Deep Trouble for the Deep Self

David Rose, Jonathan Livengood, Justin Sytsma, Edouard Machery

The folk concept of intentional action has been the subject of extensive research by experimental philosophers and psychologists (e.g., 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. For example, researchers have been concerned with ordinary judgments about the intentional status of side effects, as in the case described in Knobe (2003a), where a negative foreseen side effect of a CEO's action—harming the environment—is judged to be intentional while a positive foreseen side effect—helping the environment—is judged to be unintentional. Call any puzzling asymmetry in ordinary judgments about intentional action (whether or not they involve 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 prescriptive judgment (Alicke, 2008; Knobe, 2003a, 2004, 2006; Mele, 2006; Mele & Cushman, 2007; Nadelhoffer, 2004, 2006; Pettit & Knobe, 2009; Uttich & Lombrozo, 2010; Wright & Bengson, 2009) or rather in terms of the interplay between various descriptive judgments (Guglielmo, Monroe, & Malle, 2009; Machery, 2008; Malle, 2006; Nanay, 2010; Sripada, forthcoming). To adjudicate between these two types of accounts—what we will call *prescriptivist* accounts and *descriptivist* accounts—Chandra Sripada and Sarah Konrath (forthcoming) used structural equation models (SEMs) to support a descriptivist account put forward by Sripada (forthcoming)—the Deep Self Concordance

Account (DSCA).[1] Sripada and Konrath argue that their data support the DSCA and also undermine the prescriptivist accounts they consider.

In the present paper, we show that Sripada and Konrath are wrong to think that their data support the DSCA. Rather, while Sripada and Konrath's data *do* undermine the prescriptivist accounts they consider, they *also* undermine the DSCA. In fact, the DSCA should even be rejected in light of Sripada and Konrath's data. Our paper proceeds as follows. In Section 1, we describe Sripada and Konrath's Deep Self Concordance Account and the claims they make on the basis of their modeling work. In Sections 2 to 4, we argue that Sripada and Konrath's data undermine the DSCA. In Section 2, we show that a search procedure for the best fitting causal models does not output Sripada and Konrath's models. In Section 3, we show that the part of Sripada and Konrath's models that embody the positive causal hypotheses made by the DSCA should be rejected in light of their own data. In Section 4, we demonstrate that, in light of the conditional dependencies and independencies among the variables measured by Sripada and Konrath, at least one of the two positive causal hypotheses made by the DSCA is unacceptable. In Section 5, we offer some concluding thoughts.

1. The Deep Self Concordance Account

Most philosophers and psychologists who have written about intentional action asymmetries have concluded that people's prescriptive, normative, or moral judgments influence their judgments about intentional action. By contrast, Sripada and Konrath's DSCA offers an account

[1] Sripada and Konrath call their account the Deep Self Concordance Model, but in order to avoid some confusion in what follows, we will reserve the term "account" for psychological theories and the term "model" for a specific structural equation instantiation of a theory.

that does not call on prescriptive judgments in explaining these asymmetries. According to the DSCA, an agent is said to have performed action φ intentionally if and only if the action φ "concords with the agent's underlying core values, attitudes, and behavioral dispositions, which constitute the agent's 'Deep Self'" (6). Applied to the CEO case, for example, the DSCA predicts that if a participant judges the CEO to have both anti-environment attitudes and stable dispositions to act against the interests of the environment, then the participant will believe that the CEO'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 CEO intentionally harmed the environment (since the outcome concords with the CEO's attitudes and values) but unlikely to say that the CEO intentionally helped the environment (since the outcome does not concord with the CEO's attitudes and values).

This account stands in sharp contrast with the prominent prescriptivist accounts examined by Sripada and Konrath. As Sripada and Konrath rightly argue, these prescriptivist accounts make different causal hypotheses about which variables influence judgments about whether a given outcome is brought about intentionally. Thus, Knobe's Good/Bad Account assert 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 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 that outcome intentionally.

In order to test the various accounts they consider, Sripada and Konrath presented 240 students at the University of Michigan with Knobe's CEO story along with six questions—the original question about whether the CEO intentionally helped or harmed the environment and

five additional questions intended to measure other judgments that might explain the asymmetry between judgments in the help condition and in the harm condition. Five of the questions as well as the assignment to a harm or help condition were represented as variables in two separate structural equation models. The questions and variable labels that we will use throughout the rest of the present paper are given in Table 1 below.

