View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by ScholarWorks @ Georgia State University

Georgia State University
ScholarWorks @ Georgia State University

Philosophy Theses

Department of Philosophy

Spring 4-13-2011

Normative Judgments, 'Deep Self' Judgments, and Intentional Action

Jason S. Shepard
Georgia State University

Follow this and additional works at: https://scholarworks.gsu.edu/philosophy_theses

Part of the Philosophy Commons

Recommended Citation

Shepard, Jason S., "Normative Judgments, 'Deep Self' Judgments, and Intentional Action." Thesis, Georgia State University, 2011.
https://scholarworks.gsu.edu/philosophy_theses/86

This Thesis is brought to you for free and open access by the Department of Philosophy at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Philosophy Theses by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

NORMATIVE JUDGMENTS, 'DEEP SELF' JUDGMENTS, AND INTENTIONAL ACTION

by

JASON S. SHEPARD

Under the Direction of Eddy Nahmias

ABSTRACT

Sripada and Konrath (forthcoming) use Structural Equation Modeling techniques to provide empirical evidence for the claim that implicit and automatic inferences about people's dispositions, and not normative judgments, are the driving cause behind the pattern of folk judgments of intentional action in Knobe's (2003a) chairman case. However, I will argue that their evidence is not as strong as they claim due to the potential of methodological and statistical problems with the way they tested their model. After correcting for these problems, I show that even after accounting for the role of dispositional inferences, normative judgments are still playing a significant role in folk judgments of intentional action.

INDEX WORDS: Intentional action, Normative judgments, Moral judgments, Dispositional inferences, Deep Self, Chandra Sripada, Joshua Knobe

NORMATIVE JUDGMENTS, ‘DEEP SELF’ JUDGMENTS, AND INTENTIONAL ACTION

by

JASON S. SHEPARD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

In the College of Arts and Sciences

Georgia State University

2011

Copyright by
Jason Scott Shepard
2011

NORMATIVE JUDGMENTS, 'DEEP SELF' JUDGMENTS, AND INTENTIONAL ACTION

by

JASON S. SHEPARD

Committee Chair: Eddy Nahmias

Committee: George Graham

Chris Henrich

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2011

TABLE OF CONTENTS

LIST OF TABLES		v
LIST OF FIGURES		vi
1	**INTRODUCTION**	1
2	**THE SIDE-EFFECT EFFECT: EVIDENCE FOR BIDIRECTIONALISM**	9
3	**THE DEEP SELF MODEL**	11
4	**THE PROBLEM OF MODELING IMPLICIT INTERACTION TERMS**	22
	4.2 **Methods**	29
	4.3 **Results and Discussion***	29
5	**THE PROBLEM OF USING MANIPULATION CHECKS AS MEDIATORS***	34
	5.2 **Methods**	38
	5.3 **Results and Discussion***	38
6	**THE DEEP SELF MODEL IS STILL BIDIRECTIONAL**	40
	6.2 **Methods**	46
	6.3 **Results and Discussion***	46
7	**THE CASE OF THE CARING CHAIRMAN**	50
	7.2 **Methods**	56
	7.3 **Results and Discussion***	56
8	**A DISAPPEARING ACT?**	63

LIST OF TABLES

Table 1. Questions from Sripada and Konrath's (forthcoming) study on intentional action. 13

Table 2. Summary of fit indices for the replications of Sripida and Konrath's base model using the recoded chairman values/attitudes judgments to reflect an implicit interaction term and the analogous model using an explicit interaction term. 33

Table 3. Means for intentionality judgments and chairman values/attitudes judgments 47

Table 4. Means for intentionality judgments and chairman values/attitudes judgments in the caring chairman case 57

LIST OF FIGURES

Figure 1. Sripada and Konrath's (forthcoming) base model. 17

Figure 2. Sripada and Konrath's (forthcoming) final model. 19

Figure 3. Replication of Sripida and Konrath's base model using the recoded chairman values/attitudes judgments to reflect an implicit interaction term. 31

Figure 4. Replication of Sripida and Konrath's base model using an explicit interaction term. 32

Figure 5. The Deep Self Model predicted simple slopes of intentionality judgments regressed on case at different levels of chairman values/attitudes judgments controlling for all the other variables in the model. 44

Figure 6. The observed simple slopes of intentionality judgments regressed on case at different levels of chairman values/attitudes judgments controlling for all the other variables in the model. 51

Figure 7. The observed simple slopes of intentionality judgments regressed on case at different levels of chairman values/attitudes judgments controlling for all the other variables in the model for the caring chairman cases. 61

1. INTRODUCTION

Two children, Elijah and Sacha, are playing on a playground when Elijah runs into Sacha, knocking her down and causing minor injury. The caregivers rush over to tend to Sacha. After assessing the degree of the injury, the caregivers find themselves faced with the task of trying to determine the wrongness of Elijah's act, and also trying to determine what sort of punishment (if any) would be appropriate. There are many facts about Elijah, Sacha, and the situation that the caregivers may consider when trying to assess the wrongness of Sacha's action and what kind of punishment (if any) would be appropriate. One of these considerations—perhaps one that would figure prominently in these assessments—would be whether or not Elijah *intentionall*y knocked down Sacha. If Elijah was deemed to have intentionally knocked down Sacha, the act may be considered more wrong (and it may deemed that a harsher punishment is appropriate) than if the act was deemed to have been unintentional.

What this simple story illustrates is that judgments of intentional action—that is judgments related to answering questions like "Did person *y* intentionally do *x*?"—are important for making normative judgments. Here I mean to use the term 'normative judgments' fairly broadly to include judgments about wrongness, responsibility, and punishment, along with valuation judgments like judgments concerning goodness or badness. In the context of various theories that I will be discussing later, I take it that judgments of norm violations—that is judgments concerning when people act in a way that is contrary to the way we expect people *should* act—are of primary importance. However, given the dialect to which this thesis belongs, the primary concern is the role of normative judgments (like the ones listed above); the primary concern is not which of these normative judgments types are of primary importance.

This inference from judgments of intentional action to normative judgments like wrongness of act or appropriateness of punishment seems to be a natural and appropriate order of inference. However, what about the reverse interest—that is, from normative judgments like wrongness of act or appropriately of punishment to judgments of intentional action? For example, what if the caregivers first made a judgment based on a valuation of the badness of Sacha's injury, then used these judgments to inform their judgments of intentional action. Say, if Sacha broke her arm, then the caregivers would conclude that Elijah intentionally knocked down Sacha, but if Sacha only received a scratch, then the caregivers would conclude that Elijah did not intentionally knock down Sacha. This sort of inference should seem really peculiar, and if we witnessed the caregivers reasoning in this way, we would be liable to say that the caregivers' reasoning was in gross error.

However, a growing number of philosophers and psychologists have recently put forth views that argue that we often do make inferences from normative judgments to judgments of intentional action (Alicke, 2008; Knobe, 2003a, 2006, 2010; Mele, 2006; Nadelhoffer, 2004a, 2006; Nado, 2008; Uttich & Lombrozo 2010; Wright & Bengson, 2007).

Before continuing, it will be helpful to introduce some terminology and some distinctions. I will use the term *unidirectional account* to refer to those accounts that assume that there is a unidirectional relationship between intentional action judgments and normative judgments. This unidirectional assumption comes in two forms: descriptive unidirectionality and normative unidirectionality. *Descriptive unidirectionality* holds that, under normal conditions, people do in fact use intentional action judgments as inputs into normative judgments but do not use normative judgments as inputs into intentional action judgments. *Normative unidirectionality* holds that normative judgments cannot appropriately act as inputs into intentional action

judgments. If, in some given case normative judgments did influence intentional action judgments, then this influence would be an error. These two components can come apart, as one can easily hold normative unidirectionalism while admitting that descriptive unidirectionalism is often violated.

I will use the term *bidirectional account* to refer to those accounts that assume there is a bidirectional relationship between intentional action judgments and normative judgments—that is, to refer to those accounts that hold that normative judgments can, and often do, act as inputs into intentional action judgments. For most proponents of bidirectionality, this bidirectionality claim is primarily a descriptive claim that can be seen as a rejection of the descriptive component of unidirectionalism, at least across certain cases (Alicke, 2008; Mele, 2006; Nadelhoffer, 2004a, 2006; Nado, 2008; Wright and Bengson, 2007). Most of these researchers tend to be either agnostic about normative unidirectionalism or accept normative unidirectionalism, explicitly stating that descriptive bidirectionalism occurs due to some kind of bias or error. However, Joshua Knobe has extensively argued that the bidirectional relationship between normative judgments and many folk psychological judgments are reflective of underlying competencies, not caused by bias or error (Knobe, 2006; Knobe 2010). This claim by Knobe can be seen as a rejection of the normative component of unidirectionalism.

These unidirectional and bidirectional accounts are about the relationships between normative judgments and judgments of intentional action. These accounts attempt to answer the question: When people make judgments of intentional action, are they using (should they be using) normative judgments as inputs into their judgment of intentional action? This question is a very different from the question: When people make judgments concerning whether or not something is the appropriate sort of thing to which intentional action can be attributed, are they

using (should they be using) normative judgments as inputs into their judgments of agenthood. The sorts of unidirectionalists and bidirectionalists account being discussed in this thesis are concerned with the former question not the latter. In other words, neither the unidirectional nor the bidirectional accounts discussed in this thesis make any claims concerning whether or not there are (or should be) normative constraints on agenthood (or on judgments concerning agenthood).

