Deep Trouble for the Deep Self[1]

[Forthcoming in *Philosophical Psychology*]

David Rose, Jonathan Livengood, Justin Sytsma and Edouard Machery

Abstract: Chandra Sripada's (forthcoming) Deep Self Concordance Account aims to explain various asymmetries in people's judgments of intentional action. On this account, people distinguish between an agent's active and deep self; attitude attributions to the agent's deep self are then presumed to play a causal role in people's intentionality ascriptions. Two judgments are supposed to play a role in these attributions—a judgment that specifies the attitude at issue and one that indicates that the attitude is robust (Sripada and Konrath, forthcoming). In this article, we show that the Deep Self Concordance Account, as it is currently articulated, is unacceptable.

The folk concept of intentional action has been the subject of extensive research by experimental philosophers and psychologists (Alicke, 2008; Knobe, 2003a, 2003b, 2006; Machery, 2008; Malle, 2006; Mele, 2006; Nadelhoffer, 2004, 2006; Nichols & Ulatowski, 2007; Wright & Bengson, 2009). This research has focused primarily on puzzling asymmetries in ordinary people's judgments about intentional action. Researchers have been particularly concerned with ordinary judgments about the intentional status of side effects, as in Knobe's (2003a) harm and help cases. Consider first Knobe's harm case:

> 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

[1] We would like to thank Josh Knobe, Shaun Nichols, Chandra Sripada and Liane Young for their helpful comments. We would also like to thank David Danks for his very helpful discussions.

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.

When experimental participants were presented with this case and asked to rate their agreement with the claim that the chairman intentionally harmed the environment, 82% *agreed* that the chairman intentionally harmed the environment. Consider now Knobe's help case:

> 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, and it will also help the environment." The chairman of the board answered, "I don't care at all about helping 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 helped.

Notice that, in this case, everything is the same as in the harm case except the outcome: in the help case, the environment is helped as a result of starting the program. Strikingly, when participants were asked to rate their agreement with the claim that the chairman intentionally helped the environment, 77% *disagreed* that the chairman intentionally helped the environment. Ordinary judgments in Knobe's harm and help cases, thus, reveal a striking asymmetry: a negative foreseen side effect of a chairman's action—harming the environment—is judged to have been brought about *intentionally*, while a positive foreseen side effect—helping the environment—is judged to have been brought about *unintentionally*. Call any puzzling asymmetry in ordinary judgments about intentional action (whether or not it involves side effects) an *intentionality judgment asymmetry*.

Debate has turned on whether intentionality judgment asymmetries are best explained in terms of the influence of some type of normative judgment (Alicke 2008; Knobe 2003a, 2004,

2006; Mele 2006; Mele and Cushman, 2007; Nadelhoffer 2004, 2006; Pettit and Knobe 2009; Uttich and Lombrozo, 2010; Wright and Bengson, 2009) or rather in terms of the interplay between various descriptive judgments (Guglielmo, Monroe and Malle 2009; Machery 2008; Malle 2006; Nanay 2010; Scaife and Weber forthcoming; Sripada forthcoming). We will refer to explanations of the former type as *prescriptivist accounts* and explanations of the latter type as *descriptivist accounts*.

Our concern in this paper is with one descriptivist account in particular, the Deep Self Concordance Account put forward by Chandra Sripada (Sripada forthcoming; Sripada and Konrath forthcoming).[2] Sripada claims that, in reading stories like those used in Knobe's harm and help cases, people intuitively form opinions about the agent's "deep self," which is a part of the agent's psychology containing her "stable and central psychological attitudes, including the agent's values, principles, life goals, and other more fundamental attitudes" (Sripada forthcoming, p. 18). Sripada then argues that people's judgments about whether an agent intentionally brought about an outcome depend on whether that outcome is concordant with their sense of her deep self. As such, descriptive judgments about the agent's deep self are thought to play a causal role in people's intentionality judgments, while normative judgments, characteristic of prescriptivist accounts, are thought to play no causal role.

In this article, we argue that the current empirical data undermine the Deep Self Concordance Account, as it has been articulated to date. Surprisingly, the data we appeal to are produced by Sripada himself in a follow-up article with Sara Konrath (Sripada and Konrath

[2] Sripada (forthcoming) refers to this account as the "Deep Self Model," while Sripada and Konrath (forthcoming) call it the "Deep Self Concordance Model." In order to avoid some confusion in what follows, however, we will reserve the term "account" for psychological theories and the term "model" for a specific structural equation instantiation of a theory.

forthcoming). Although Sripada and Konrath naturally take their data to support the Deep Self Concordance Account, we contend that they are mistaken.

Here is how we will proceed. In Section 1, we further describe Sripada's Deep Self Concordance Account, arguing that as Sripada has articulated it to date it entails two positive causal hypotheses. In Section 2, we describe Sripada and Konrath's reasons for believing that their data support these hypotheses. Then, in Sections 3 through 5, we argue that their analysis of the data is mistaken, presenting three objections to their analysis and arguing that the data actually undermine the positive causal hypotheses noted in Section 1.

1. The Deep Self Concordance Account

1.1 Sripada's Account

Most philosophers and psychologists who have written about intentionality judgment asymmetries have concluded that people's normative judgments influence their judgments about intentional action. In contrast to such prescriptivist accounts, Sripada (forthcoming) and Sripada and Konrath (forthcoming) advocate a descriptivist account, the Deep Self Concordance Account, that does not call on normative judgments in explaining these asymmetries. This account is premised on people intuitively distinguishing between an agent's active and deep self. Sripada (forthcoming, p.7) writes:

> According to the [Deep Self Concordance Account], people utilize a naive theory of the structure and contents of the mind and this theory guides judgments about intentionality. The key feature of this theory is that it posits that behind the agent's Acting Self, i.e., the narrow set of outcome-directed proximal desires, means-end beliefs, and intentions, that are the immediate causal source of the action, lies a much larger set of more stable,

enduring and fundamental attitudes. These attitudes collectively constitute the agent's Deep Self.