Variable Name	**Question**	**Anchors**
Case	N/A – Participants were assigned to a "harm" condition or to a "help" condition.	N/A
Int	How much do you agree with the statement 'The Chairman intentionally harmed [helped] the environment'?	Strongly Agree, Strongly Disagree
GB	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
Att	What are the Chairman's values and attitudes towards the environment?	Very Pro-environment, Very anti-environment
Gen	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 structural equation models were meant to simultaneously test all of the examined accounts of the intentionality judgment asymmetries. Variable *GB* 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 DSCA predicts that neither *GB* nor *Moral* cause *Int.* Call this Sripada and Konrath's *negative causal hypothesis*. The variables *Gen* and *Att* are directly relevant to testing the DSCA. Specifically, variable *Att* is supposed to

measure participants' views about the CEO's core values and attitudes (with respect to the environment), and variable *Gen* is supposed to measure participants' views about how likely the CEO is to act in the same way in a range of other contexts, i.e. whether the CEO has a stable disposition to help or harm the environment. As Sripada and Konrath put it, "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" (p.7). The DSCA predicts that both *Att* and *Gen* cause *Int*. Call this Sripada and Konrath's *positive causal hypothesis*.

Sripada and Konrath fit an initial model in order to test the causal hypotheses made by the DSCA and the prescriptivist accounts. After examining the fit of their initial model, they considered what are called 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.[2] 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. In agreement with their positive and negative causal hypotheses, these models have a positive part and a negative part. The positive part shows that *Att* and *Gen* cause *Int*, while the negative part shows that neither *GB* nor *Moral* causes *Int*. One of Sripada and Konrath's two models is pictured (with unstandardized path coefficients) in Figure 1. (Their other model is obtained by reversing the edge from *GB* into *Att*.)

[2] 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.

	Case	GB	Moral	Att	Gen	Int
Case	0.251					
GB	0.843	4.877				
Moral	0.13	0.493	1.422			
Att	0.732	2.74	0.389	3.581		
Gen	-0.707	-2.542	-0.35	-2.275	3.814	
Int	-0.703	-2.381	-0.456	-2.572	2.51	4.46

Table 2: Covariance Matrix from Sripada and Konrath

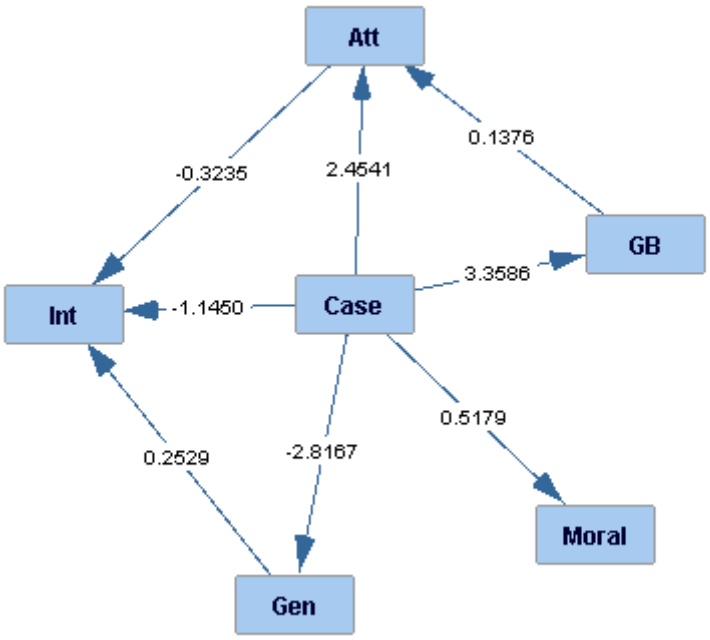

Figure 1: One of Sripada and Konrath's Models[3]

Sripada and Konrath's models have very good overall fit to their data, providing support for both their negative and positive causal hypotheses.[4] Thus, it seems that the causal hypotheses made

[3] Each edge in the causal model shown above represents a direct causal connection, and the numbers on each edge are linear coefficients. For example, according to the model in Figure 1, the expected value of *Gen* given that *Case* takes the value *c* is $E(Gen \mid Case = c) = -2.8167 \cdot c$. In the case of ordinary regression, the conditional expectation is observational in character, meaning that it tells us what value we can expect *Gen* 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 their model to count as structural (or causal), a number of strong assumptions need to be satisfied. The substantive variables need to be linearly related and free of measurement error (since this is a single-indicator path model). The error terms need to be normally distributed and uncorrelated with each other. Furthermore, the substantive 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. We doubt that all or even most of these assumptions are satisfied for Sripada and Konrath's model. However, in the present paper, we set these doubts aside in order to give them the strongest possible position from which to defend their model.

