness/Badness Judgments and Moral Status Judgments and Intentionality Judgments are all causally influenced by Case, *but Goodness/Badness Judgments and Moral Status Judgments are not themselves causes of Intentionality Judgments*.

In contrast, Case is a significant predictor of Chairman Values/Attitudes Judgments and Generalizability Judgments, and these latter two variables remain significant predictors of Intentionality Judgments, after controlling for the effects of other explanatory variables including Case. This pattern of results is consistent

[10] In another study of the Chairman case, Wright and Bengson (2009) also found the relationship between judgments of the goodness or badness of the outcome and intentionality judgments was *not* statistically significant after controlling for the effects of the case manipulation, lending additional support for the similar finding in the current study.

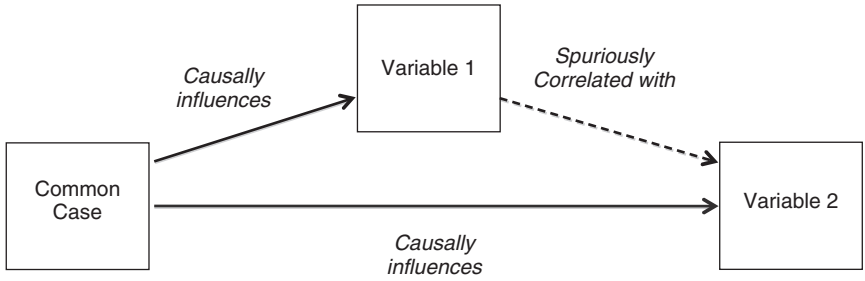

Figure 2 *Schematic diagram of two variables influenced by a common cause. The relationship between Variable 1 and Variable 2 is spurious*

with the hypothesis that Chairman Values/Attitudes Judgments and Generalizability Judgments are both causes of Intentionality Judgments.

Let us return to an issue that we deferred earlier—the issue of the overall goodness of fit for a structural path model. We can assess how well a structural path model fits with the observed data using a family of statistical measures called *goodness of fit indices*.[11] Using these indices, we found the model in Figure 1 has 'good' overall fit with the data.[12] An additional tool available for the researcher consists of *modification tests* that assess how the overall goodness of fit of a model would change if new paths were added to the model and/or existing paths were dropped from the model. Using modification tests, we found that model fit is significantly improved if two existing paths are removed from the model and one new path is added to the model. The two paths recommended for removal are the *Goodness/Badness Judgments* → *Intentionality Judgments* and *Moral Status Judgments* → *Intentionality Judgments* paths.[13] The recommendation that these paths be removed is consistent with our antecedent hypothesis that only Deep Self variables, but not

[11] There are a large number of goodness of fit indices available, each emphasizing certain dimensions of good fit. We follow published guidelines (Raykov *et al.*, 1991) in reporting five standard goodness of fit measures: the chi-square test, normed fit index (NFI), non-normed fit index (NNFI), comparative fit index (CFI), and root mean square error of approximation (RMSEA).

[12] $X^2(6, N = 240) = 12.00$, $p = 0.06$; NFI = .985; NNFI = .981; CFI = .992; RMSEA = .065. Using 'rule of thumb' cutoff values for the RMSEA (Hu and Bentler, 1999), the model displays 'good' overall fit with the data. Note, the X^2 test is a 'badness of fit test', and a non-significant p value indicates good fit of the model to the data.

[13] All modification tests were implemented in *EQS* (Multivariate Software Inc, Encino CA), and decisions to add or remove paths were evaluated in light of antecedent hypotheses and theoretical plausibility. The multivariate Wald Test (Bentler, 1995) for removing paths from the model shows that removing the paths between the two normative variables and Intentionality Judgments increases the overall model X^2 (an index of lack of fit of the model to the data) by 2.31, which is not a statistically significant difference ($p = 0.320$). The resulting model shows improvements in goodness of fit measures (such as the NNFI and RMSEA) that are sensitive to parsimony (i.e. the number of paths in the model).

normative variables, influence judgments of intentionality. The path recommended for addition to the model links Goodness/Badness Judgments with Chairman Values/Attitudes Judgments.[14] This path too has antecedent theoretical justification [for example, it has been discussed at length in prior work (Sripada, 2010)], and we expand on the meaning of this path a bit later when we discuss the 'Indirect Influence Hypothesis'. The resulting model, shown in Figure 3, has near perfect fit with the data.[15]

In addition to providing information about the significance of individual paths, structural path analysis also allows calculation of *mediation effects*, the influence of one variable on another that is mediated through one or more intervening variables (Baron and Kenny, 1986). Using calculations of mediation effects,[16] we found that 55% of the influence of the case manipulation on Intentionality Judgments is mediated through Chairman Values/Attitudes Judgments and Generalizability Judgments, the two Deep Self variables. Just 4% of the influence of Case on Intentionality Judgments is mediated through Goodness/Badness Judgments. As shown in Figure 3, this variable influences Intentionality Judgments only *indirectly* by first influencing the Deep Self Variables (i.e. through the *Chairman Values/Attitudes Judgments* → *Intentionality Judgments* path). There is no effect of Case on Intentionality Judgments that is mediated by Moral Status Judgments, as this variable does not have any paths connecting it with Intentionality Judgments at all. These mediation effects calculations provide us with strong evidence that *at least* a majority of the asymmetry effect in intentionality judgments in the Chairman case is explained by Deep Self variables, and normative factors play a lesser role in explaining the asymmetry. But notice that 40% of the effect of Case on Intentionality Judgments is still unaccounted for. This 'unexplained variance' of Case on Intentionality Judgments