Sripada then goes on to argue that "in making judgments about intentionality, subjects are, inter alia, assessing the concordance between the outcomes an agent brings about and the relatively deep and enduring parts of the agent's underlying psychology, and this concept [i.e., intentionality] is applied only when such concordance obtains" (p. 8). Applied to Knobe's harm and help cases, then, the Deep Self Concordance Account predicts that, if a participant judges the chairman to have anti-environment attitudes, and judges those attitudes to be *robust* (in the sense that they are stable and enduring), then the participant will believe that the chairman's action concords with her Deep Self in the harm condition but not in the help condition. Consequently, participants will be likely to say that the chairman intentionally harmed the environment (since the outcome concords with her robust attitudes) but unlikely to say that the chairman intentionally helped the environment (since the outcome does not concord with her robust attitudes).

Sripada's account appears to be a clear and straightforward descriptivist account of intentionality judgment asymmetries: An agent is more likely to be said to have performed an action intentionally when that action concords with her "stable and central psychological attitudes" (Sripada forthcoming, p. 18). But is this account correct? To answer this question, we need to do two things: First, we need to clearly articulate what the hypothesized attribution of attitudes to the deep self consists in for at least one prominent asymmetry in the literature (we will focus on Knobe's help and harm cases); second, we need empirical data that indicate whether or not, in the case of this asymmetry, those supposed attributions cause people's intentionality judgments. We examine these two issues in turn in the remainder of this section.

1.2 The Two Positive Causal Hypotheses of the Deep Self Concordance Account

Although Sripada's (forthcoming) presentation of the Deep Self Concordance Account is clear and straightforward, it is also somewhat underspecified. Particularly, it is unclear what the attribution of attitudes to an agent's deep self consists in. What judgments do people make when they ascribe an attitude to, e.g., the chairman's deep self? Fortunately, Sripada and Konrath's (forthcoming) articulation of the Deep Self Concordance Account largely remedies this problem. Sripada and Konrath propose that the attribution of attitudes to someone's deep self involves making *two* distinct judgments. First, people are hypothesized to attribute some specific attitudes to some other individual. Second, people are supposed to judge that these attitudes are robust, in the sense given above. A robust attitude would lead the individual to act in the same particular way across various situations. To illustrate, when considering Knobe's harm case, people are first hypothesized to attribute an anti-environmental attitude, such as the view that the environment is not worth helping, to the chairman. Second, people are also supposed to judge that this anti-environmental attitude is robust, that is, that it would lead the chairman to harm the environment in a variety of situations.

Thus, the Deep Self Concordance Account, as it is currently articulated, makes two positive causal hypotheses about people's judgment of intentionality in the harm and help cases: First, it hypothesizes that people's attribution of a pro- or anti-attitude towards the environment to the chairman causally influences their judgments about the intentional nature of her action. Call this the *first positive causal hypothesis* of the Deep Self Concordance Account. Second, it hypothesizes that people's judgment concerning how likely the chairman is to harm or help the environment in other circumstances causally influences their judgments about the intentional

nature of her action. Call this the *second positive causal hypothesis* of the Deep Self Concordance Account.

1.3 Assessing the Positive Causal Hypotheses of the Deep Self Concordance Account

As noted above, in addition to specifying the Deep Self Concordance Account, we need data to assess the causal hypotheses made by this account. Sripada (forthcoming) reports some data showing that people tend to judge that the chairman has anti-environmental attitudes (regardless of whether they are given the help or the harm cases), and concludes that they provide support for the causal claims made by the Deep Self Concordance Account. We demur. The two hypotheses made by the Deep Self Concordance Account (as it is currently articulated) are causal: Sripada appears to hold both that the ascription of attitudes to others causally affects intentionality judgments and that judgments about the robustness of these attitudes causally affect intentionality judgments. Unfortunately, the data reported by Sripada are silent about these hypotheses; after all, it is certainly possible that people judge that the chairman has anti-environmental attitudes and yet that these attributions do not play a causal role in their intentionality judgments.

Sripada and Konrath's (forthcoming) work is a notable effort to provide some evidence that bears directly on the two positive causal hypotheses of interest. They asked participants several questions related to what attitude the chairman has, the robustness of his attitudes, etc. They then employed the quantitative method of *structural path analysis* to determine the causal relations between these variables.[3] We now show that, far from supporting Sripada's Deep Self

[3] For Sripada and Konrath's models to count as *causal models* (or structural models), a number of strong assumptions need to be satisfied. Specifically, the substantive variables need to be linearly related and free of measurement error (since their models are single-indicator path models). The errors on exogenous variables have to be uncorrelated with each other and also uncorrelated with the errors on the endogenous variables. The errors for

Concordance Account (as it is currently articulated), these new data provide evidence against the positive causal hypotheses.

2. Sripada and Konrath's Causal Models

Sripada and Konrath set out to simultaneously test Sripada's Deep Self Concordance Account as well as several prominent prescriptivist accounts from the literature. Like the Deep Self Concordance Account, these prescriptivist accounts entail some specific causal hypotheses regarding intentionality judgments. For example, Knobe's (2006) Good/Bad Account asserts that people's judgments about whether the outcome of an action is good (or bad) causally influence their judgments about whether the agent brought about that outcome intentionally, and Alicke's (2008) Moral Status Account asserts that people's judgments about whether an agent is a bad person who is worthy of blame for an action causally influence their judgments about whether the agent brought about the outcome of this action intentionally.

In order to test the various accounts of intentionality judgment asymmetries that they consider, Sripada and Konrath presented 240 students at the University of Michigan with Knobe's help and harm cases along with five questions—the original question about whether the chairman intentionally helped or harmed the environment and four additional questions that are assumed to measure candidate explanatory variables. The questions and variable names that we will use are given in Table 1 below.

Variable	Question	Anchors

two endogenous variables can only be correlated if neither substantive variable causes the other. (Usually, however, it is assumed that the whole body of error terms is uncorrelated. Sripada and Konrath appear to make this assumption.) Furthermore, the variables need to measure properties that could actually be causally related in the real world, which is not as trivial an assumption as it might appear. Although we have some doubts as to whether these assumptions actually hold, for the sake of argument we will follow Sripada and Konrath in assuming that they do.