The debate between unidirectionalists and bidirectionalists has potential implications for several debates within philosophy, the law, and other domains. For example, many philosophical accounts of intentional action explicitly view *conservatism* as being a virtue of theorizing about the philosophically relevant concept 'intentional action'. Conservatism is the view that folk belief and folk concept application is an important constraint on philosophical theorizing (Doris, Knobe, Woolfolk, 2007). For example, Al Mele states, "[A] philosophical analysis of intentional action that is wholly unconstrained by that concept [the folk concept] runs the risk of having nothing more than a philosophical fiction as its subject matter" (Mele, 2001, p. 27). If it can be convincingly shown that normative judgments pervasively influence judgments of intentional action, this would lead proponents of conservatism to a potentially uneasy choice: (a) Admit that a philosophically appropriate account of intentional action must account for the influence of normative judgments on intentional action, (b) convincingly argue that the folk are pervasively and systematically confused in these situations and provide some kind of error theory that would allow the conservative to remain conservative while rejecting the folk's application of intentional action when these applications are (wrongly) influenced by normative judgments, (c) or reject

conservatism when it comes to the influence of normative judgments on judgments of intentional action.[1]

Another philosophically relevant debate for which these views are potentially important concerns the debate over the fundamental nature of folk psychology and the pervasiveness of moral cognition. By *folk psychology*, I simply mean the pre-theoretical system of inferences that we make about the mental states, behavior, causes, etc. of others and their actions. Many in philosophy and psychology have proffered the view that folk psychology is fundamentally a system employed for explanation and prediction (Churchland, 1989; Gopnik & Melzoff, 1997, Gordon, 1986; Goldman, 1989).[2] However, on the basis that normative judgments pervasively influence judgments of intentional action and other folk psychological judgments, Joshua Knobe (Knobe, 2006; Knobe, 2010) has argued that the evidence should lead one to a completely different conclusion: At its base, folk psychology has a fundamentally normative component that is completely independent of any role for explanation and prediction.[3]

Additionally, as is evident with the story of Elijah and Sacha and perhaps even more prominently displayed in legal examples, central to our everyday and legal understanding of responsibility and punishment include issues involving judgments of intentional action. Just as Elijah would have been considered more fully responsible if he intentionally knocked down

---

[1] Even revisionists, who are willing to flaunt certain supposed constraints of folk usage of a concept, should find the patterns of folk application of a concept relevant, as I take it that it is important for the revisionists (and philosophy in general) to understand just how revisionary and in what regards their accounts are revisionary.

[2] Though generally recognized as competing theories about folk psychology, theory theorists, simulation theorists, and eliminativists as represented by Churchland share the commonality of viewing folk psychology as being fundamentally a system employed for explanation and prediction.

[3] The 'completely independent' clause of Knobe's account is a very important one, as other researchers who admit that normative judgments pervasively influence folk psychological judgments have tried to work this fact into a framework of explanation and prediction (see for example Uttich & Lombrozo, 2010, Pizarro & Tannenbaum, forthcoming).

Sacha, judgments concerning one's degree of responsibility for acts ranging from violating the dress code of a party to harming the environment, from walking in on someone using the restroom to the destruction of property, and so forth generally depend, at least in part, on judgments concerning whether the act was committed intentionally. However, if judgments of intentional action are being influenced by normative judgments, then this bidirectional relationship between judgments of intentional action and normative judgments calls into question whether judgments of intentional action can function as the sorts of judgments that can, independently and impartially, guide judgments of responsibility (or wrongness of act or etc.)

A related philosophically relevant issue—one with practical implications for law—has to do with issues this debate raises for the potential of jury impartiality (Nadelhoffer, 2006). Potential criminal acts are, in part, differentiated by *mens rea* conditions even in cases where the outcome is the same. *Mens rea* is Latin for guilty mind and refers to the mental states an offender had in relation to the act. *Mens rea* conditions can refer to whether an offender committed an act knowingly, willingly, purposefully, intentionally, with forethought, with malice, etc. For example, consider a case in which a jury is faced with determining the appropriate conviction for someone who has taken the life of another. Assuming that the taking of the life did not occur in the commission of another felony, the choice of conviction may come down to a choice between murder or manslaughter. To put the weight of this decision in practical terms, a conviction of murder is typically accompanied by a punishment ranging from 20 years to life in prison. A conviction of manslaughter is typically accompanied by a punishment of up to five years in prison. In this case, one way in which murder may be distinguished from manslaughter is whether the defendant took the life *intentionally* as in the case of murder (Model Penal Code, § 24.02) or whether the taking of the life was merely *reckless* (and not intentional) as in the case of

manslaughter (Model Penal Code, § 24.05). Thus, getting the conviction correct comes down to a determination of whether or not the defendant intentionally took the life or whether the taking of the life was merely reckless. However, if normative judgments sometimes influence our judgments of intentional action, the ability of jurors to impartially determine whether or not the defendant acted intentionally is extremely suspect, as it seems that we may be predisposed to judge norm violations, like the taking of another's life, as being brought about intentionally. In other words, due to the influence of normative judgments, a jury would be more likely to convict the defendant of murder than manslaughter all other things being equal.

All of these philosophically relevant problems would disappear[4]if it could be shown that, in fact, normative judgments do *not* influence folk judgments of intentional action. One recent and prominent account that attempts to show that normative judgments are not influencing folk judgments of intentional action—and thus, if correct, would make the above-mentioned problems disappear—is Chandra Sripada's Deep Self Model of Intentionality and Responsibility (hereafter, 'Deep Self Model') (Sripada, 2010; Sripada & Konrath, forthcoming). In brief, Sripada claims that normative judgments do not play a role in the production of judgments of intentional action (above and beyond the influence of the relevant, non-normative judgments); rather what he calls 'deep self concordance' judgments are of primary importance for the production of judgments of intentional action ('deep self concordance' will be explained in greater detail in Section 3).

[4] At the very least, if these problems persisted, it would be due to content from sources outside of the debate of current topic. For example, the concern of juror impartiality will not go away if it could be shown that normative judgments are not influencing intentional action judgments. Likely, the problem would still persist due to issues such as stereotyping, prejudice, etc. Rather, showing that normative judgments are not influencing intentional action judgments would show that problems of juror impartiality qua normative-judgments-acting-as-inputs-into-intentional-action-judgments go away.

In this thesis, I will argue that Sripada's model falls short of being able to convincingly demonstrate that normative judgments are not influencing folk judgments of intentional action, and as a consequence, the above-mentioned philosophical problems remain real problems. Along the way, I will offer a more plausible version of a deep-self model—one that is explicitly bidirectional.

The thesis is organized as follows: In the next section, I briefly present some evidence for the claim that normative judgments do in fact influence judgments of intentional action. In Section 3, I offer a somewhat detailed account of the Deep Self Model. In Sections 4 and 5, I outline two methodological issues with Sripada's attempt to empirically test his model (Sripada and Konrath, forthcoming). The dialectical role of these sections will be two-fold: First, the problems shed doubts on Sripada's empirical findings and his interpretation of his findings, hence raising doubts concerning the empirical support for his theoretical model. Second, the problems raised are meant to motivate an empirical re-analysis of his model using corrected methodology. In Section 6, I offer a sketch of a plausible bidirectional deep-self model, and then demonstrate that once the methodological issues raised in Sections 4 and 5 are corrected, the evidence favors the view that, at least in some cases, normative judgments are influencing judgments of intentional action above and beyond the influence of deep-self judgments. In other words, the evidence favors the predictions made by a bidirectional deep-self model over Sripada's Deep Self Model. In Section 7, I present the results of a new experiment that was designed to test additional predictions of the kind of bidirectional deep-self model that I outline in Section 6. In Section 8, I conclude with a brief discussion concerning how the results reported in this paper may shed light on the status of the above-mentioned philosophically relevant debates about people's judgments of intentional action.

2. THE SIDE-EFFECT EFFECT: EVIDENCE FOR BIDIRECTIONALISM

A major part of the bidirectionalists' argument against the descriptive component of unidirectionalism relies on a growing body of experimental evidence that appears to support a bidirectional account.

Take for example, the now classic study on intentional action by Joshua Knobe (2003a). In this study, Knobe randomly presented participants with one of two vignettes. Both vignettes involved a decision made by the chairman of the board of a company. The only thing that differed about the vignettes was the moral valence of a foreseen side effect of the chairman's decision. In this case, it was whether the environment was harmed or helped as a result of the chairman's decision to implement a new policy. The harm vignette read as follows:

> 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.

The help vignette read exactly as the harm vignette, except all instances of 'harm' language were replaced with 'help' language.

After participants were presented with the vignette, they were asked if the chairman intentionally harmed [helped] the environment. When presented with the harm condition, 82

percent of the participants judged that the chairman intentionally harmed the environment, whereas in the help condition, only 23 percent judged that the chairman intentionally helped the environment. This asymmetrical pattern of attribution based on changes in the moral valence of a side effect has become known as the *side-effect effect*.

These results have been replicated using a diverse array of vignettes (see for example, Knobe 2003b, Mallon 2008, Nadelhoffer 2004b), in a diverse array of populations, including with Hindi speakers when the vignettes are translated into Hindi (Knobe & Burra 2006), with four-year old children (Leslie et al. 2006), and with subjects who suffer from deficits in emotional processing due to lesions in the ventromedial prefrontal cortex (Young et al. 2006). Similar asymmetrical patterns of attribution have also been found for the attributions of knowledge (Beebe & Buckwalter, 2010), valuing (Knobe & Roedder, 2009), decided, advocated, and in favor of (Pettit and Knobe, 2009), causation (Hitchcock & Knobe, 2009), and act individuation (Ulatowski, ms).