[4] A summary of the usual fit indices for Sripada and Konrath's model is given in Table 3 below.

by the DSCA are supported, whereas the prescriptivist accounts are not supported.[5] What's not to like? Surprisingly, quite a lot.

2. Trouble from Alternative Models

Sripada and Konrath employ a method typical in social-scientific use of SEMs: 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 true model 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 true model by conducting modification tests is not good. Even in small search spaces, the number of possible structural equation models may be very large. 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. Assuming that *Case* is not caused by anything, there are still $2^5 \cdot 3^{10}$ distinct models. Restricting attention to directed acyclic graphs still leaves 936,992 admissible models.[6] In search spaces this large, it

[5] We agree with Sripada and Konrath that their data undermine the prescriptivist accounts they consider. We will not challenge this conclusion in what follows (but see our concerns with the modeling assumptions in footnote 3).

[6] 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

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. A more principled approach to search seems to be called for.

We used a Greedy Equivalence Search (GES) in Tetrad IV to search for 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.[7] The output of the constrained GES algorithm is given in Figure 3. GES succeeds in orienting all but one edge in the graph: the edge between *Int*

---

corresponding to a structural equation model is acyclic (and the error terms in the model are uncorrelated), then the model is called *recursive*.

[7] 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/.

and *Gen*. The edge is not oriented because the two orientations correspond to statistically equivalent structural equation models.

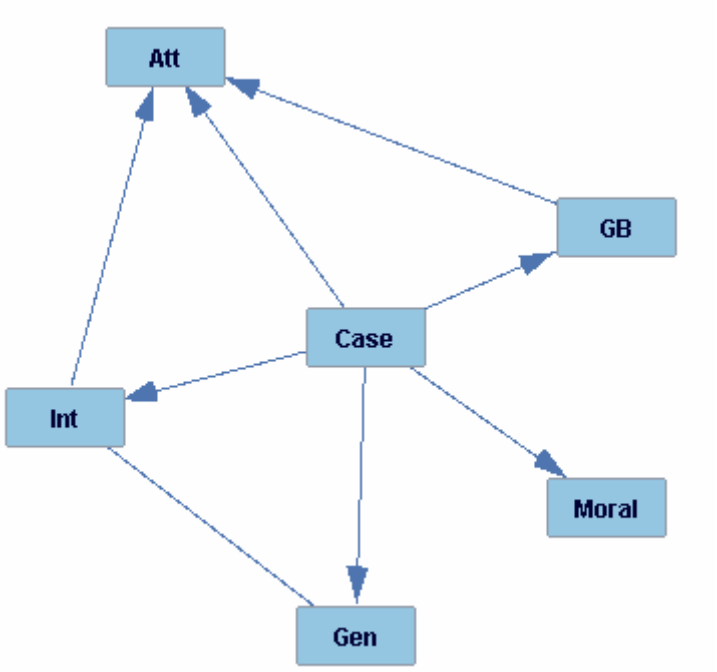

Figure 3: GES Output (Tetrad Models)

Call the two equivalent models output by GES the *Tetrad Models*. From Figure 3, we can see that the edges about which the Tetrad Models and Sripada and Konrath's models disagree are exactly those edges that matter for the positive causal hypothesis of the DSCA. Contrary to what the DSCA predicts, the Tetrad Models indicate that judgments about whether the CEO acted intentionally cause judgments about the CEO's attitude towards the environment, *not the other way round*. Furthermore, the GES-based analysis shows that the data are silent about whether or not judgments about the generalizability of the chairman's actions affect judgments of intentionality. Perhaps even more surprising is that the only variable that is known by GES to causally influence *Int* is *Case*.

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.[8]

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: Side-by-Side Indices of Fit

Hence, the Tetrad Models have two advantages over Sripada and Konrath's models. First, they fit the data better. Second, they are the products of a reliable search procedure: a procedure that

[8] For example, Kline (1998, 137-138) writes, "Sometimes researchers specify alternative models that are not hierarchically related. Although the values of the χ^2 statistics from two nonhierarchical models can still be compared, the difference between them cannot be tested for significance. … Given two nonhierarchical models, the one with the lowest AIC is preferred." Raftery (1995, 138-141) attempts to quantify BIC differences in terms of posterior probabilities. According to his system, the BIC difference between Sripada and Konrath's model and the Tetrad model corresponds to the claim that the posterior probability of the Tetrad model (given that one of the models is true) falls in the range 0.75 to 0.95.