Case	N/A; participants were assigned to a "harm" condition or to a "help" condition.	N/A
Intentional	How much do you agree with the statement 'The chairman intentionally harmed [helped] the environment'?	Strongly Agree, Strongly Disagree
Environment	In your view, how good or bad is the outcome that the environment is harmed [helped]?	Very Good, Very Bad
Moral	In your view, what is the chairman's moral status?	Very Moral, Very Immoral
Attitude	What are the chairman's values and attitudes towards the environment?	Very Pro-environment, Very anti-environment
General	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

Table 1: Sripada and Konrath's questions and corresponding variable labels.

The variable *Environment* is relevant to testing Knobe's Good/Bad Account, while *Moral* is relevant to testing Alicke's Moral Status Account. Against these two prescriptivist accounts, the Deep Self Concordance Account predicts that neither *Environment* nor *Moral* cause *Intentional*. Call these predictions Sripada and Konrath's *negative causal hypotheses*. The variables *Attitude* and *General* are relevant to testing the two positive causal hypotheses discussed in Section 1: If the first positive causal hypothesis is correct, then *Attitude* causes *Intentional*; and, if the second positive causal hypothesis is correct, then *General* causes *Intentional*.

Sripada and Konrath attempted to test all four of these causal hypotheses—both the two negative causal hypotheses and the two positive causal hypotheses—at the same time. To do this, they fit an initial structural equation model to the data they collected.[4] Sripada and

[4] The fit indices reported by Sripada and Konrath (forthcoming) are as follows: $X^2(6, N = 240) = 12.00$, $p=0.06$; NFI = .985; NNFI = .981; CFI = .992; RMSEA =.065 (p. 11).

Konrath's initial model is shown in Figure 1. This model posits that each of the four candidate explanatory variables is caused by *Case* and that each explanatory variable causes *Intentional*.

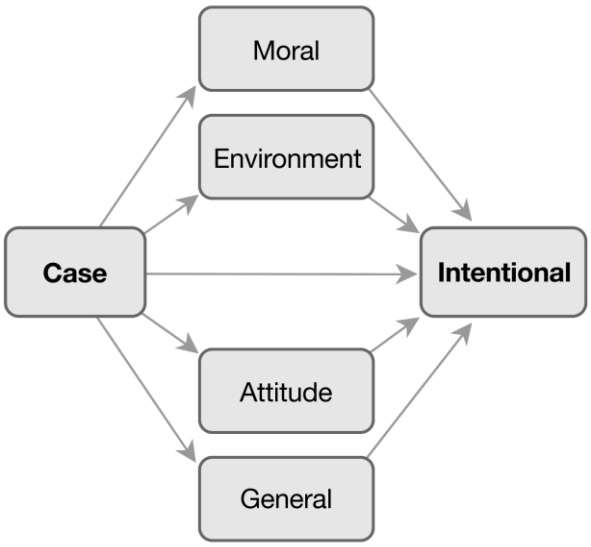

Figure 1: Sripada and Konrath's initial model.

After examining the fit of their initial model, Sripada and Konrath conducted modification tests. A modification test is a statistical test of whether a model's overall fit would improve significantly if an edge were added or removed in the model's corresponding graph. Only models that are hierarchically related can be compared by modification tests.[5] Following some discussion of the results of the significant modification tests for their initial model, Sripada and Konrath settle on two statistically equivalent models that fit the data significantly better than their initial model. Call these *Sripada and Konrath's causal models*. A summary of the usual fit indices for these models is shown in Table 2. Corresponding with the positive and negative causal hypotheses discussed above, each model has both a *positive part* and a *negative part*. The positive part shows that both *Attitude* and *General* cause *Intentional*, while the negative part

[5] Two models are hierarchically related if and only if the graph of one model is a proper sub-graph of the graph of the other model.

shows that neither *Environment* nor *Moral* cause *Intentional*. The models are shown in Figure 2, with the positive part depicted with solid black edges and the negative part depicted with solid gray edges (the two additional edges are depicted with dotted lines).[6]

	Case	***Environment***	***Moral***	***Attitude***	***General***	***Intentional***
Case	0.251					
Environment	0.843	4.877				
Moral	0.13	0.493	1.422			
Attitude	0.732	2.74	0.389	3.581		
General	-0.707	-2.542	-0.35	-2.275	3.814	
Intentional	-0.703	-2.381	-0.456	-2.572	2.51	4.46

Table 2: Covariance matrix for Sripada and Konrath's models.