[14] The Lagrange multiplier test (Bentler, 1995) was used to identify candidate paths to be added to the model. To avoid capitalizing on chance relationships in the data, and following a practice recommended in the literature (Bentler, 1995), we required a more conservative threshold for addition of new paths of $p < 0.01$. The Lagrange multiplier test indicates that adding a path linking Goodness/Badness Judgments and Chairman Values/Attitudes Judgments reduces the overall model X^2 (an index of lack of fit of the model to the data) by 6.37, which is statistically significant ($p = 0.012$). The direction of this path cannot be determined from our data because the reduction in X^2 is identical when a path is drawn between these variables in either direction. Thus in Figure 3, we chose to display the path in a direction that seems to us to be most psychologically plausible, though we cannot rule out the influence runs in the opposite direction. Nothing of importance hangs on the question of which direction this particular path runs.

[15] $X^2(7, N = 240) = 7.01$; $p = 0.43$; NFI = .991; NNFI = 1.00; CFI = 1.00, RMSEA = .002. Using 'rule of thumb' cutoff values for the RMSEA (Hu and Bentler, 1999), the model displays 'exact' overall fit with the data.

[16] The percentage of the effect of a predictor on an outcome that is mediated, sometimes referred to as an *mediation effect ratio*, is defined as: (the *indirect effect* of the predictor on the outcome)/(the *total effect* of the predictor on the outcome) (Shrout and Bolger, 2002). An *indirect effect* refers to the influence of a predictor on an outcome transmitted through some third variable (the mediator). All mediation effects were calculated using the total, direct, and indirect effect outputs in *EQS* (Multivariate Software Inc, Encino CA).

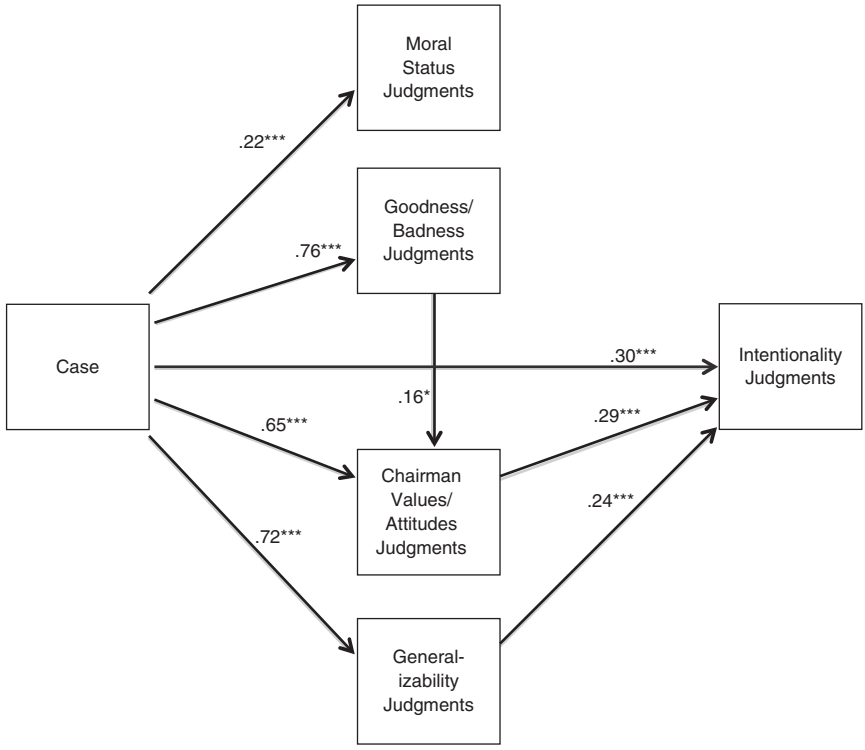

Figure 3 *Final structural path model of four candidate variables that potentially explain the relationship between the case manipulation and judgments of intentional action in the Chairman case. Statistically significant paths are shown as solid arrows, while non-significant paths are shown as dashed arrows.* $^{*} = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$

can be attributed to just one of two sources. First, some unexplained variance might be due to *systematic effects* of other variables that mediate the relationship between Case and Intentionality Judgments but are not measured in the present study. These other variables might be variables related to normative factors, Deep Self factors, or factors unrelated to either. Alternatively, some unexplained variance might be due to *unsystematic effects* (i.e. noise) associated with measurement of the variables in the model. Of course, all variables are measured with at least *some* error, and this is especially likely to be true in the current study in which all variables are measured with just one (as opposed to multiple) ratings.[17] If we (quite plausibly) suppose that a fair portion of the remaining unexplained effect of Case on Intentionality Judgments is due to measurement error of the candidate mediator variables, then it

[17] Measuring each variable using multiple closely related ratings *substantially* reduces measurement error, and will be implemented in future studies.

follows that the Deep Self variables explain *more than* 55% of the asymmetry effect in the Chairman case.