These above-cited examples of the replications and generalizations of the side-effect effect (and side-effect effect-like effects) are but a small fraction of the entire body of literature covering the persistence and pervasiveness of the side-effect effect, not to mention sources of evidence outside of side-effect effect studies that suggest that normative judgments can, and often do, influence judgments of intentional action. This growing body of evidence requires explanation. The consensus view is that descriptive bidirectionalism is the best explanation. If these researchers are right, then the philosophically relevant issues raised in the introduction of this thesis remain very real problems with which philosophers (and others) will have to contend. However, in spite of this growing consensus, some researchers have attempted to maintain a descriptive unidirectional account of the data. One of these researcher’s is Chandra Sripada, who

offers a descriptive unidirectional model that he calls the Deep Self Concordance Model of Intentionality and Responsibility ('Deep Self Model'). In the following section, I will turn to giving a detailed exposition of Sripada's Deep Self Model.

3 THE DEEP SELF MODEL

In rejection of a bidirectional account of the side-effect effect, in particular, and judgments of intentional action, more generally, Sripada offers the Deep Self Model. The Deep Self Model consists of two main theses: First, normative judgments are not influencing judgments of intentional action. Second, rather, "deep self concordance" judgments are primarily responsible for our judgments of intentional action. *Deep-self judgments* are attributions about an agent's stable attitudes, values, and behavioral dispositions (Sripada, 2010; Sripada & Konrath, forthcoming). This notion of a 'deep self' is borrowed form Hume and can be contrasted with one's acting self. Additionally, these deep-self judgments are similar to what some psychologists call dispositional attributions and are meant to be trait-like attributions. Deep-self concordance occurs when the psychological attitudes of the agent's deep self concords, or matches up with, an act. So, deep-self concordance judgments are judgments, generally made by a third party, that an agent's deep self concords with some act.

Sripada's Deep Self Model proposes that when someone believes that an agent's actions are in concordance with the agent's deep self (i.e., when the action concords with the agent's underlying values and attitudes), she will be more likely to say that the agent intentionally committed the act in question. According to the model, when trying to figure out whether an agent's deep self is in concordance with an action, people consider non-normative facts about the situation and are not making judgments based on normative considerations.

In the case of Knobe's chairman vignettes, in both the harm and the help scenarios, the chairman expresses not caring about bringing about the side effect of harming [helping] the

environment. According to Sripada, the statement of indifference toward the environmental outcome may be taken to express general contempt or hostility toward the environment. Thus, in both conditions, participants attribute anti-environmental attitudes toward the chairman. However, only in the harm condition does the chairman bring about an outcome that is in concordance with the attribution of anti-environmental attitudes. Thus, according to the model, it should be more likely that the participants will say that the chairman intentionally harmed the environment in the harm condition than the participants would be willing to say the chairman intentionally helped the environment in the help condition. Sripada's model would also predict that if participants attributed pro-environmental attitudes to the chairman, then intentionality attributions would be more likely in cases where the chairman helped the environment.

To test the Deep Self Model, Sripada and Konrath (forthcoming) randomly assigned 240 participants to the harm or the help condition of Knobe's chairman vignette. Following the vignette, participants were presented with six questions: one question concerning whether the chairman intentionally harmed [helped] the environment, two questions that were meant to measure key candidate explanatory variables for two different bidirectional accounts, two questions that were meant to measure candidate explanatory variables for Sripada's Deep Self Model, and a question about the participants' own values concerning the environment. See Table 1 for a summary of the questions asked and the model that each question was meant to represent.

Sripada and Konrath analyzed their data using a Structural Equation Modeling (SEM) method known as structural path analysis. SEM is a statistical technique that allows for the simultaneous measurement of multiple and complex linear relationships among observed or latent variables. Before continuing, it will be helpful to review some basic SEM terminology. *Structural path analysis* is an SEM technique that only uses observed variables. A *path diagram*

Table 1. Questions from Sripada and Konrath's (forthcoming) study on intentional action.

Question	Anchors for 7-point scale	Abbreviated Variable Name	Model (author to whom Sripada ascribed model)
How much do you agree with the statement 'The Chairman intentionally harmed [helped] the environment'?	Strongly Agree, Strongly Disagree	Intentionality Judgments	N/A
In your view, how good or bad is the outcome that the environment is harmed [helped]?	Very Good, Very Bad	Goodness/Badness Judgments	Good/Bad Model, bidirectional (Knobe, 2005, 2006, no longer endorses model; Buckwalter & Beebe, 2010)
In your view, what is the Chairman's moral status?	Very Moral, Very Immoral	Moral Status Judgments	Moral Status Model, bidirectional (Alicke, 2008, 1982)
What are the Chairman's values and attitudes towards the environment?	Very Pro-environment, Very Anti-environment	Chairman Values/Attitudes Judgments	Deep Self Model, unidirectional (Sripada, forthcoming)
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	Deep Self Model, unidirectional (Sripada, forthcoming)
What are your own values and attitudes towards the environment?	Very Pro-environment, Very Anti-environment	Personal Values/Attitudes Judgments	Indirect Influence Model, bidirectional (Petit & Knobe 2009)

is a pictorial representation of the structural model. A *structural model* is the set of equations that represent the relationships among the variables in the model. So, a path diagram is a pictorial representation of the relationships among the variables in the model. Each structural path in a path diagram is meant to represent a causal relationship between two variables. Associated with each structural path is a path coefficient. The *path coefficient* represents the magnitude of the relationship between the two variables a structural path connects, controlling for the relationships of the other variables in the model. If certain assumptions are met, namely that the structural model has not mis-specified the causal relationships among the variables, the path coefficients can be interpreted as representing the strength of the causal relationship an independent variable has on a dependent variable, controlling for the effects of all the other variables in the model. The interpretations of these causal relationships are also expressed in the vocabulary of the path coefficient representing the *unique* causal contribution (or effect) of an independent variable on a dependent variable or in the vocabulary of the path coefficient representing the causal contribution (or effect) of an independent variable on a dependent variable, *above and beyond the causal contribution (or effects) of all other variables on in the model*. These path coefficients can be represented in *standardized* or *unstandardized* terms. If the path coefficient is a *standardized path coefficient*, then the path coefficient can be interpreted as the expected change in the dependent variable per standard deviation unit change in the independent variable, controlling for all the other variables in the model. If the path coefficient is an *unstandardized path coefficient*, then the path coefficient can be interpreted as the expected change in the dependent variable per unit change (using the original units in which the variables were measured) in the independent variable, controlling for all the other variables in the model.

So far all this terminology should sound familiar to those familiar with standard

regression. In one sense, structural path analysis just is a set of regression formulas. So, if you are familiar with regression and if it helps you to think about structural path analysis in terms of regression analyses, it is perfectly acceptable to interpret the path coefficients just as you would interpret regression coefficients. However, SEM does have some advantages over basic regression analyses. First, SEM allows the modeling of relationships among latent variables. *Latent variables* are variables that are not directly observed, but are variables that underlie the production of directly observable variables. Latent variables are not being utilized in any of the analyses reported or conducted in this paper, so a more in-depth explanation or understanding of the nature of latent variables is not needed. Second, SEM allows for the modeling of more complex relationships than what standard regression allows. Third, SEM provides a method to evaluate the 'fit' of a structural model against the data. Every structural model implies a specific covariance matrix. A *covariance matrix* is simply an unstandardized correlational matrix, and thus represents the relationships of associations among all the variables in the model. To test the fit of a structural model, the structural model's implied covariance matrix is compared to the covariance matrix of the actual data set. There are a number of different fit indices. *Fit indices* are indices of how well the structural model's implied covariance matrix fits the covariance matrix of the actual data set. Or, in other words, the fit indices are indices of whether or not, or the extent to which, the structural model's implied covariance matrix varies from the covariance matrix of the actual data set. The most widely reported fit index is the X^2 statistic. A significant X^2 statistic implies that the structural model's implied covariance matrix varies from the covariance of the actual data set with statistical significance. However, there remain various debates in the literature concerning the use of the X^2 statistic as an absolute arbitrator between those models that have "good fit" and those that do not. Often other fit indices are used as

evidence of good fit, either as a supplement to the evidence provided by a X^2 statistic or sometimes in lieu of a X^2 statistic.

With this technical apparatus under our command, we are now in a position to be able to understand and interpret Sripada and Konrath's models. Figure 1 is a path diagram of the first model they tested. As predicted by Sripada and Konrath's Deep Self Model, the analysis showed that the two normative variables had no effect on intentionality judgments above and beyond the effects of the deep-self variables and the experimental manipulation. This can be seen by looking at the path coefficients associated with the structural paths from moral status judgments to intentionality judgments and from goodness/badness judgments to intentionality judgments. Neither of these path coefficients represents a statistically significant relationship. Thus, there is no evidence that the strength of these relationships differs from zero, when controlling for all the other variables in the model. In other words, there is no evidence of an effect of moral status judgments on intentionality judgments above and beyond the effect of the other variables in the model nor is there evidence of an effect of goodness/badness judgments on intentionality judgments above and beyond the effect of the other variables in the model. Additionally, the analysis showed that the deep-self variables and the experimental manipulation had significant effects on intentionality judgments above and beyond the effects of all the other variables in the model.