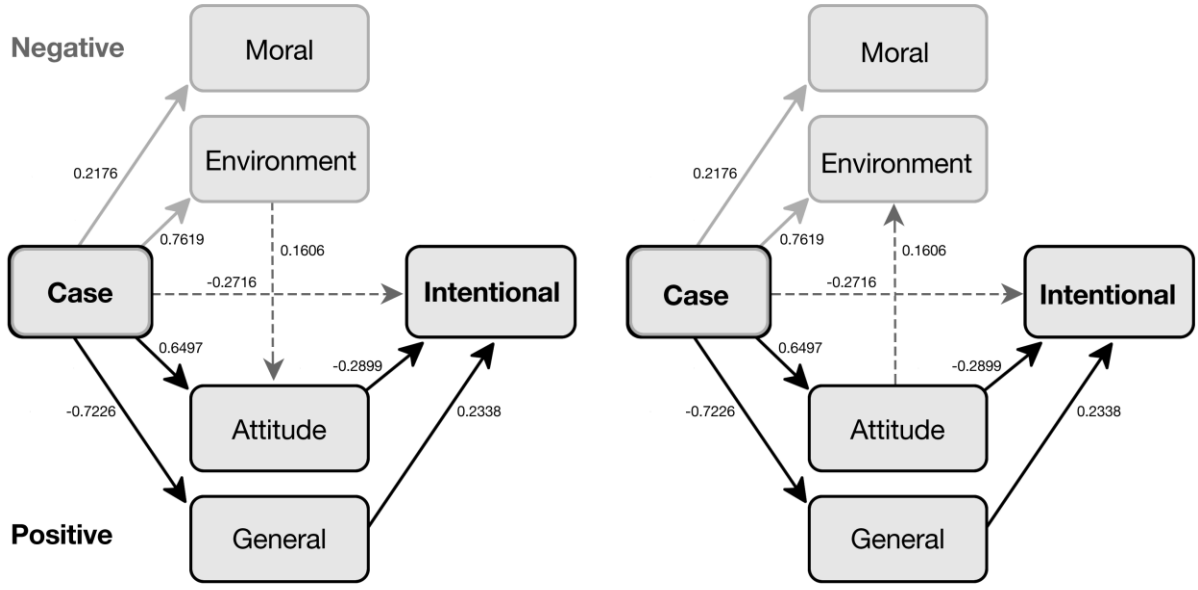

Figure 2: Sripada and Konrath's models with standardized path coefficients.

[6] Each edge in the causal models shown in Figure 2 represents a direct causal connection, and the numbers on each edge are linear coefficients. For example, according to the model in Figure 2, the expected value of *General* given that *Case* takes the value *c* is E(*General* | *Case* = *c*) = -.7226 · *c*. In the case of ordinary regression, the conditional expectation is observational in character, meaning that it tells us what value we can expect *General* to take if we *passively observe* a given value of *Case*. However, Sripada and Konrath want more from their model; they want their model to have *causal* content, meaning roughly that the equations also tell us what to expect given that some variable is *set* to a specified value. For this, they must make the non-trivial assumptions about their data noted in footnote 3.

Sripada and Konrath's models have very good overall fit to their data. Further the models appear to support both the negative and positive causal hypotheses. Thus, Sripada and Konrath take the data to undermine the prescriptivist accounts that they examined and to support the Deep Self Concordance Account. What's not to like? Surprisingly, quite a lot: When the data are analyzed more carefully, what we find is that, far from supporting the positive causal hypotheses of Sripada's Deep Self Concordance Account (as it is currently articulated), the data actually undermine them.

Before turning to the technical details of our alternative analysis of Sripada and Konrath's data, however, it is worth pointing out that the analysis is motivated by two basic concerns about their statistical analysis. First, while Sripada and Konrath performed a search to arrive at their models, they did not perform a full search; thus, while they managed to find two models that both fit their data and support the positive causal hypotheses of the Deep Self Concordance Account, they did not rule out that there are other models that fit their data better and that contradict those hypotheses. In effect, our concern was that Sripada and Konrath cherry picked their models. We explore this concern in Section 3 by conducting a full search to find those models that best fit the data. It turns out that those models are not Sripada and Konrath's models.

Second, Sripada and Konrath attempted to answer two distinct questions via their structural path analysis: whether normative judgments influence intentionality judgments and whether the two judgments relevant to the Deep Concordance Self Account influence them. As a result, the models produced by their modeling work include variables associated with Sripada and Konrath's negative causal hypotheses (*Moral* and *Environment*), as should be evident from

Figure 2. Thus, before accepting the conclusion that their data support the Deep Self Concordance Account, as it is currently formulated, it is important to check whether the data still support the positive causal hypotheses we are concerned with once the variables related to the negative causal hypotheses are removed. We explore this concern in Section 4 by splitting Sripada and Konrath's models into their two component parts and testing the fit of each. It turns out that when we do so the positive sub-model—the sub-model that embodies the positive causal hypotheses of the Deep Self Concordance Account—is not a good fit. This finding is then explored further in Section 5.

3. Trouble from Alternative Models

Sripada and Konrath employ a method typical in social-scientific use of structural equation models: derive a model from one's preferred theory and test whether it fits the data; if the theory-derived model fits, then conduct modification tests in order to examine whether models similar to the theory-derived model fit better; stop at the best-fitting model; if the theory-derived model does not fit, then go back and theorize some more. How likely are we to hit on the model that best fits the data using this guess-and-check method? While that depends, in part, on how good one's preferred theory is, we suspect that in general the likelihood of hitting on the model by conducting modification tests is not good. The reason is that even in relatively small search spaces, the number of possible structural equation models can be quite large. For example, assuming that there is at most a single edge connecting any two variables, there are 3^{15} distinct models over the six variables considered by Sripada and Konrath. Even assuming that *Case* is not caused by anything, there are still $2^5 \cdot 3^{10}$ distinct models. Further restricting attention to

directed acyclic graphs leaves almost a million admissible models (936,992 models to be exact).[7] In search spaces this large, it will often happen (as it does in the case of Sripada and Konrath's models) that multiple competing models have acceptable fit to the data, but only some of these will be found by conducting modification tests. As such, we believe that a more principled approach to search than the one employed by Sripada and Konrath is called for.

To search in a principled way, we used the Greedy Equivalence Search algorithm (GES) in Tetrad IV to identify the best-fitting models consistent with the covariance matrix reported in Table 2 and the constraint that *Case* is not caused by any other variable in the model.[8] GES succeeds in orienting all but one edge in the graph: the edge between *Intentional* and *General*. (That edge is not oriented because the two orientations correspond to statistically equivalent structural equation models.) Call the two equivalent models output by GES the *Tetrad models*. These two models are pictured in Figure 3 (the edges that contradict the positive causal hypotheses of the Deep Self Concordance Account are shown in black).

[7] A *directed graph* is a graph in which every edge has a single arrowhead giving it a direction. A directed *acyclic* graph has no cycles. That is, one cannot begin at a vertex in the graph, move through the graph by following the arrows, and return to the initial vertex. If the directed graph corresponding to a structural equation model is acyclic (and the error terms in the model are uncorrelated), then the model is called *recursive*.