1.4 The Indirect Influence Model

But a potential alternative explanation of our results remains. It is possible that the Deep Self Concordance Model *is itself a version of a normative factor model*, in which normative factors operate *indirectly* to influence Deep Self-related variables.[18] For example, suppose that people's judgments of whether the Chairman is pro- or anti-environment are significantly influenced by the person's *own* values and attitudes towards the environment. That is, people who are more pro-environment tend to see the Chairman's behavior as normatively inadequate, thus rating the Chairman as *more* anti-environment. And, people who are more anti-environment tend to see the Chairman's behavior as more normatively adequate, thus rating the Chairman as *less* anti-environment. If normative factors such as these influence the variables associated with the Deep Self Concordance Model, then it can be argued that at least part of the success of the Deep Self Concordance Model should be attributed to normative factors, making the Deep Self Concordance Model simply a version of a normative factor model.

We already tested one version of this Indirect Influence Model earlier when we discussed mediation effects. There we noted that just 4% of the influence of Case on Intentionality Judgments is accounted for by the *Case* → *Goodness/Badness Judgments* → *Chairman Values/Attitudes Judgments* → *Intentionality Judgments* path. Thus though there *is* a statistically significant indirect effect of Goodness/Badness Judgments on Deep Self variables, the size of this effect is rather small—roughly fourteen times smaller than the effect of Deep Self Variables on Intentionality Judgments that is *not* mediated through Goodness/Badness Judgments. We performed a second test of the Indirect Influence Hypothesis using question 6 in Table 1, which asked participants to rate their own values and attitudes towards the environment. The Indirect Influence Model predicts that these Self Values/Attitudes Judgments will strongly influence Deep Self variables. However, results showed that the correlation between Self Values/Attitudes Judgments and Chairman Values/Attitudes Judgments was close to zero, and was non-significant ($r = 0.02$, $p = 0.780$). A similar pattern held for the correlation between Self Values/Attitudes Judgments and Generalizability Judgments ($r = 0.04$; $p = 0.580$).[19] This result provides strong evidence that

[18] In Knobe's latest writings, he rejects the Good/Bad Model which he originally proposed, and formulates a new model in which moral attitudes influence intentionality judgments *indirectly*, by first influencing attitude ascriptions to the agent [e.g. '. . .moral judgments can shift people's representations of an attitude along a scale from 'con' to 'pro.'' (Pettit and Knobe, 2009)]. Thus Knobe would presumably want to endorse the Indirect Influence Model that we are presently discussing.

[19] Of note, Self Values/Attitudes Judgments were also not *directly* correlated with Intentionality Judgments ($r = -.03$, $p = 0.648$).

participants' own values and attitudes towards the environment are *not* influencing Deep Self variables, and thus not serving to indirectly influence intentionality judgments.[20]

2. People's Intuitive Explanations of Asymmetric Judgments in the Chairman Case: The Role of Salience and Accessibility

In the previous part of the paper, we provided strong evidence that Deep Self-related factors explain a majority of the asymmetric judgment effect in the Chairman Case, while normative factors account for a small minority of the asymmetry. However, based on the experience of one of us (C.S.S.) in presenting the Deep Self Concordance Model to audiences in papers and talks, in terms of intuitive appeal, the order of priority is fully reversed—most people find normative factor models highly intuitive (and indeed some think it is simply *obvious* that these models are correct), while they find the Deep Self Concordance Model far less intuitive. In this part of the paper, we develop an explanation for this striking phenomenon that though Deep Self factors *actually* explain the majority of the Chairman asymmetry, people nevertheless *think* that the asymmetry is driven by normative factors.

To test the hypothesis that people intuitively tend to explain the Chairman asymmetry in terms of normative factor variables rather than Deep Self variables, we conducted an additional study. We presented 31 undergraduates at the University of Michigan with *both* the harm condition and the help condition of the Chairman case, shown on the left and right side of a page, respectively. Below the two cases, participants read: 'Researchers have found that if people are presented with the case on the left, they tend to say that the Chairman *intentionally* harmed the environment. However, if people are instead presented with the case on the right, then they tend to say the Chairman *did not intentionally* help the environment.' Participants were then asked to write a paragraph answering the question, 'What do you think explains why people answer differently in the two cases?' Participants' responses were then coded by two undergraduate research assistants blinded to the hypothesis of this study. Coders were presented with descriptions for the Good/Bad Model and Moral Status Model (grouped together and labeled as a 'Type 1 Explanation'), and the Deep Self Concordance Model (labeled as a 'Type 2 Explanation'), and asked to classify whether responses corresponded to one of these two kinds of

[20] In a separate study of the Chairman case, one of us (C.S.S.) asked 240 participants questions that probed their own view of the moral status of harming the environment (7-point scale ranging from 'Very Moral' to 'Very Immoral') as well as their agreement with the statements 'I believe that the environment ought to be protected' and 'Most people believe that the environment ought to be protected'. Ratings did not significantly correlate with Deep Self variables (or with intentionality judgments), providing additional evidence against various versions of the Indirect Influence Model.

explanations or 'None of the Above'. Coders agreed on 24 of 31 responses, and after being allowed to discuss codings with each other, agreed on 29 of 31 responses. Of the 29 responses for which there was agreement, 22 responses corresponded to a 'Type 1' normative factor explanation, and just 2 responses corresponded to a 'Type 2' Deep Self explanation, with the remaining 5 responses coded as 'None of the Above'.

The results of the preceding study, when combined with the results of the study reported in Part I of the paper, provide compelling evidence that people are prone to misidentifying the actual causes of intuitive judgment in the Chairman case. In Part I of the paper, we showed that compared to normative factor variables, Deep Self variables have much greater importance in driving asymmetric judgments in the Chairman case. However, in our second study, we showed that, by a very large margin, *the very normative factors* that we have shown (in Part I) are *not* driving asymmetric judgments in the Chairman case are overwhelmingly cited by people as explaining why people make asymmetric judgments.