The final model Sripada and Konrath tested was their hypothesized Deep Self Model (the base model minus the paths between the normative variables and intentionality judgments) plus the addition of a single empirically driven modification (an added path between goodness/badness judgments and chairman values/attitudes judgments). Based on all reported fit indices, their final model fit the data well, thus providing additional support for the two major

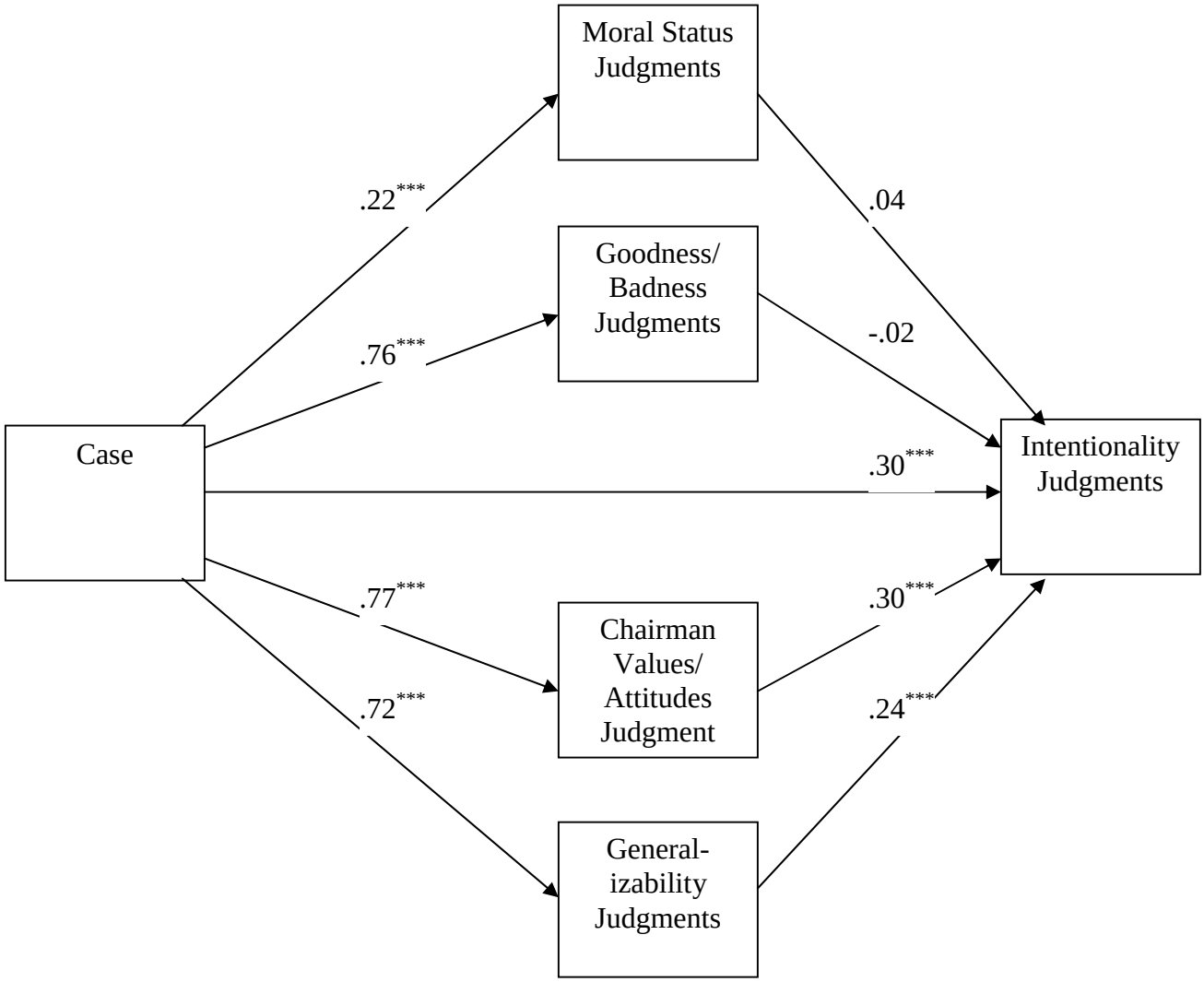

Figure 1. Sripada and Konrath's (forthcoming) base model.
Note: All path coefficients are expressed in standardized units. * p< .005, ** p <.01, *** p < .001.

hypothesis of the Deep Self Concordance Model—that is, for the hypothesis that deep-self judgments are a major driving force behind the pattern of intentional action judgments and for the hypothesis that normative judgments do not have an impact on intentional action judgments above and beyond the impact of deep-self judgments. See Figure 2 for the path diagram of their final model.

Sripada and Konrath followed this study with a second study that examined what people think drives the asymmetric intentionality responses in the chairman case. This second study was motivated by the fact that although Sripada and Konrath "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[,] … based on the experience of [Sripada] in presenting the Deep Self 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 Model far less intuitive" (Sripada & Konrath, forthcoming, draft p. 15).

Of the participants who were a part of Sripada and Konrath's second study, 76% gave open-ended responses that implicated normative variables as the probable cause of the asymmetry while 7% gave open-ended responses that implicated deep-self variables as the probable cause of the asymmetry. The remainder gave responses that were classified as not falling into either category. In light of these results, Sripada and Konrath concluded that their participants, along with audience members at their talks, were experiencing a *tracking problem* when trying to explain the reasons why the asymmetry arises. To justify this tracking problem explanation, Sripada and Konrath paid homage to Nisbett and Wilson's (1977) influential

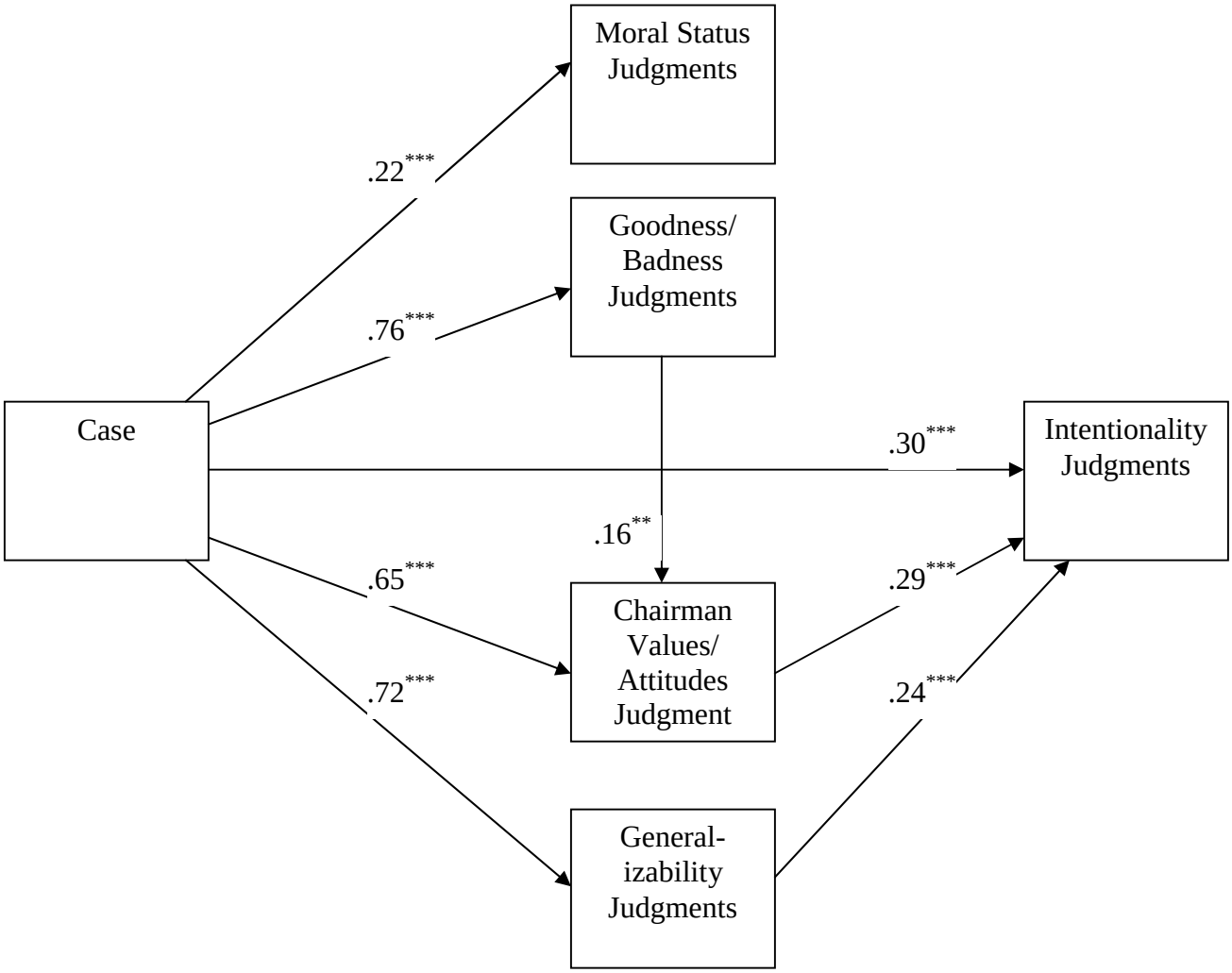

Figure 2. Sripada and Konrath's (forthcoming) final model.
Note: All path coefficients are expressed in standardized units. * p< .005, ** p <.01, *** p < .001. Overall model fit, $\chi^2(6) = 12.00$, p= .06; NFI = .985; NNFI = .981; CFI = .992; RMSEA = .065.

article “Telling more than we can know: Verbal reports on mental processes.” Nisbett and Wilson’s review of hundreds of studies provided extensive evidence that even though people are generally able to accurately state 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. Mis-implications of certain judgments tend to occur when (1) the critical features that drive the judgment are 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 met, people have a greater tendency to implicate the feature that is high in salience even if this highly salient feature plays no role in the actual judgment outcome, while ignoring the low-salient feature.