[8] GES searches over equivalence classes of models (graphical *patterns*) by assigning an information score, like AIC or BIC, to each pattern that it considers. Beginning with the completely disconnected or null graph, GES first finds the edge (if there is one) that most improves the score over not adding an edge at all, adds it to the pattern, and applies the edge-orientation rules in Meek (1997). The algorithm iterates this procedure until no additions improve the score. Once no *additions* improve the score, GES considers *deletions*. GES finds the edge (if there is one) that most improves the score over not deleting an edge at all, deletes it from the pattern, and applies Meek's orientation rules. When no further deletions would improve the score, GES stops. Chickering (2002) proved that the GES procedure is pointwise consistent if the true model is recursive and omits no common causes. In other words, if the assumptions Sripada and Konrath make are correct, then GES is guaranteed to find the truth given enough data. Tetrad IV is available for free download at http://www.phil.cmu.edu/projects/tetrad/.

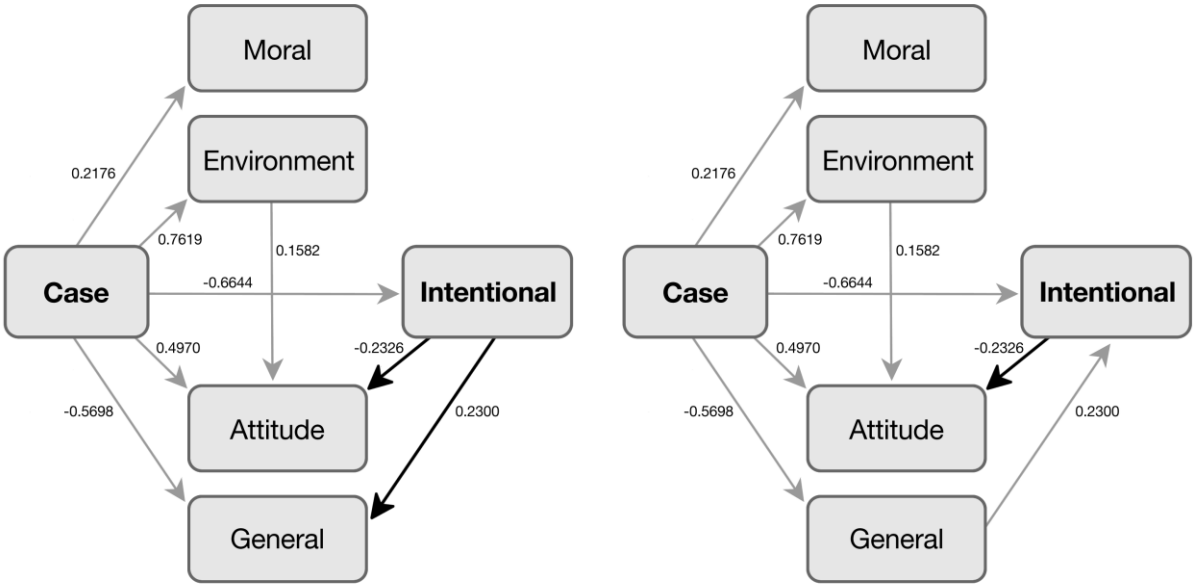

Figure 3: GES output (Tetrad models) with standardized path coefficients.

Consider the model on the left in Figure 3. In this model, the arrows between *Intentional* and *Attitude* and between *Intentional* and *General* are oriented opposite of what the Deep Self Concordance Account predicts. In the model on the right, only one edge, the arrow from *Intentional* to *Attitude*, is opposite of what the Deep Self Concordance Account predicts. Thus, contrary to the first positive causal hypothesis, the models that fit Sripada and Konrath's data best indicate that judgments about whether the chairman acted intentionally cause judgments about the chairman's attitude towards the environment, not the other way round. Furthermore, the models that fit Sripada and Konrath's data best are silent about the second positive causal hypothesis: These models are silent about whether or not people's judgments about the robustness of the chairman's attitudes affect their intentionality judgments.[9]

[9] The reason we say that the GES analysis is "silent" is because the two equivalent Tetrad models show that the causal direction between *General* and *Intentional* could go either way. Tetrad cannot discern between either of these possibilities and so is "silent" on the issue.

The Tetrad models fit the data better than Sripada and Konrath's models, as can be seen in a side-by-side comparison of typical fit indices in Table 3. However, the models are not hierarchically related (i.e. neither model is nested in the other), so the difference in fit cannot be tested for significance. When faced with non-hierarchical models, one typical practice is to choose the model with the best AIC or BIC score (Kaplan 2009; Klein 1998; Loehlin 2004; Rafferty 1995; Raykov and Marcoulides 2000; Rust et al. 1995; Schreiber et al. 2006). Following this practice, we would pick the Tetrad models over Sripada and Konrath's models.

Fit Index	**S&K**	**Tetrad**
Chi-Square (DF)	6.999 (7)	3.6081 (7)
p-value	0.42898	0.82365
Adjusted GFI	0.97154	0.98513
Bentler-Bonnett NFI	0.9912	0.99547
Tucker-Lewis NNFI	1	1.0093
SRMR	0.018832	0.014072
BIC	-31.365	-34.756

Table 3: Typical fit indices for Sripada and Konrath's models and Tetrad models.

The upshot is that the Tetrad models have two distinct advantages over Sripada and Konrath's models. First, they fit the data better. Second, they are the products of a reliable search procedure; that is, a procedure that is guaranteed to find the truth in the large-sample limit if the modeling assumptions made by Sripada and Konrath are satisfied. Since the models

produced by GES are inconsistent with the first positive causal hypothesis of the Deep Self Concordance Account (*Attitude* → *Intentional*) and silent on the second positive causal hypothesis (*General* → *Intentional*), we conclude that Sripada and Konrath's data undermine Sripada's Deep Self Concordance Account, as it is currently formulated.[10]

Nonetheless, Sripada might respond that, although the Tetrad models fit the data *better* than Sripada and Konrath's models, their models are at least *consistent* with the data. After all, their models have admissible fit indices and p-values. As such, Sripada might argue that at the very least the data do not undermine the Deep Self Concordance Account. We have two replies to this response. First, even if we were to grant that Sripada and Konrath's data do not undermine Sripada's account, it would also be the case that they also fail to support it since, plausibly, data that support two incompatible models provide no positive evidence for any of them. Second, two other serious problems (Sections 4 and 5) show that Sripada and Konrath's data actually undermine the Deep Self Concordance Account, as it is currently articulated.