Why are people's explanations for the Chairman asymmetry so inaccurate? We now sketch an account that draws heavily on the seminal work of Richard Nisbett and Timothy Wilson, and emphasizes a critical role for *salience* and *accessibility* in biasing people's explanations about the sources of their intuitive judgments. In their highly influential 1977 article, 'Telling more than we can know: verbal reports on mental processes', Nisbett and Wilson reviewed hundreds of studies that documented striking limitations in people's ability to report about the mental processes that give rise to their intuitive judgments (Nisbett and Wilson, 1977). While people are able to state the *outcomes* of judgment processes, they have very little ability to introspectively identify the *intervening judgments and inferences* that played a role in producing judgment outputs.

For example, in one study, Nisbett and Wilson presented passersby at a commercial establishment with four *identical* pairs of pantyhose, where the hose were always arranged in a left to right sequence. Participants were then asked which item was of best quality, and, after announcing their choice, they were then asked why they had chosen as they had. There was a pronounced position effect with the rightmost item chosen by a 4:1 margin over the leftmost item. But when asked why they had chosen as they had, only one participant (out of hundreds) spontaneously mentioned the position of the item in the array as influencing her judgment. Most participants cited some other superficially salient attribute of the pantyhose, such as knit strength, elasticity, or sheerness, as being the crucial variable that drove their judgments, despite the fact that all the pantyhose were in fact identical. In another set of studies, Latane and Darley (1970) investigated participants' reactions to a person (actually a confederate in the study) who appeared to be in serious distress. The experimenters systematically manipulated the number of bystanders (also study confederates) present. Results showed highly reliable effects that helping behavior substantially diminished as the number of bystanders increased. But extensive probing revealed that participants were consistently unaware of the effect the number of bystanders was exerting on their choices to help. Instead, participants

routinely cited superficially salient features of the situation to explain why they did not help.

Based on their review of prior studies, Nisbett and Wilson identified two factors that play a central role in why people routinely misidentify which are the critical features of a situation that influence their intuitive judgments: 1) the actual critical features that drive judgment are often *low in salience* and the influences of these features on judgment processes are *not easily accessible to conscious awareness*; and 2) there is some *other* feature of the situation that is *high in salience* and that is *more readily accessible to awareness*. When these two conditions are present, people tend to 'Tell more than they can know' about the causes of their intuitive judgments. That is, they ignore the importance of the low salience feature, and instead identify the high salience feature as the cause of their intuitive judgment. We believe that both these '*conditions for misattribution*' are present in the Chairman case, and may help explain why philosophers have focused on normative factors in explanations for asymmetric intentionality judgments in the Chairman case.

Deep Self variables may be underemphasized in explanations of the Chairman asymmetry because these variables are relatively low in salience, and a large literature in social psychology about spontaneous trait inferences (STIs) supports this claim (see Uleman *et al.*, 2008 for a review). STIs are attributions of attitudes, values, and traits that occur *implicitly* during the course of observing others' behavior. Social psychologists have shown that STIs occur *ubiquitously*, whenever trait-relevant information about others' behavior is presented, *automatically*, without the need for conscious initiation or control, and *implicitly*, with minimal conscious awareness or accessibility. We believe it is plausible that in studies of the Chairman case, when participants are presented with written descriptions of the Chairman's behavior, they automatically make STIs directed at the Chairman about his values and attitudes towards the environment and his propensity to bring about type-similar outcomes in the future. However, since these STIs proceed implicitly, the fact that they are making these inferences is not very salient to participants.

In contrast, normative variables may be overemphasized in explanations of the Chairman Asymmetry because these variables are very high in salience. It is noteworthy that there is one respect in which both the Good/Bad Model and Moral Status Model are absolutely correct: these models are right that people do in fact make profoundly different normative judgments in the two versions of the Chairman case. Indeed, one the strongest effects between measured variables we detected in our entire study is the influence between the case manipulation and Goodness/Badness Judgments (Figure 1 and Figure 3). Moreover, these normative judgments are directed towards a Chairman who expresses contempt for the environment. Thus it is highly plausible that these normative judgments are associated with strong affect and high levels of arousal, making them highly salient and readily accessible to conscious awareness.

We can sum up the conclusions of this section of the paper in terms of a distinction between two different kinds of intuitive judgments. People not only

have first-order *content intuitions* about the Chairman case (i.e. intuitions about whether the Chairman brought about the outcome intentionally in the Harm and Help conditions), they also form second-order intuitive judgments about what features of the cases their (first-order) intuitions track. Moreover, the results of our second study suggest these second-order intuitive '*tracking judgments*' are highly inaccurate. People *think* their intuitions are tracking normative differences in the two versions of the case, but their intuitions *actually* track Deep Self variables. We propose inaccuracies in tracking judgments in the Chairman case arise because of a more general phenomenon: people have only limited abilities to introspectively identify the sources of their intuitions and they are highly susceptible to the biasing effects of salience and accessibility. In short, when asked to identify what features of cases their intuitions track, people often 'Tell more than they can know'.