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 one of the two positive causal relations hypothesized by the DSCA (*Att* → *Int*) and agnostic about the other causal hypothesis (*Gen* → *Int*), we conclude that the data reported by Sripada and Konrath fail to support the DSCA. In the next two sections, we argue that Sripada and Konrath's DSCA are actually undermined by their own data.

3. Trouble from Model P-Values

Sripada and Konrath may reply to our first line of argument that, although the Tetrad Models fit the data *better* than their models, their models are at least *consistent* with the data (in contrast to the causal models derived from the prominent prescriptivist accounts of intentional action judgment). After all, their models have admissible fit indices and p-values.

This reply would raise a difficult question about the nature of empirical support: Does empirical support for a model increase along with its fit to the data, as Bayesians and likelihoodists might maintain, or are different models equally well-supported whenever their fit indices are acceptable (e.g., when their p-value is above some threshold) independently of how well these hypotheses fit the data, as falsificationists might maintain? Fortunately, we do not have to solve this difficult question since the DSCA should be rejected by either standard in light of Sripada and Konrath's own data.

Sripada and Konrath's models have good overall fit to their data. However, fit indices (including but not limited to the p-value) for a model indicate how well the model *as a whole* fits the data but not how well any particular *component* of the model fits the data. A model might

have great overall fit and still have some component or sub-model that does not have good or even acceptable fit.[9] Given that model p-values work this way, we decided to test whether the part of Sripada and Konrath's models that embodies their positive causal hypothesis (*Att* → *Int* and *Gen* → *Int*) fits the data. Thus, we split Sripada and Konrath's models into two sub-models: *the positive sub-model* (including the variables *Att*, *Gen*, *Int*, and *Case*) and *the negative sub-model* (including the variables *Moral*, *GB*, *Att*, and *Case*). The positive sub-model and the negative sub-model are pictured in Figure 2.

[9] All structural equation modeling assumes that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, 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 $\mathbf{S}$. 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 $\mathbf{S}$ or because some of the entries in $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ are extremely close to $\mathbf{S}$, even though other entries in $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ are not even acceptably close to $\mathbf{S}$. See Bollen (1989, 104 ff. and 256 ff.) for gory details.