4. Trouble from Model P-Values

Sripada and Konrath's models have good overall fit to their data. However, fit indices for a model (including, but not limited to, the p-value) indicate how well that model fits the data as a whole; they do not indicate how well any particular component of the model fits the data. Thus, a model might have great overall fit, while some of its components or sub-models do not have

[10] Perhaps Sripada would hold that the models output by Tetrad are implausible or even absurd on a priori grounds. He would then conclude that the fact that they fit the data better does not undermine the Deep Self Concordance Account. But it would surely beg the question for Sripada to simply hold that the orientation of the *Intentional—Attitude* and *Intentional—General* edges in the Tetrad models is absurd. Further, stepping away from Sripada's account for a second, it does not seem implausible that people's judgments that the chairman brought about the outcome intentionally or unintentionally would causally affect their judgments about both her attitudes toward the environment as well as their judgments about the likelihood that she would act in ways that are apt to bring about similar outcomes in the future. For example, in the harm condition, it is plausible that having judged that the chairman intentionally harmed the environment, people would then be more likely to judge that she has anti-environment attitudes and that, in the future, she is likely to act in ways that harm the environment.

good or even acceptable fit.[11] If it turned out that Sripada and Konrath's models have good overall fit to their data only because of the parts of their models that embody Sripada's negative causal hypotheses, then their data would not support Sripada's Deep Self Concordance Account, as it is currently articulated.

This is exactly what we found when we investigated the fit of their models further. The easiest way to demonstrate the point is to split Sripada and Konrath's models into two sub-models corresponding with the negative and positive parts of the models indicated in Figure 2; the result is a *positive sub-model* (including the variables *Attitude*, *General*, *Intentional*, and *Case*), which embodies the two positive causal hypotheses of Sripada's current articulation of the Deep Self Causal Account, and a *negative sub-model* (including the variables *Moral*, *Environment*, *Intentional*, and *Case*), which embodies the two negative causal hypotheses. These sub-models are shown in Figure 4.

[11] All structural equation modeling assumes that $\boldsymbol{\Sigma} = \boldsymbol{(\theta)}$, i.e. the true covariance matrix $\boldsymbol{\Sigma}$ is a function of the model parameters $\boldsymbol{\theta}$. The parameters $\boldsymbol{\theta}$ are estimated by minimizing some fitting function (usually the maximum likelihood function). Given parameter estimates, $\hat{\boldsymbol{\theta}}$, the model implies a covariance matrix, $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$. Fit indices measure the distance between the model-implied covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ and the observed covariance matrix, denoted by **S**. Roughly, a model fit index is a function of the sum of either the absolute values of the entries or the squares of the entries in the residual covariance matrix $\mathbf{R} = \mathbf{S} - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$. Thus, a fit index might be acceptable because all of the entries in $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ are acceptably close to **S** or because some of the entries in $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ are extremely close to **S**, even though other entries in $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ are not even acceptably close to **S**. See Bollen (1989, 104 ff. and 256 ff.) for gory details.

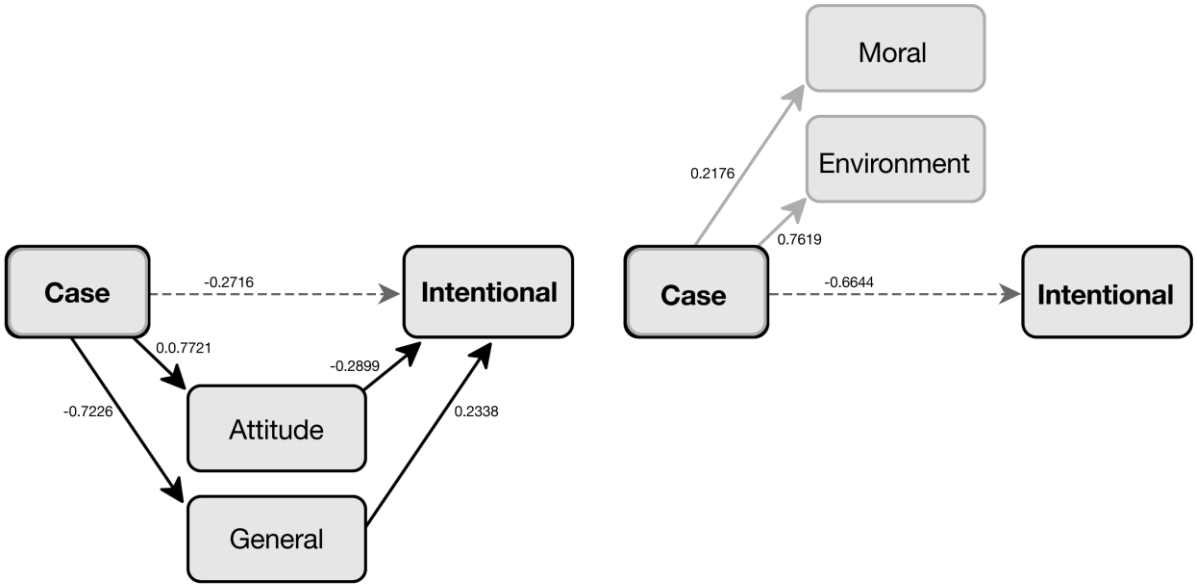

Figure 4: Positive and negative sub-models with standardized path coefficients.

When we test the fit of each sub-model, what we find is that, while the negative sub-model fits the data extremely well, the positive sub-model does not. In fact, the positive sub-model is actually rejected by a chi-square test at the 0.05 significance level.[12] A side-by-side comparison of standard fit indices for the two sub-models is given in Table 4. As the positive causal hypotheses are embodied by the positive sub-model, we conclude that the data undermine those hypotheses and, thus, undermine Sripada's Deep Self Concordance Account, as it is currently articulated.