3. The 'Critical Features Problem' and the Role of Structural Modeling Methods in Philosophical Theorizing

In the previous section, we argued that people's tracking judgments in the Chairman case are highly inaccurate. But difficulties in figuring out what features of a case one's intuitions track is by no means restricted to the Chairman case, but rather are *routinely* confronted during the course of theorizing in other philosophical domains. Let us use the term '*critical features problem*' to refer to the problem of figuring out which among the myriad features present in hypothetical cases are the critical ones that our intuitions are responsive to. In this final part of the paper, we argue that, structural path modeling (and related statistical techniques which we refer to collectively as '*structural modeling methods*') may be particularly useful as a tool for philosophers in addressing the critical features problem. In particular, we argue that structural modeling methods have significant advantages over both '*armchair*' *methods* commonly used by traditional philosophers, as well as *basic survey methods* commonly used by experimental philosophers.

Let us begin by getting a deeper sense of what the critical features problem is and why it represents an important problem for philosophical inquiry. It is widely accepted that intuitions about hypothetical cases play a central role in philosophical theory construction. But it is less well appreciated that intuitions about hypothetical cases, *by themselves*, are often not very helpful as they are only the *rawest* form of data. In order for intuitions about hypothetical cases to be useful as data or inputs into philosophical theorizing, the theorist must often clarify, *among the myriad features present in hypothetical cases, which are the critical ones that are the sources of her intuitive reactions*.

We can readily see this point if we reflect for a moment on well known thought experiments in the history of philosophy. Consider for example Edmund Gettier's famous counter-example to the 'justified, true belief' analysis of knowledge (Gettier,

1963).[21] Gettier described a hypothetical case in which a man named Smith believes a certain proposition *p*, and the belief appears to be fully justified. In addition, *p* is true, but in the circumstances described by Gettier, *p*'s being true arises almost entirely by sheer coincidence. So even though Smith's belief that *p* is justified and true, we intuitively feel that Smith does not *know* that *p*. Gettier's point in presenting this case is not to simply notice and catalog the contents of our intuitions in this *individual* case, and then stop there. Rather, Gettier uses this case to advance a much more *general* thesis about what our intuitions about knowledge do or do not track. In particular, Gettier advances the thesis that our intuitions about knowledge cannot simply be tracking the presence of *justification*, *truth*, and *belief*, as these features are in fact present in the case. Therefore it follows that there must be some *other* critical feature(s) that our intuitions are tracking. The pattern of reasoning used in Gettier's counter-example is by no means unique. That is, it is a widespread and well-accepted practice in philosophy to use intuitions about hypothetical cases as a basis to infer more general theses about which are the critical features our intuitions do or do not track. Thus the critical features problem is an important problem in philosophy, and methods that help philosophers solve this problem can contribute to success in philosophical inquiry.

There are some cases in which it may be relatively easy to identify which are the critical features of hypothetical cases our intuitions are responsive to. But there are also some important cases that are much like the Chairman case where the critical features our intuitions track are more obscure and difficult to recognize. For example, Peter Singer (Singer, 1972) famously observed that if we encounter a small child about to drown in a shallow pond, we intuitively feel that we are morally obligated to rescue the endangered child. However, when many, many more small children are imperiled in a famine in Bangladesh, we (that is, we who are observing the famine from a distance) do not intuitively feel that we are morally obligated to rescue these endangered Bangladeshi children, or at least, we feel the pull of this intuition much more weakly. Here, it is not at all obvious why our intuitive reactions diverge in the two cases, so much so that sustained efforts have been devoted by other philosophers in chasing down what is the crucial difference (for example, see Unger, 1996).

The standard 'armchair' method that philosophers utilize to solve the critical features problem is the '*the method of dissociation*'. In order to identify whether our intuitions track some feature F versus some feature G, one can systematically construct hypothetical cases that dissociate these two features. That is, one can construct cases in which F is present but G is not, and cases in

[21] There is a sophisticated philosophical literature on the proper way to formalize the argumentative structure of Gettier's thought experiment (Williamson, 2007; Pust, Unpublished Manuscript). Our discussion is compatible with a number of different formalization strategies, and we use an intuitive characterization of the structure of Gettier's argument in what follows.

which G is present while F is not. By checking one's intuitive reactions to this overall set of cases, it can often be discerned whether our intuitions track F versus G.

While the method of dissociation is often effective, and it is certainly deployed adeptly by philosophers, the method is not infallible and indeed may have serious limitations in some cases. In particular, the method of dissociation is conspicuously less effective when two or more features are *highly correlated* across a range of hypothetical cases. In this circumstance, it may be difficult to construct *plausible* hypothetical cases in which these correlated features are truly dissociated. For example, it may hard to find plausible cases that *cleanly* and *completely* pull apart the notion of rightness associated with deontological theories versus utilitarian theories (especially rule utilitarian theories), even though the features these theories claim our intuitions track are indeed quite different. This problem of *correlated features* is compounded by the issue of *salience*. Recall from our discussion of the psychological literature on subjective reports on mental processes, the underlying psychological processes that cause our intuitions to track particular features of hypothetical cases are often implicit, and the features these intuitions track may be of low salience. Furthermore, other features of hypothetical cases may be much higher in salience. When correlated features of hypothetical cases differ in salience, people's intuitive judgments about what features of cases their intuitions track may become highly inaccurate. That is, there may be a very strong tendency to 'Tell more than we can know' and erroneously identify the more salient features as being the ones that drive intuitions. For these reasons, the method of dissociation is not always effective and there is a need to develop additional approaches for solving the critical features problem that go beyond what armchair methods can deliver.