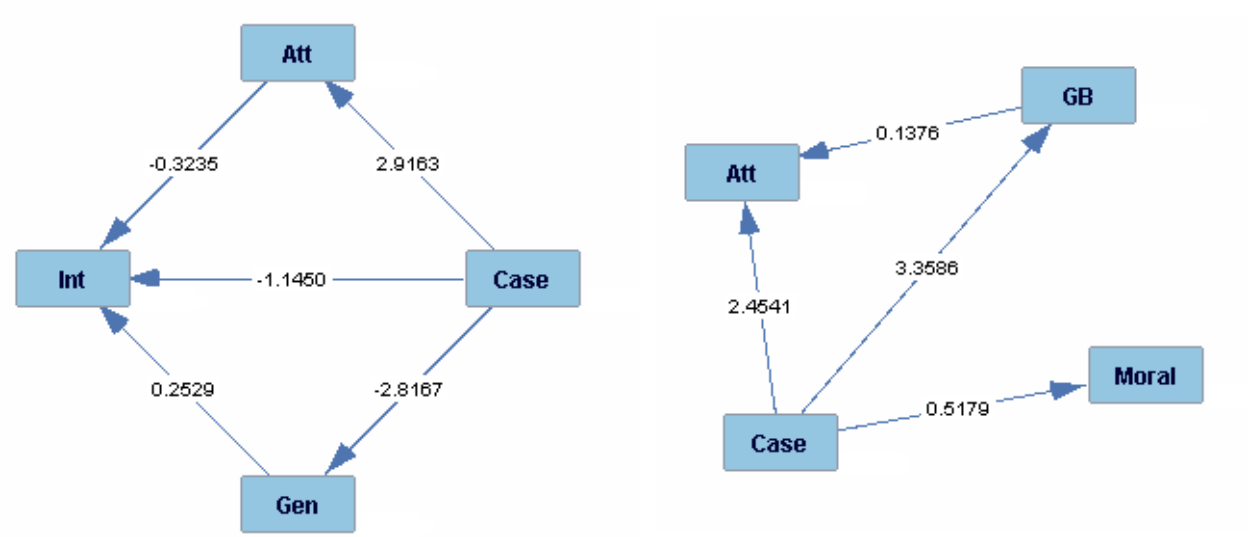

Figure 2: Positive (Left) and Negative (Right) Sub-Models

The negative sub-model fits the data extremely well.[10] However, the positive sub-model is rejected by a chi-square test at the 0.05 significance level.[11] Thus, the exceptionally good fit of the negative sub-model explains why Sripada and Konrath's models fit the data. This can be seen from the side-by-side fit indices for the negative and positive sub-models pictured in Table 4.

Fit Index	S&K	Positive	Negative
Chi-Square (DF)	6.999 (7)	4.156 (1)	0.275 (2)
p-value	0.42898	0.0415	0.87157
Adjusted GFI	0.97154	0.91455	0.99713
Bentler-Bonnett NFI	0.9912	0.99268	0.99938

[10] The model chi-square statistic for the negative sub-model is 0.275 with two degrees of freedom (p=0.872).

[11] The model chi-square statistic for the positive sub-model is 4.156 with one degree of freedom (p=0.042).

Tucker-Lewis NNFI	1	0.96628	1.0119
SRMR	0.018832	0.019619	0.006911
BIC	-31.365	-1.3252	-10.6864

Table 4: Fit Indices for the Positive and Negative Sub-Models

Despite the fact that Sripada and Konrath's models fit the data, the positive sub-model, which embodies the positive causal hypotheses of the Deep Self Concordance Account, should be *rejected* on the basis of the data.

4. Trouble from Colliders

Things only get worse for Sripada and Konrath when we look more closely at the positive sub-model. 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.[12] Assuming the Markov and Faithfulness

[12] 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

conditions, Sripada and Konrath's positive sub-model entails (i) that *Att* is independent of *Gen* conditional on *Case* and (ii) that *Att* is associated with *Gen* conditional on *Case* and *Int*. However, neither (i) nor (ii) is satisfied by the data.[13] The upshot is that at least one of the edges *Gen—Int* or *Att—Int* cannot be oriented in the way required by the DSCA. Hence, at most only one variable, either *Att* or *Gen,* is a cause of *Int*, and possibly, neither is a cause of *Int*. Since the DSCA predicts that both *Att* and *Gen* cause *Int*, the DSCA is undermined by the data.

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. But as long as these conditions are assumed, the causal facts given the data simply cannot be as they are required to be on the basis of the DSCA. Hence, the DSCA should be rejected on the basis of Sripada and Konrath's data.

5. Trouble Explaining Intentionality Ascription Asymmetries

Let's review the three troubles we have just presented. The two causal models derived from the Deep Self Concordance Account are not output by the search procedure we used (GES), and they

---

on *C* is in ***S***. Assuming the graph ***G*** is Markov and Faithful to its corresponding probability distribution, the vertices V_A and V_B are d-separated by ***S*** in ***G*** if and only if they are independent conditional on ***S***, denoted $V_A \perp\!\!\!\perp V_B \mid \boldsymbol{S}$. For further discussion see Spirtes, P. et al. (1993/2000) and Pearl (2000).

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

are not the best fitting models over the variables Sripada and Konrath measured. Furthermore, the good overall fit displayed by Sripada and Konrath's models is entirely due to the good fit of the model defined over variables irrelevant to the truth of the Deep Self Concordance Account (the negative sub-model), while the fit of the sub-model embodying the positive causal hypotheses made by Sripada and Konrath (the positive sub-model) is so low that it should be rejected in light of their data. Finally, the conditional dependencies and independencies among the variables relevant to the DSCA (the positive sub-model) are such that the two positive causal hypotheses made by the Deep Self Concordance Account (*Att* → *Int* and *Gen* → *Int*) cannot both be true. We conclude that Sripada and Konrath's data undermine the Deep Self Concordance Account and that, just like the most prominent prescriptivist accounts rightly criticized by Sripada and Konrath, this account does not seem able to explain the puzzling asymmetries in ordinary judgments about intentional action.

References

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

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

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.

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.

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.

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

Spirtes, P., Glymore, C, & Scheines, R. (1993/2000) *Causation, Prediction, and Search*, 2nd Ed. 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.*

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

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.