[12] The model chi-square statistic for the positive sub-model is 4.156 with one degree of freedom (p=0.0415). In comparison, the model chi-square statistic for the negative sub-model is 0.275 with two degrees of freedom (p=0.872).

Fit Index	S&K	Positive	Negative
Chi-Square (DF)	6.999 (7)	4.156 (1)	0.885 (3)
p-value	0.42898	0.0415	0.8289
Adjusted GFI	0.97154	0.91455	0.9938
Bentler-Bonnett NFI	0.9912	0.99268	0.99754
Tucker-Lewis NNFI	1	0.96628	1.0120
SRMR	0.018832	0.019619	0.013447
BIC	-31.365	-1.3252	-15.5565

Table 4: Fit indices for the positive and negative sub-models.

In response, Sripada (personal communication) has argued that it is inappropriate to rely on the chi-square test when the sample size is large. If he is correct, then we cannot reject the positive submodel by means of a chi-square test; as such, we cannot conclude that the current articulations of the Deep Self Concordance Account are undermined by the data.

There are two replies to this response. First, it is at the very best unclear that Sripada and Konrath's sample (N=240) was large enough to make the chi-square test unreliable. In this connection, Paul Barrett remarks (2007, p. 820):

> The χ^2 test is the only statistical test for a SEM model fit to the data. A problem occurs when the sample size is "huge", as stated succinctly by Burnham and Anderson (2002). They note that "model goodness-of-fit" based on statistical tests becomes irrelevant when sample size is huge. ... [However] the numbers being used in examples of "huge" datasets by Burnham and Anderson are of the order of 10,000 cases or more. Not the 200's or so

which seems to be the "trigger" threshold at which many will reject the χ^2 test as being "flawed"!

Other contributors to the special issue of *Personality and Individual Differences* on structural equation modeling in which Barrett's article appeared take issue with many of Barrett's suggestions, but none of them argues that samples in the order of 200 prevent using chi-square tests. If one agrees with Barrett about the range of sample sizes that allow the use of the chi-square test, then one ought to reject the positive sub-model and with it Sripada's Deep Self Concordance Account, as it is currently articulated.

Second, even if one takes a more conservative attitude toward indices of model fit, holding in particular that models with sample sizes greater than 200 prevent using chi-square tests, it is still the case that the positive sub-model is undermined by Sripada and Konrath's data (even if it not rejected in light of the data anymore). For the crucial point is that the very good overall fit of Sripada and Konrath's models is principally due to the negative sub-model, not to the positive sub-model. The positive sub-model itself, which embodies the positive causal hypotheses of the Deep Self Concordance Account, does not fit the data presented by Sripada and Konrath very well, and surely a poor fit is evidence against a model.

5. Trouble from Colliders

Graphical structure is related to conditional independence constraints by the causal Markov and causal Faithfulness conditions. The causal Markov condition entails that for recursive models a variable is independent of its non-effects (its non-descendants) conditional on the set of all of its direct causes (its graphical parents). The causal Faithfulness condition entails that two variables are statistically independent (or conditionally independent) only if that independence (or

conditional independence) is entailed by the causal Markov condition. Roughly, the causal Markov and causal Faithfulness conditions require that statistical associations be explained by causal structure.[13] Though we will not defend them here, the Markov and Faithfulness conditions are very plausible assumptions about the relationship between causation and statistical association. They should not be rejected without strong reasons to think that they fail.

Assuming the Markov and Faithfulness conditions, the positive sub-model discussed in Section 4 entails (i) that *Attitude* is independent of *General* conditional on *Case* and (ii) that *Attitude* is associated with *General* conditional on *Case* and *Intentional*. However, neither (i) nor (ii) is satisfied by the data.[14] Thus, either the edge *General—Intentional* or the edge *Attitude—Intentional* cannot be oriented in the way predicted by the positive causal hypotheses of the Deep Self Concordance Account. Hence, at most only one variable, either *Attitude* or *General*, is a cause of *Intentional*, and possibly, neither is a cause of *Intentional*. That is, at least one of the two positive causal hypotheses is incorrect, and it might be that both are incorrect. We conclude that, if we assume the Markov and Faithfulness conditions, Sripada's Deep Self Concordance Account, as it is currently formulated, is undermined by Sripada and Konrath's data.

Sripada might respond by modifying the way the Deep Self Concordance Account has been articulated; specifically, he could reject the second positive causal hypothesis—viz. the hypothesis that judgments about the robustness of people's attitudes (measured by the variable

[13] More precisely, a *chain* of length n connecting vertices V_1 and V_{n+1} in the graph $\boldsymbol{G}$, denoted $v_1 \leftrightarrow v_{n+1}$, is a sequence $V_1, V_2, \ldots, V_{n+1}$ of vertices such that either $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ for $i = 1, \ldots, n$. A vertex V_i is a *collider* on the chain C if and only if $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$ in C. Vertices V_A and V_B are *d-separated* by the set $\boldsymbol{S}$ of vertices in $\boldsymbol{G}$ if and only if there is no chain C between V_A and V_B such that (i) every collider on C is in $\boldsymbol{S}$ or has a descendant in $\boldsymbol{S}$, and (ii) no other vertex on C is in $\boldsymbol{S}$. Assuming the graph $\boldsymbol{G}$ is Markov and Faithful to its corresponding probability distribution, the vertices V_A and V_B are d-separated by $\boldsymbol{S}$ in $\boldsymbol{G}$ if and only if they are independent conditional on $\boldsymbol{S}$, denoted $V_A \perp\!\!\!\perp V_B \mid \boldsymbol{S}$. For further discussion see Spirtes et al. (1993/2000) and Pearl (2000).

[14] Using Fisher's exact test, the hypothesis that *Attitude* is independent of *General* conditional on *Case* is rejected (p=0.0425) while the hypothesis that *Attitude* is independent of *General* conditional on *Case* and *Intentional* fails to be rejected (p=0.2995).

General in Sripada and Konrath's study) influence intentionality judgments. This modification would allow him to hypothesize that it is the *General*→*Intentional* edge in the positive sub-model that is incorrect while maintaining the correctness of the *Attitude* →*Intentional* edge.