Experimental philosophers frequently criticize armchair methods of philosophical theorizing as being ineffective or unreliable. But with regard to the critical features problem specifically, it is not at all clear how typical methods used by experimental philosophers represent much of an improvement over armchair methods. The vast majority of experimental studies by philosophers investigating the Chairman case (though by no means all!) have used *basic survey methods* that count and compare the number of people who make one kind of intuitive judgment versus another. Surveys such as these can clearly be useful for some purposes (for example, to document that there is variability within or across cultures in philosophically-relevant intuitions). But in cases such as the Chairman case where our primary interest is in finding out the sources or causes of our intuitions, basic survey methods are less helpful. These surveys can only document that asymmetric judgments of intentionality are observed, but they cannot explain what causes this asymmetry. If the results of a survey show that more people think the Chairman intentionally harmed the environment in the harm condition than helped the environment in the help condition, the experimenter still has to figure out what features of the case generated this asymmetric pattern of intuitions. So the experimenter is still confronted with the problem that it is quite hard to introspectively identify the sources of one's intuitions, and quite easy to fall prey to biasing effects of factors

with high salience and high accessibility. Thus, the use of basic survey methods appears to offer little help in solving the critical features problem.

In contrast to basic survey methods, which do not appear to add much to standard armchair methods as a means to address the critical features problem, structural modeling methods of the kind utilized in our study of the Chairman case *do* deliver important information that is not accessible from the armchair. Structural modeling methods work by taking advantage of natural patterns of variation in intuitive judgments across a very large number (typically hundreds) of people in order to arrive at conclusions that are not easily accessible to any one of those persons taken alone. For example, in the Chairman case, suppose one person makes an intuitive judgment about whether the Chairman intentionally brought about the outcome, as well as a normative judgment about the goodness or badness of this outcome. That lone person would be very hard-pressed to say with any confidence whether the judgment of intentionality is being causally driven by the goodness/badness judgment, or whether these two judgments simply co-occur without the former being directly causally influenced by the latter. But if we assemble hundreds of people together, the aggregate will exhibit patterns of variation with respect to each of these judgments. That is, people will differ on how strongly they feel that that the Chairman intentionally brought about the outcome, and people will differ on how good or bad they rate the outcome brought about by the Chairman. By analyzing these naturally occurring patterns of variation across hundreds of people with structural modeling, we can often gain valuable evidence for whether one judgment is causally driving the other or whether the two judgments merely co-occur. Based on this kind of evidence, we can form more reliable assessments of whether intuitions about intentionality track the goodness or badness of outcomes, or whether intuitions about intentionality track other features (such as Deep Self-related features) that are often correlated with the goodness or badness of outcomes. Thus quantitative methods such as structural modeling provide a powerful means for addressing the critical features problem, a means that is quite distinct from, and in some ways goes significantly beyond, what can be accomplished from the armchair.

One of the most common complaints heard about the use of experimental methods in philosophy starts with the observation that experimental methods study the intuitions of *ordinary people*. Then, the question is posed, 'Why should we care about the intuitions of ordinary folk? Since philosophers seek to construct *reflective* theories in which ordinary opinions are critically scrutinized and frequently revised, why will simply polling the opinions of ordinary folk contribute much to philosophical inquiry?' This '*Who cares about the folk?*' objection may have some bite when directed at the *basic survey methods* discussed earlier that seek to count and compare the number of people that have one intuition versus another (though how much bite it actually carries is open for debate). But the 'Who cares about the folk' objection wholly misses the mark when directed against structural modeling methods. The point of structural modeling is not to simply catalog the *contents* of ordinary people's intuitions, or to quantify the number of people who have one intuition versus another. Rather, the point is to *illuminate underlying*

causal relationships between various implicit judgment processes that play a role in producing philosophically-relevant intuitions. By helping us to identify which among a number of potential factors do or do not contribute to the formation of intuitive judgments, these experimental methods help solve the problem of figuring out what features of hypothetical cases our intuitions track. It is precisely when we are aware of the features our intuitions track that we are able to reflectively criticize whether these intuitions are warranted, and whether these intuitions should carry weight in a mature philosophical account. Thus quantitative methods such as structural modeling do not seek to substitute the reflective opinions of philosophers with the untutored opinions of laypeople. Rather, by illuminating underlying tracking relationships, these quantitative methods serve as a tool for advancing philosophers' own reflective agendas.

4. Conclusion

In this paper, we challenged a widely held view among philosophers that normative considerations influence judgments of intentionality. We conducted a study of people's intuitions in the Chairman case, one of the most influential in the philosophical literature, using a statistical method called *structural path modeling*. The results of our study showed that people's attributions of intentional action to an agent are not primarily driven by their normative assessments, but are instead mainly driven by assessments of the underlying values and characterological dispositions of the agent, as predicted by the alternative Deep Self Concordance Model. In a second study, we examined people's judgments about what they *think* drives asymmetric intuitions in the Chairman case and found that people are highly inaccurate in identifying which features of the case their intuitions track. Invoking the work of Richard Nisbett and Timothy Wilson from their famous paper 'Telling more than we can know: Verbal reports on mental processes', we argued that people's tracking judgments are inaccurate because these judgments are often inappropriately biased by factors such as salience and accessibility. We then discussed the ways in which structural path modeling can supplement existing methods used by philosophers, including existing methods used by experimental philosophers, in helping to solve the critical features problem—the problem of figuring out which among the myriad features present in hypothetical cases are the critical ones our intuitions are responsive to. Thus, structural path modeling and related techniques may serve as an important new *additional* tool for philosophers that can complement existing methods of philosophical inquiry.