Sripada and Konrath see the chairman case as clearly meeting these features. First, the manipulation of the moral valance of the scenario is highly salient. Second, the attributions of dispositions, such as the attribution of underlying values and attitudes and behavioral dispositions, are often spontaneous, implicit, and automatic (see Uleman, et al., 2008, for a recent review of the spontaneous trait inferences literature). To help avoid the tracking problem when theorizing and when testing theories concerning complex judgment processes such as those that may appear in the side-effect effect, Sripada and Konrath recommend that more of the field move toward the more sophisticated statistical tools of Structural Equation Modeling, since SEM is particularly well suited for separating out unique influences of multiple variables on some given (set of) dependent variable(s).

Going forward, I will argue that, even though it may appear that Sripada and Konrath have provided strong evidence that deep-self judgments play a significant role in the asymmetric

pattern of intentionality judgments in Knobe's chairman case, while normative judgments play no (or at most an irrelevantly small) role in the asymmetric pattern of intentionality judgments in Knobe's chairman case, Sripida and Konrath's experimental evidence simply does not justify the conclusion that normative judgments are playing no role in the production of intentionality judgments in Knobe's chairman case. Rather, I will argue that a more plausible version of a deep-self model is one that is explicitly bidirectional. That is, I will argue that a more plausible deep-self model is one that admits that both deep-self judgments and normative judgments play an important role in the production of judgments of intentional action.

In the next two sections, I outline two methodological issues with Sripada's attempt to empirically test his model (Sripada and Konrath, forthcoming). The first issue, taken up in Section 4, involves the problem of implicitly modeling interaction terms. The second issue, taken up in Section 5, involves the problem of using manipulation checks as mediators. The dialectical role of these sections will be two-fold: First, the problems shed doubts on Sripada's empirical findings and his interpretation of his findings, hence raising doubts concerning the empirical support for his theoretical model. Second, the problems raised are meant to motivate an empirical re-analysis of his model using corrected methodology. In Section 6, I offer a sketch of a plausible bidirectional deep-self model, and then demonstrate that once the methodological issues raised in Sections 4 and 5 are corrected, the evidence favors the view that, at least in some cases, normative judgments are influencing judgments of intentional action above and beyond the influence of deep-self judgments. In other words, the evidence favors the predictions made by a bidirectional deep-self model over Sripada's Deep Self Model. In Section 7, I present the results of a new experiment that was designed to test additional predictions of the kind of bidirectional deep-self model that I outline in Section 6. In Section 8, I conclude with a brief discussion

concerning how the results reported in this paper may shed light on the status of the above-mentioned philosophically relevant debates about people's judgments of intentional action.

4 THE PROBLEM OF MODELING IMPLICIT INTERACTION TERMS

*** TECHICAL DISCUSSION WARNING: This section involves either a fairly technical discussion of methodological and statistical issues or the reporting of various statistical results. While it is not necessary to understand all the technical methodological and statistical points, understanding these points can help the reader have a greater appreciation and understanding for what is going. Additionally, in order to keep the results as transparent as possible, I am not relegating the reporting of the more technical aspect of the statistics to footnotes and appendices. Those without expertise in statistical interpretation should not be off-put or overwhelmed by the reporting of these statistics. Rather, one without expertise in statistical interpretation can gloss over the more technical aspects of the reporting and simply focus on the English interpretations of the statistics. Subsequent section that contain technical discussions of methodological or statistical points will be marked by an asterisk. ***

The problem highlighted in this section might be summarized in a very general form as follows: The way Sripada and Konrath use the chairman's values/attitudes judgments variable in their model does not appear to be the best way to conceptualize the variable given their theoretical model. Their conceptualization and use of the chairman's values/attitudes judgments variable causes a mismatch between their theoretical model and their statistical model. In the last subsection of this section, I demonstrate that when using a conceptualization of the chairman's values/attitudes judgments variable that better matches the theoretical model, my replication of Sripada and Konrath's model no longer fits the data as well. This in turn provides some evidence against Sripada and Konrath's theoretical model.

Before continuing, it will be helpful to get a little bit of terminology out of the way regarding path coefficients and fit indices, both of which were briefly defined above (pp. 12-13), and mediator variables and moderator variables. A *mediator variable* is a causal intermediary variable through which the effect of an independent variable on a dependent variable can be accounted for or explained. *Mediation* is said to occur to the extent that the mediator variable can account for or explain the effect of the independent variable on the dependent variable. *Partial mediation* occurs when the mediator partially accounts for or explains the effect of an independent variable on a dependent variable. *Full mediation* occurs when the mediator fully accounts for or explains the effect of an independent variable on a dependent variable.

A *moderator variable* is a variable that affects the direction or strength of the relationship between an independent variable and a dependent variable. *Moderation effects* are also known as *interaction effects*. When demonstrating moderation in a regression, SEM, or other similar frameworks, the standard methodology requires one to create an *interaction term*, which, in the case of linear moderation effects, is created by multiplying the moderator variable and the independent variable together and then using this product as the newly formed interaction term. Moderation and mediation are two very distinct types of effects. In spite of their conceptual distinctness, moderation and mediation are often confused in much of the social science literature. Confusing moderators and mediators can lead to serious problems with one's theoretical model, statistical model, or interpretation of the models (depending at which point the conceptual confusion occurred). (See Baron & Kenny, 1986 for a detailed discussion of the moderator-mediator distinction, along with a discussion of standard methodology for testing moderation and mediation; see Shrout & Bolger, 2002 for a detailed discussion of mediation and various ways mediation effects can be measured.)

With this terminology out of the way, let's consider the relevance of these issues to the current discussion. The primary claim of this section is that Sripada and Konrath recoded the chairman values/attitudes judgments to implicitly reflect an effect of moderation, instead of explicitly modeling a moderation effect. Because of this decision to use a recoded variable, the chairman values/attitudes judgments, which theoretically is meant to be a moderator variable, ends up being treated as a mediator variable in the statistical model. Furthermore, this recoding of the chairman values/attitudes judgments leads to several interpretative and statistical difficulties. However, even ignoring the specific interpretative and statistical inaccuracies that I point out below, Sripada and Konrath's decision to use a variable that implicitly reflects a moderation effect remains problematic, if for no other reason than this recoding caused a variable that is theoretically a moderator to be treated as a mediator in their statistical model.

Given Sripada and Konrath's explanation of the application of the Deep Self Model to the chairman case, the path coefficient that represents the relationship between case and chairman values/attitudes judgments appears not to support part of this explanation.

> Applying the Deep Self 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 toward the environment. (Sripada and Konrath, forthcoming, draft p. 6)

Here, Sripada and Konrath predict that there should be little-to-no effect of case on chairman values/attitudes judgments, yet in their statistical model, the path coefficient associated with the relationship between case and chairman values/attitudes judgments is representative of a large effect. However, this seemingly contradictory result may not be due to a failure of prediction; rather it may be due to the fact that Sripada and Konrath recoded the chairman values/attitudes judgments variable to implicitly reflect a moderation effect (instead of explicitly modeling a moderation effect). Here is there explanation of their decision:

> The predictions of the Deep Self 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 Model predicts correlations of *opposite directions* in the two conditions. That is, according to the Deep Self 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 intentionality helped the environment. (note 8)

Here, Sripada and Konrath's application of the Deep Self Model to the chairman case predicts that case moderates the effect of chairman values/attitudes judgments on intentionality judgments. This should be clear, as Sripada and Konrath explicitly state that the direction of the

effect of chairman values/attitudes judgments on intentionality judgments depends on level of case (case having two levels: harm, help). This moderation claim is statistically equivalent to the claim that chairman values/attitudes judgments moderate the effect of case on intentionality judgments.[5] If we expressed the moderation effect in this way, we would want to say that the direction (or strength) of the effect of case on intentionality judgments depends on level of chairman values/attitudes judgments (chairman values/attitudes judgments having continuum of levels ranging from 1 to 7). To put this a little more concretely, under the predictions of Sripada and Konrath's model, we would say that when people attribute anti-environmental values and attitudes to the chairman, they should be more likely to attribute intentionality to the chairman when the chairman harms the environment than when he helps the environment. On Sripada's account, this expectation would be because anti-environmental attitudes and harming the environment concord. On the other hand, when people attribute pro-environmental values and attitudes to the chairman, they should be more likely to attribute intentionality to the chairman when the chairman helps the environment. Again, on Sripada's account, this expectation would be because pro-environmental attitudes and helping the environment concord.[6] Instead of including an explicit interaction term in their model, Sripada and Konrath recoded the chairman

[5] This fact of equivalence is important for the purposes of this paper, as in Sections 6 & 7, I will choose to conceptualize the moderation effect as chairman values/attitudes judgments moderating the effect of case on intentionality judgments. Given that these claims are equivalent, nothing substantial rides on the choice of the ways to express the relationship. However, given what I show in Sections 6 & 7, the interpretation becomes a bit easier if the moderation is expressed in this manner. However, the exact same effect – and problems for Sripada's account – could be shown if I chose to express the moderation effect as case moderating the effect of chairman values/attitudes judgments on intentionality judgments.

[6] Additionally, this moderation effect strictly implies that when people attribute ambivalent environmental values and attitudes to the chairman, there should be no difference in intentionality attributions by case.

values/attitudes judgments to reflect the predicted moderation effect. The choice not to use an explicit interaction term leads to both interpretive and statistical inaccuracies.

On the interpretative side, one of the problems this recoding creates is the production of a model that is incapable of straightforwardly capturing the hypothesized relationship between case and chairman values/attitudes judgments. Sripada and Konrath's statistical model makes it appear as if there is a strong relationship between case and chairman values/attitudes judgments; however, their explanation of the application of their theoretical model to the chairman case predicts that there should be *little to no* effect of case on chairman values/attitudes judgments, since the chairman 'does not care at all' about the effects on the environment in either case. That is, their model predicts that people should respond that he has negative attitudes towards the environment in both cases. By using a recoded chairman values/attitudes judgments variable, there is no straightforward way to interpret the path coefficients such that one can discern the relationship between case and chairman values/attitudes judgments.