However, it is far from clear that this move is open to Sripada. It appears to be central to the Deep Self Concordance Account that whether people view an attitude as being robust determines whether they ascribe it to the agent's deep self. Certainly, it seems that not just any attitude should be associated with a person's assessment of the agent's deep self, on pain of undermining the distinction between the active self and the deep self that the Deep Self Concordance Account is built on. As such, it would be quite puzzling for Sripada's account if people's intentionality judgments (measured by the variable *Intentional*) causally influenced their judgments about the robustness of the agent's attitudes (measured by the variable *General*).

Alternatively, Sripada could argue that people's ascription of attitudes to an agent, and their judgments about the robustness of her attitudes are not two distinct causes of people's intentionality judgments. Instead, they are the expression of a single cause that influences people's intentionality judgments—viz., how people conceive of the agent's deep self. On this view, the question that was taken to measure the ascription of attitudes (the variable *Attitude*) and the question that was taken to measure people's judgments about the robustness of the agent's attitudes (the variable *General*) are actually different measures of a single cause, people's conception of the agent's deep self. In fact, this is arguably suggested by Sripada and Konrath's discussion, even though it is inconsistent with their causal models: They write that "asking whether the chairman will bring about similar outcomes in other contexts and 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" (forthcoming p. 11).[15]

There are two main issues with this response. First, by making this move, Sripada would concede that the Deep Self Concordance Account *as it is currently articulated* is undermined by Sripada and Konrath's data. Second, it does not appear that this new articulation of the Deep Self Concordance Account can be satisfactorily evaluated with the data currently at hand. If Sripada were to embrace this articulation, then this would mean that there is currently no empirical support for the Deep Self Concordance Account.

7. Conclusion

Contrary to appearances, Sripada's Deep Self Concordance Account, as it is currently articulated, is undermined by Sripada and Konrath's own data. There are better models than Sripada and Konrath's, and these models are inconsistent with one of the two positive causal hypotheses found in the current articulations of the Deep Self Concordance Account (people's attribution of attitudes to an agent influences their judgments about the intentional nature of that agent's action) while being silent about the other (people's judgments about the robustness of an agent's attitudes influence their judgments about the intentional nature of the agent's action). Second, the good fit of Sripada and Konrath's models is explained by variables that are irrelevant to the evaluation of the positive claims made by the Deep Self Concordance Account, while the poor fit of the sub-model that embodies the positive causal hypotheses of the Deep Self Concordance Account indicates that that this account, as it is currently articulated, is undermined by the data and should perhaps even be rejected in light of the data. Finally, the conditional dependencies and independencies among the variables relevant to the positive hypotheses of the

[15] Sripada has also argued for this response in personal communication.

Deep Self Concordance Account are such that the two positive causal hypotheses made by the Deep Self Concordance Account cannot both be true. As it is currently articulated, the Deep Self Concordance Account is unacceptable.

References

Alicke, M. (2008). Blaming badly. Journal of Cognition and Culture, 1-2, 179-186.

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. Personality and Individual Differences, 42, 815-824.

Bollen, K. (1989). Structural Equations with Latent Variables. New York: John Wiley & Sons.

Burnham, K., and Anderson, D. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. New York: Springer.

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

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. Inquiry, 52, 449-466.

Kaplan, D. (2009). Structural Equation Modeling: Foundations and Extensions, Second Edition. Thousand Oaks: Sage.

Kline, R. (1998). Principles and Practice of Structural Equation Modeling. New York: Guilford Press.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. Analysis, 63, 190-193.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. Philosophical Psychology, 16, 309-324.

Knobe, J. (2004). Intention, intentional action and moral considerations. Analysis, 64, 181-187.

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. Philosophical Studies, 130, 203-231.

Loehlin, J. (2004). Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis, Fourth Edition. Mahwah: Lawrence Erlbaum.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. Mind & Language, 23, 165-189.

Malle, B. F. (2006). The relation between judgments of intentionality and morality. Journal of Cognition and Culture, 6, 61-86.

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

Mele, A. (2006). The folk concept of intentional action: A commentary. Journal of Cognition and Culture, 6, 277-290.

Mele, A. R., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. Midwest Studies in Philosophy, 31, 184-201.

Nadelhoffer, T. (2004). Praise, side effects, and intentional action. Journal of Theoretical and Philosophical Psychology, 24, 196-213.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. Philosophical Explorations, 9, 203-219.

Nanay, B. (2010). Morality of modality? What does the attribution of intentionality depend on? Canadian Journal of Philosophy, 40, 28-40.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: the Knobe effect revisited Mind & Language, 22, 346-365.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. Mind & Language, 24, 586-604.

Pearl, J. (2000). Causality. Cambridge University Press

Phelan, M., & Sarkissian, H. (2008). The folk strike back; Or, why you didn't do it intentionally, though it was bad and you knew it. Philosophical Studies, 138, 291-298.

Raftery, A. (1995). Bayesian model selection in social research. Sociological Methodology, 25, 111-163.

Raykov, T., & Marcoulides, G. (2000). A First Course in Structural Equation Modeling. Mahwah: Lawrence Erlbaum.

Rust, R.T., Chol, L., & Valente, Jr., E. (1995). Comparing covariance structure models. International Journal of Research in Marketing, 12, 279-291.

Scaife, R., & Weber, J. (Forthcoming). Intentional side-effects of action. Journal of Moral Philosophy.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. The Journal of Education Research, 99, 323-337.

Spirtes, P., Glymour, C., & Scheines, R. (1993/2000). Causation, Prediction, and Search, 2nd Ed. Cambridge, MA: MIT Press.

Sripada, C. S. (Forthcoming). The Deep Self Model and asymmetries in folk judgments about intentional action. Philosophical Studies.

Sripada, C. S., & Konrath, S. (Forthcoming). Telling more than we can know about intentional action. Mind & Language.

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

Wright, J. C., & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. Mind & Language, 24, 24-50.