C. Sripada, Department of Philosophy, University of Michigan, Ann Arbor and Department of Psychiatry, University of Michigan, Ann Arbor

S. Konrath, Institute for Social Research, University of Michigan, Ann Arbor and Department of Psychiatry, University of Rochester, Rochester, New York

Appendix A: Testing Nadelhoffer's 'Interaction Model'

In a series of publications (Nadelhoffer, 2004a and b, 2006), Thomas Nadelhoffer has put forward an intriguing model for explaining asymmetries in intentionality judgments in cases such as the Chairman case. While we did not have Nadelhoffer's model in mind when we designed the present study, the data we collected did allow us to perform an additional test of his model, and we report the results in this appendix section. Nadelhoffer's model of intentionality judgments incorporates elements of the Moral Status Model and the Good/Bad Model (and an important feature of the Deep Self Concordance Model as well). In agreement with the Moral Status Model, Nadelhoffer believes that seeing an agent behave in an immoral way biases people's judgments of intentionality. But unlike the Moral Status Model, Nadelhoffer contends that the Chairman asymmetry doesn't arise because people view the Chairman as more immoral in the harm condition and less immoral in the help condition (though he doesn't explicitly deny that this may be the case). Rather, he focuses on the observation that people will tend to view the Chairman as more immoral than moral in *both* the harm and the help condition. This is because the Chairman expresses contempt and hostility towards the environment both when he actually harms it (in the harm condition), and *also* when he says he doesn't care at all about helping it (in the help condition). Additionally, Nadelhoffer puts forward the idea that moral status judgments of the agent *interact* with the evaluative valence of the outcome that the agent brings about, which is why we dub his model the 'Interaction Model'. According to this model, if an agent is judged to be immoral, then people are biased to judge that the *bad* outcomes the agent brings about are intentional, but the *good* outcomes the agent brings about are not intentional. The Interaction Model thus contains the idea that intentionality judgments depend, at least in part, on the presence of concordance between the moral valence assigned to the agent and the evaluative valence assigned to the outcome, and in this respect this model has some similarities with the Deep Self Concordance Model.

To test the Interaction Model, we computed an additional variable that represents the magnitude of the difference (i.e. the absolute value of the difference) between each participant's ratings of the Moral Status of the Chairman (Question 3) and the Goodness/Badness of the outcome (Question 2), and we refer to this additional variable as the 'MSGB' variable. The MSGB variable represents the concordance between the moral valence assigned to the agent and the evaluative valence assigned to the outcome. The Interaction Model predicts that the MSGB variable will take lower values in the harm condition than the help condition (there is greater concordance between the evaluative valence of the agent and the outcome in the harm condition than the help condition because in the harm condition, typically, the Chairman is judged to be very immoral and the outcome is judged to be very bad), and that this variable will predict intentionally judgments. One way to test the Interaction Model is simply to place the MSGB variable in the original structural path model presented in Figure 1 alongside our other candidate mediator variables.

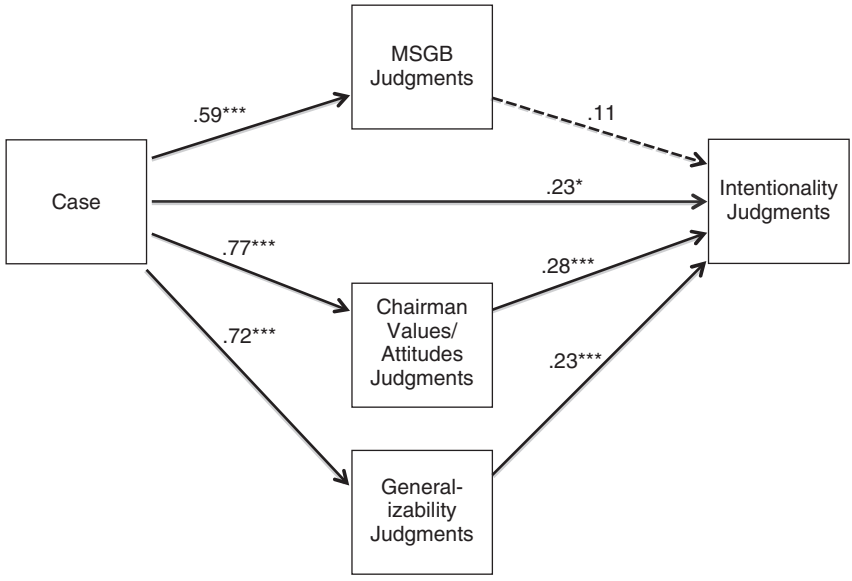

Figure 4 *Structural path model of candidate mediators of the relationship between the case manipulation and judgments of intentional action in the Chairman case. Statistically significant paths are shown as solid arrows, while non-significant paths are shown as dashed arrows.* * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$

However, since the MSGB variable is directly derived from the Moral Status Judgments and Goodness/Badness Judgments variables, putting all three normative variables in the same model results in unacceptably high levels of intercorrelations between the predictor variables in model (i.e. the problem of 'multicollinearity'). To avoid this problem, we entered the MSGB variable in a new model from which the Goodness/Badness Judgments and Moral Status Judgments variables were omitted, and performed a new structural path analysis of the relationships between the variables in this new model.