Another interpretative difficulty appears in Sripada and Konrath's final model in which they model a path between goodness/badness judgments and the recoded chairman values/attitudes judgments. It is not even clear what this structural path would represent. Statistically, this method of modeling the paths is not the same as modeling a path between goodness/badness judgments and an explicit interaction term. Furthermore, theoretically, this method of modeling the paths is not the same as modeling a relationship between goodness/badness judgments and a non-recoded chairman values/attitudes judgments. Finally, just to reiterate, the most significant problem with Sripada and Konrath's use of the recoded goodness/badness judgment is that the variable ends up being treated as a mediator in the model

when the variable is described as a moderator in their theoretical model. This confusion can lead to numerous conceptual and interpretive flaws (Baron and Kenny, 1986).

On the statistical side, the model with the implicitly recoded values/attitudes judgments variable is not statistically equivalent to a model that includes an explicit interaction term. This means the model fit will probably differ between the two models.

To test the hypothesis that the model fit between the two models differs, I will be using new data that I collected as part of a larger study that seeks to explore various relationships among normative judgments and folk psychological judgments. Specifically, I will run two sets of structural path analyses. The first analysis will be a replication of Sripada and Konrath's model as they ran it (i.e., with a recoded chairman values/attitudes judgment variable that implicitly reflects the moderation effect). The second analysis will be a replication of Sripada and Konrath's model using an explicit interaction term. I predict that a model using an explicit interaction term will better capture the predictions of Sripada and Konrath's explanation of their application of the Deep Self Model to the chairman case. That is, I predict that the path coefficient for the structural path that represents the effect of case on chairman values/attitudes judgments will not be statistically significant or will only be representative of a very small effect size, and I predict that the path coefficient representing the moderation effect will be significant. Furthermore, I predict that the model that uses the explicit interaction term will perform worse on all fit indices when compared to the model that uses the recoded chairman values/attitudes judgments. If these predictions are right, this would leave Sripada with the choice of choosing between (a) a model that more accurately captures specific predictions about certain paths but performs worse on fit indexes or (b) a potentially problematic model that cannot straightforwardly capture all the specific predictions about certain paths but performs better on

fit indexes. In other words, if these predictions are right, they would show that when an improved-theory-matching methodology is used, Sripada's model is not as well supported (via worse performing fit indices).

4.2 Methods

Participants (n = 699) were students at Georgia State University enrolled in a critical thinking course. For their participation, participants were offered extra credit at the instructor's discretion. All participants were randomly assigned to one of four vignette conditions, including the two vignette conditions used by Sripida and Konrath, as well as two other conditions to be discussed in Section 6. Following the presentation of the vignette, participants were presented with a series of questions, a subset of which were the same questions Sripida and Konrath used in their study. See Table 1 for the questions used in Sripada and Konrath's original study. In order to eliminate potential order effects and carry-over effects, a Williams design (Williams, 1949), which is Latin Square randomization balanced for first-order carry-over effects, was employed. Participants who did not finish the survey or who missed a simple comprehension check had their data dropped from analysis, resulting in n = 551. Only data from the two conditions used in Sripida and Konrath's study were used in the current analysis, resulting n = 299. All surveys were conducted online using standard survey software.

4.3 Results and Discussion*

Two structural path analyses were run. In the first analysis, Sripada and Konrath's base model was run as they ran it. That is, the chairman values/attitudes judgments were recoded to implicitly reflect the interaction. In the second analysis, a model analogous to Sripada and Konrath's base model was ran using a non-recoded chairman values/attitudes judgments variable along with an explicit case x chairman values/attitudes judgments interaction term. The

interaction term was created by multiplying together the dummy-coded case variable by the chairman values/attitudes judgments. See Figure 3 for the path diagram of the model that used the recoded chairman values/attitudes judgment variable. See Figure 4 for the path diagram of the model that used the explicit interaction term.

The statistical model with the explicit interaction term better matches the predictions of Sripada and Konrath's explanation of the application of the Deep Self Model to the chairman case. As predicted by Sripada and Konrath, there were no differences in chairman values/attitudes judgments between the harm and help conditions. Furthermore, as predicted by the Deep Self Model, there is a significant interaction between case and chairman values/attitudes judgments. However, the model with the explicit interaction term performed worse on all the available fit indices.[7] In other words, once an improved-theory-matching methodology is used, Sripida and Konrath's model no longer fits the data as well. See Table 2 for a summary of fit indices for both models.

However, this result simply sheds doubt concerning overall model fit. While a model failure can be seen as a challenge to the collection of hypotheses embedded within the model, it does not necessarily provide strong evidence against any specific hypothesis embedded in the model. Keep in mind that the Deep Self Model has two primary hypothesis: (1) deep-self variables are of primary importance in the production of judgments of intentional action, and (2) normative variables do not have an effect on judgments of intentional action above and beyond the effects of the deep-self variables. The second of these two hypotheses is the one that directly challenges the descriptive bidirectional account, and it accounts for Sripida's model being appropriately considered a unidirectional account. In the next section of this thesis, I outline a

[7] Similar results were found when comparing Sripida and Konrath's final model to its two closest corrected analogs.

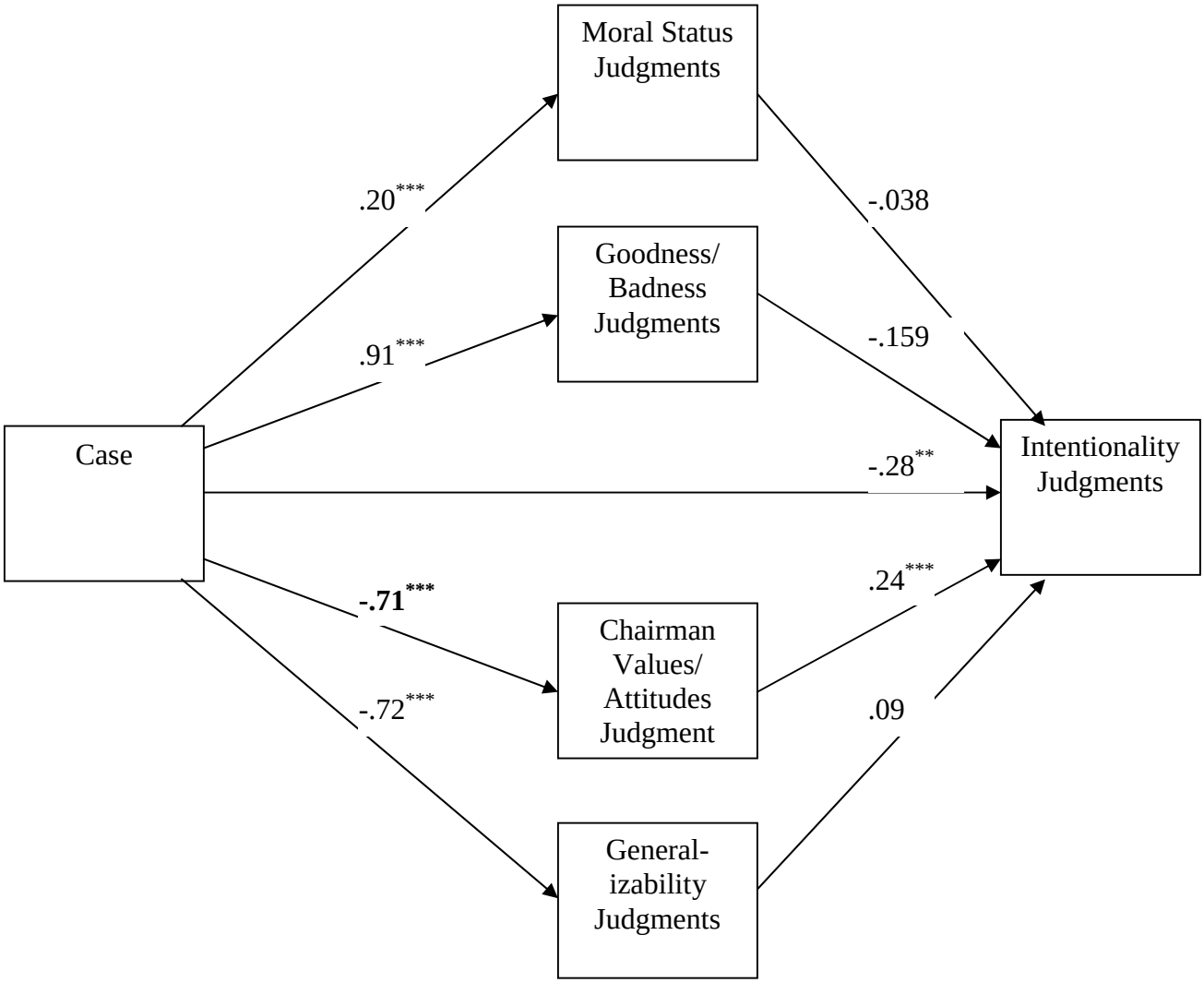

Figure 3. Replication of Sripida and Konrath's base model using the recoded chairman values/attitudes judgments to reflect an implicit interaction term.
Note: Notice that the path coefficient associated with the structural path case to chairman values/attitudes judgment is statistically significant and representative of a large effect. All path coefficients are expressed in standardized units. * p< .005, ** p <.01, *** p < .001. Overall model fit, $\chi^2(6) = 17.97$, p = .016; CFI = .990; TLI = .975: RMSEA = .082; SRMR = .020.