This new structural path model (Figure 4) exhibits 'acceptable' fit with the data,[22] and modification indices (see footnotes 11 and 12) do not identify additional paths for which there is statistical evidence that these paths should be added or removed from the model. In this new model, Case is a highly significant predictor of MSGB Judgments, but the *MSGB Judgments → Intentionality Judgments* path reaches only trend level statistical significance ($p = 0.065$). Furthermore, the model shows that MSBG Judgments are only a weak predictor of Intentionality Judgments.

[22] $X^2(3, N = 240) = 9.54$, $p = 0.02$; NFI = .986; NNFI = .967; CFI = .990; RMSEA = .095. Using 'rule of thumb' cutoff values for the RMSEA (Hu and Bentler, 1999), the model displays 'acceptable' overall fit with the data.

That is, the size of the path connecting these two variables is relatively small compared to the paths connecting Deep Self variables to Intentionality Judgments. Mediation analysis shows that 9% of the effect of Case on Intentionality Judgments is mediated through MSBG Judgments. We conclude from this additional analysis that the normative factor identified in Nadelhoffer's model may indeed contribute to Intentionality Judgments, but the size of the effect of this factor is likely to be fairly small. Moreover, our overall conclusion from Part I of the paper—that Deep Self variables explain the majority of the asymmetry effect in the Chairman case and normative variables explain a much smaller portion of the effect—remains fully supported.

Appendix B

The covariance matrix analyzed in this study is printed below:

	Case	Goodness/ Badness	Moral Status	Chairman Values/ Attitudes	Generalizability	Intentionality
Case	0.251					
Goodness/ Badness	0.843	4.877				
Moral Status	0.13	0.493	1.422			
Chairman Values/ Attitudes	0.732	2.74	0.389	3.581		
Generalizability	−0.707	−2.542	−0.35	−2.275	3.814	
Intentionality	−0.703	−2.381	−0.456	−2.572	2.51	4.46

References

Alicke, M. D. 1992: Culpable causation. *Journal of Personality and Social Psychology*, 63, 368–78.

Alicke, M. D. 2008: Blaming badly. *Journal of Cognition and Culture*, 1–2, 179–86.

Baron, R. and Kenny, D. 1986: The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–82.

Beebe, J. R. and Buckwalter, W. 2010: The epistemic side-effect effect. *Mind & Language*, 25, 474–98.

Bentler, P. 1995: *EQS Structural Equation Program Manual*. Los Angeles: BMDP Statistical Software.

Doris, J. M., Knobe, J. and Woolfolk, R. L. 2007: Variantism about responsibility. *Philosophical Perspectives*, 21, 183–214.

Gettier, E. L. 1963: Is justified true belief knowledge? *Analysis*, 23, 121–3.

Holton, R. 2010: Norms and the Knobe effect. *Analysis*, 70, 417–24.

Hu, L-T. and Bentler, P. 1999: Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.

Knobe, J. 2003: Intentional action and side effects in ordinary language. *Analysis*, 63, 190–3.

Knobe, J. 2005: Theory of mind and moral cognition: exploring the connections. *Trends in Cognitive Science*, 9, 357–9.

Knobe, J. 2006: The concept of intentional action: a case study in the uses of folk psychology. *Philosophical Studies*, 130, 203–31.

Latane, B. and Darley, J. 1970: *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century-Crofts.

Mele, A. 2006: The folk concept of intentional action: a commentary. *Journal of Cognition and Culture*, 6, 277–90.

Nadelhoffer, T. 2004a: Praise, side effects, and intentional action. *Journal of Theoretical and Philosophical Psychology*, 24, 196–213.

Nadelhoffer, T. 2004b: Blame, badness, and intentional action: a reply to Knobe and Mendlow. *The Journal of Theoretical and Philosophical Psychology*, 24, 259–69.

Nadelhoffer, T. 2006: Bad acts, blameworthy agents, and intentional actions: some problems for juror impartiality. *Philosophical Explorations*, 9, 203–19.

Nisbett, R. and Wilson, T. 1977: Telling more than we can know: verbal reports on mental processes. *Psychological Review*, 84, 231–59.

Pettit, D. and Knobe, J. 2009: The pervasive impact of moral judgment. *Mind & Language*, 24, 586–604.

Pust, J. Unpublished Manuscript: Intuitions.

Raykov, T., Tomer, A. and Nesselroade, J. R. 1991: Reporting structural equation modeling results in psychology and aging: some proposed guidelines. *Psychology and Aging*, 6, 499–503.

Shrout, P. and Bolger, N. 2002: Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods*, 7, 422–45.

Singer, P. 1972: Famine, affluence, and morality. *Philosophy and Public Affairs*, 229–43.

Sinnott-Armstrong, W., Mallon, R., Mccoy, T. and Hull, J. G. 2008: Intention, temporal order, and moral judgments. *Mind & Language*, 23, 90–106.

Sripada, C. S. 2010: The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, 151, 159–76.

Uleman, J., Saribay, S. A. and Gonzalez, C. 2008: Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, 59, 329–60.

Unger, P. 1996: *Living High and Letting Die*. New York: Oxford University Press.

Williamson, T. 2007: *The Philosophy of Philosophy*. New York: Routledge.

Wright, J. C. and Bengson, J. 2009: Asymmetries in judgments of responsibility and intentional action. *Mind & Language*, 24, 24–50.