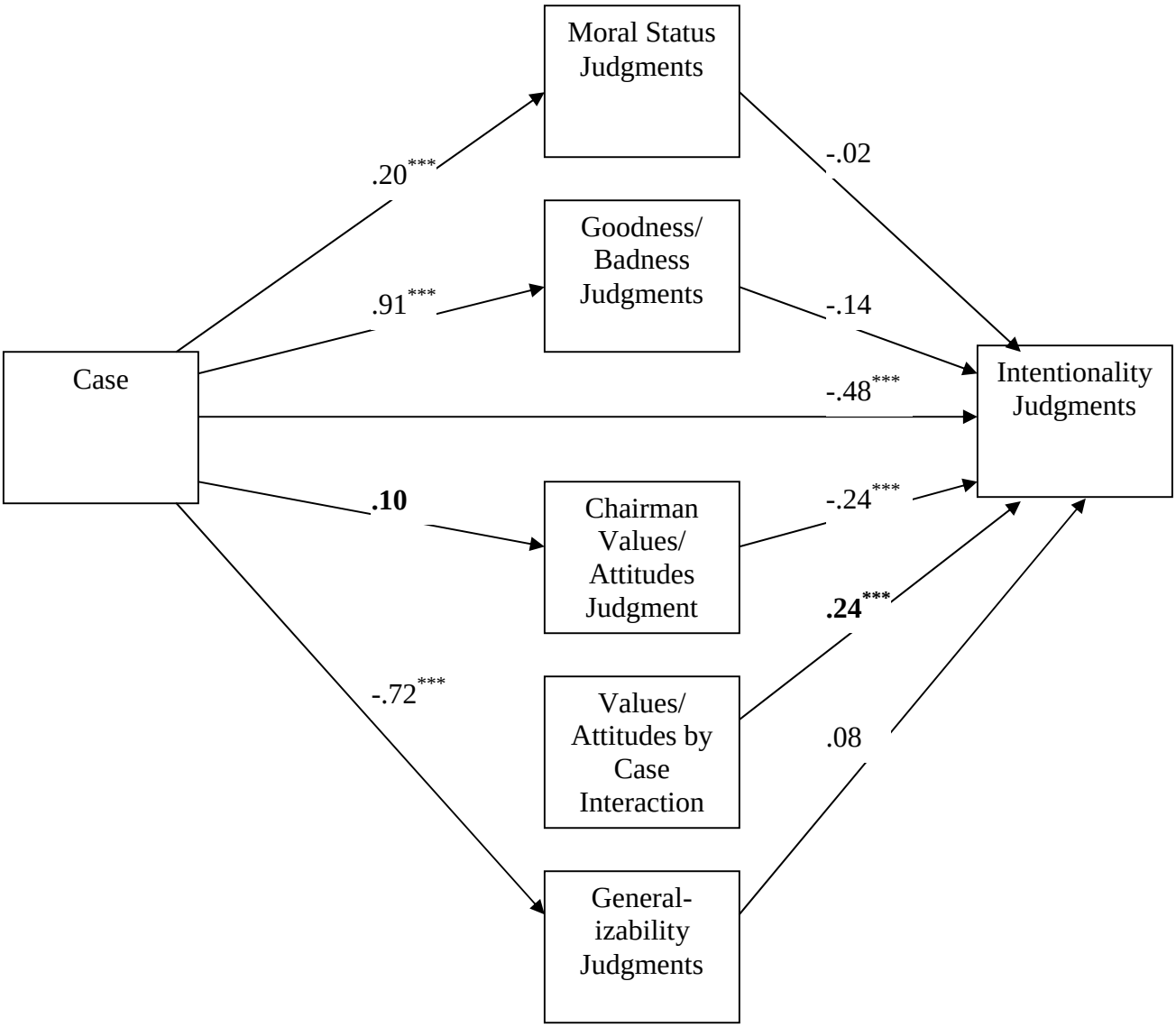

Figure 4. Replication of Sripida and Konrath's base model using an explicit interaction term. Note: Notice that the path coefficient associated with the structural path case to chairman values/attitudes judgment is *not* statistically significant. Also notice that the structural path values/attitudes by case interaction to intentionality judgments is statistically significant. Both of these results would be predicted based on Sripada's theoretical model. All path coefficients are expressed in standardized units. * p< .005, ** p <.01, *** p < .001. Overall model fit, $\chi^2(9) = 49.45$, p < .001; CFI = .967; TLI = .926: RMSEA = .123; SRMR = .056. Interaction term was allowed to covary freely with both of its components.

Table 2. Summary of fit indices for the replications of Sripida and Konrath's base model using the recoded chairman values/attitudes judgments to reflect an implicit interaction term and the analogous model using an explicit interaction term.

Fit Index	Implicit Interaction Term	Explicit Interaction Term
Chi-Square Test of Model Fit	$X^2(6) = 17.972$, $p = .006$	$X^2(9) = 49.452$, $p < .001$
CFI	.990	.967
TLI	.975	.926
AIC	5414.752	5987.269
BIC	5485.060	6076.079
RMSEA	.082	.123
SRMR	.020	.056

Note: What is important here is the comparative fit between the two models. Lower scores are signs of better fit for the Chi-Square Test of Model Fit, the AIC, the BIC, the RMSEA, and SRMR. Higher scores are signs of better fit for the CFI and TLI.

second methodological issue with the way Sripada and Konrath empirically tested their theoretical model—the problem of using manipulation checks as mediators. This second methodological issue speaks directly to the strength of the evidence that Sripada and Konrath's model provides for the hypothesis that normative judgments are not influencing judgments of intentional action.

5 THE PROBLEM OF USING MANIPULATION CHECKS AS MEDIATORS*

The primary claim of this section is that the observed variable goodness/badness judgments conceptually is a *manipulation check*. Here I use the term 'manipulation check' to refer to any observed variable that acts as a measure of some conceptual variable whose variation is (thought to be) a constitutive part of the experimental manipulation. In other words, it is a measurement of what the experimental manipulation is thought to be manipulating. This use of manipulation check is also referred to as an 'independent variable check' (Sigall and Mills, 1998). Because the observed variable goodness/badness judgments is a manipulation check, the causal information provided by the goodness/badness judgments variable and the causal information provided by the manipulation of case are, in large part, redundant. Given this redundancy, it should not be expected that there should be a unique effect of goodness/badness judgments on intentionality judgments, above and beyond the effects of all the other variables in the model whenever the manipulation of case is included as one of the variables in the model. However, this failure of expectations to find a unique effect has little to nothing to do with the role of the deep-self variables; rather it has much more to do with the fact that the causal information provided by the observed variable goodness/badness judgments and the causal information provided by the manipulation of case are, in large part, redundant. This redundancy

occurs because the variation of goodness/badness judgments is a constitutive part of the manipulation of case.

For our current purposes, it may be helpful to review the general form of Sripada and Konrath's base model. In the base model, there is a dichotomous experimental manipulation. The experimental manipulation involves giving participants a vignette in which the only thing that changes is the moral valence of a foreseen side effect of an agent's action—that is, the manipulation of whether the environment was harmed or helped as the result of the chairman's decision to implement a profit-generating policy. In the base model there are also four variables that are construed as mediator variables.[8] The experimental manipulation along with the four potential mediator variables all act as potential predictors of intentionality judgments. One of the potential mediator variables is goodness/badness judgments. The goodness/badness judgment variable is a rating of the participants' judgments about the moral valence of the side effect. It asks whether the harming [helping] of the environment is good or bad. With this in mind, the goodness/badness judgments conceptually just is a manipulation check. After all, the experimental manipulation was designed to manipulate the perceived moral valence of the side effect, and the goodness/badness judgments is meant to be a measurement of whether participants perceived the moral valence to have been manipulated. The other three potential mediator variables are ratings about various perceptions about the agent—the values and attitudes of the agents, the behavioral dispositions of the agent, and the moral status of the agent.

[8] As pointed out in Section 4, Sripida and Konrath's recoded chairman values/attitudes judgments played the role of a mediator in their statistical model even though the variable should have been treated at a moderator. For the specific purposes of this section nothing of importance rides on forcing this correction upon Sripida and Konrath's model, so for present purposes, I will refer to the variables as they were used in Sripida and Konrath's model.

In Sripida and Konrath's base model, the path coefficient associated with the structural path goodness/badness judgments to intentionality judgments does not statistically differ from zero. This result leads Sripida and Konrath (forthcoming) to conclude, "This pattern is consistent with the hypothesis that goodness/badness judgments … are causally influenced by case, *but goodness/badness judgments … are not themselves causes of intentionality judgments*" (emphasis original).

However, this conclusion is not warranted. The reason goes back to the fact that, conceptually, the goodness/badness judgments variable just is a manipulation check. When using a proximal mediator—a manipulation check is the prototypical example of the most proximal of proximal mediators[9]—it is common to receive results that lead to a large path coefficient for the path that represents the relationship between the independent variable, in this case the experimental manipulation of case, and the proximal mediator, in this case goodness/badness judgments. Further, using a proximal mediator generally leads to a very small or null coefficient for the path that represents the relationship between the proximal mediator and the outcome variable. This leads to very small power to detect genuine indirect effects (Hoyle and Kenny, 1999). This decreased power is generally caused by inflated standard errors due to high multi-collinearity. These statistical results are common for proximal mediators in both experimental and non-experimental research; however, the problem of using a manipulation check as a mediator goes beyond these commonly noted problems associated with using proximal mediator.

[9] Though manipulation checks are often given as examples as a proximal mediators, and I follow suit here, it may not be accurate to refer to manipulation checks as proximal mediators. However, an in-depth discussion of these issues is well beyond the scope of this current paper. For our current purposes, I will stick with the convention of classifying manipulation checks as a type of proximal mediator.

Recall that the path coefficients represent the *unique* effect of an independent variable on a dependent variable, *above and beyond the effects of all other variables in the model*. In Sripada and Konrath's base model, the path coefficient that is associated with the structural path goodness/badness judgments to intentionality judgments represents the unique effect of goodness/badness judgments on intentionality judgments, above and beyond the effects of all the other variables in the model. Given the results of Sripada and Konrath's model, it may be accurate to say that there is no unique effect of goodness/badness judgments on intentionality judgments above and beyond the effect of the experimental manipulation and the effects of the other potential mediators. However, considering that the measurement of goodness/badness judgments conceptually just is a manipulation check, there should not be an expectation of a unique effect of goodness/badness judgments on intentionality judgments above and beyond the effects of all the other variables in the model whenever all the other variables *include the experimental manipulation*. This failure of expectations occurs because the causal information conveyed by the goodness/badness judgments variable and the causal information conveyed by the manipulation of case are, in large part, redundant. This redundancy occurs because the variation of goodness/badness judgments is a constitutive part of the experimental manipulation. That is, the experimental manipulation is a manipulation of goodness/badness judgments.

If my claims above are correct, then several empirical results should follow. First, if the experimental manipulation is acting as a successful manipulation of goodness/badness judgments (and if the observed variable goodness/badness judgments is an accurate measurement of this manipulation), then we should expect to find a strong positive correlation between the manipulation of case and goodness/badness judgments. Second, if the goodness/badness judgments are acting as a manipulation check of the experimental manipulation, we should be

able to demonstrate that the goodness/badness judgments provide little to no causal information above and beyond the causal information provided by the effect of the experimental manipulation on intentionality judgments. This could be demonstrated by showing that the unique effect of goodness/badness judgments on intentionality judgments should be dramatically reduced when *only* the experimental manipulation is added to the model as a predictor of intentionality judgments.

I predict that the statistical results will bear out the above empirical predictions. That is, I predict that the correlation between case and goodness/badness judgments will be strong and positive. Furthermore, I predict that the unique effect of goodness/badness judgments on intentionality judgments will be dramatically reduced when only the experimental manipulation is added to the model as a predictor of intentionality judgments.

5.2 Methods

The methods used here are the same as the methods mentioned in Section 4.2.

5.3 Results and Discussion*

In order to test the hypothesis that experimental manipulation is acting as a successful manipulation of goodness/badness judgments (and the observed variable goodness/badness judgments is an accurate measurement of this manipulation), a point-biserial correlation was computed. There was a strong, positive correlation between the experimental manipulation and goodness/badness judgments, $r_{pb} = .914$, $p < .001$.

In order to test the hypothesis that goodness/badness judgments provide little to no causal information above and beyond the causal information provided by the experimental manipulation, a two-step hierarchical regression was conducted. In the first step, goodness/badness judgments was entered into the regression formula as the sole predictor of

intentionality judgments. In the second step, the experimental manipulation was added to the regression formula as a second predictor of intentionality judgments. The first step yielded the regression formula $y = -.581x_1 - 0$, $F(1, 297) = 213.380$, $p < .001$, $R^2 = .418$. Goodness/badness judgments had a significant relationship with intentionality judgments such that, on average, the more one judged the side effect to be good, the less one judged the side effect to have been brought about intentionally, $\beta = -.647$, $p < .001$. The second step yielded the regression formula $y = -.196x_1 - 2.093x_2 + 1.036$, $F(2, 296) = 123.156$, $p < .001$. The addition of the experimental manipulation to the model improved overall model fit, $\Delta R^2 = .036$, $p < .001$. Goodness/badness judgments retained a significant effect on intentionality judgments in the same direction as the first step, but just marginally so, $\beta = -.218$, $p = .041$, $r_{sp} = -.088$. For the purposes of the current hypothesis, it is important to notice that the unique effect of goodness/badness judgments, in standardized terms, dropped from $\beta = -.647$ to $\beta = -.218$ (and semi-partial correlations dropped from -.647 to -.088). Due to error, differences in scale, and other factors, one should not necessarily have expected that the inclusion of the experimental manipulation to decrease the unique effect of goodness/badness judgments on intentionality judgments to a nil effect. However, the fact that the experimental manipulation alone reduced the unique effect of goodness/badness judgments on intentionality judgments from a very large effect to a near-nil effect supports the hypothesis that the causal information provided by the goodness/badness judgments and the information provided by the experimental manipulation are, in large part, redundant. This in turn provides empirical support for the fact that the observed goodness/badness judgments variable is acting as nothing more than a manipulation check.

These results motivate the removal of the observed variable goodness/badness judgments from an appropriately testable model (because the goodness/badness judgments variable and the

manipulation of case are redundant). Additionally, these issues motivate an interpretation of the effect of case on intentionality judgment that respects the fact the manipulation of case is a successful manipulation of goodness/badness judgments (because the manipulation of goodness/badness judgments is a constitutive part of the manipulation of case). In subsequent sections, when testing the predictions of the Deep Self Model, I will not include the observed variable goodness/badness judgments in any of the analysis. Additionally, the effect of the manipulation of case will be interpreted as a successful manipulation of goodness/badness judgments.

6 THE DEEP SELF MODEL IS STILL BIDIRECTIONAL

Despite the criticisms I have raised above, I think there is no denying the importance of deep-self judgments when making judgments of intentional action (and other folk psychological judgments). Often when making judgments of intentional action, we implicitly or explicitly ask ourselves: Is the agent the kind of person who would want to bring about such an outcome? Intuitions about this or similarly framed questions are then used as potential evidence that a person acted intentionally. For example, returning to the story of Elijah and Sasha presented at the beginning of this paper, when the caregivers are trying to determine if Elijah intentionally knocked down Sasha, the caregivers may, explicitly or implicitly, consider what they perceive to be the values and attitudes or the behavioral dispositions of Elijah. For instance, if one of the caregivers believes that Elijah has a very negative attitude toward girls playing in his play space, or if one of the caregivers believes that Elijah is in general an aggressive child, then it is more likely that the caregiver will conclude that Elijah intentionally knocked down Sasha.

However, it seems equally clear that the preponderance of evidence favors the conclusion that, in at least some cases, normative judgments are influencing judgments of intentional action

(and other folk psychological judgments). With this in mind, the best deep-self model of intentionality will be one that admits some degree of bidirectionality.

One plausible way a bidirectional deep-self model could be cashed out would be in terms of claiming that deep-self judgments can attenuate (or, perhaps, also intensify) the effects of normative judgments on judgments of intentional action. Roughly, such a model would be committed to something along the following lines: In absence of other information, normative judgments often do impact judgments of intentional action (and perhaps other folk psychological judgments). Additionally, deep-self judgments also directly impact judgments of intentional action. However, the effect of normative judgments on judgments of intentional action will be attenuated, or will perhaps even disappear, to the extent that other judgments that provide information that conflict with the normative-driven attributions of intentional action are made.

Bringing this rough sketch to bear on the chairman case, the story might go as follows: People are predisposed to attribute greater intentionality when they judge that there has been a norm violation. The normative violation in the chairman case is the harming of the environment. Thus, there should be a predisposition to attribute greater intentionality when the chairman harms the environment than when the chairman helps the environment. However, this predisposition may be attenuated, or possibly even overridden, if people make deep-self judgments that stand in opposition to the norm-violation-driven intentionality attribution. So, for instance, if the people judged that the chairman was truly a lover of the environment, then the initial norm-violation-driven effect may lessen, or perhaps even disappear. (I directly test this latter prediction in the next section.)

In this section, my goal is to briefly review some specific predictions Sripada's Deep Self Model is committed to. Then I want to compare these predictions to the predictions a

bidirectional deep-self model (like the one outlined above) might make. Then finally, I want to turn to testing these predictions, making the methodological corrections suggested by the discussion in Sections 4 and 5.

Several of the predictions of the Deep Self Model have already been rehearsed throughout the earlier sections of this paper. However, to get clear on the contrasting predictions that would be made by Sripada's Deep Self Model and a bidirectional deep-self model, it may be helpful to review a couple key predictions. First, in the chairman case, the Deep Self Model makes a very explicit and specific prediction concerning the interaction of case and chairman values/attitudes judgments on intentionality judgments. The Deep Self Model predicts that to the extent that people attribute anti-environmental attitudes to the chairman, intentionality judgments will be higher in the cases where the chairman harms the environment (and lower in cases where the chairman helps the environment), but also, to the extent that people attribute pro-environmental attitudes to the chairman, intentionality judgments will be higher in cases where the chairman helps the environment (and lower in cases where the chairman harms the environment). This prediction entails a "crossover" effect in which the interaction crosses over at the point in which people attribute ambivalent environmental attitudes (i.e., environmental attitudes that are neither anti- nor pro-environmental) to the chairman. In other words, when people attribute ambivalent environmental attitudes to the chairman, there should be no effect of case on intentionality judgments. Second, the Deep Self Model predicts that once the effect of the deep-self variables on intentionality judgments is accounted for, there should be no effect, neither direct nor indirect, of normative variables on intentionality judgments above and beyond the effects of deep self variables. This prediction reinforces the precision of the prediction that

when people attribute ambivalent environmental attitudes, there will be no effect of case on intentionality judgments.

These predictions of Sripada's Deep Self Model can be graphically represented by a pattern of simple slopes. *Simple slopes* represent the relationship of an independent variable with a dependent variable at different levels of a moderating variable. In our case, the simple slopes would be used to represent the relationship of the manipulation of case with intentionality judgments at different levels of chairman values/attitudes judgments. Given the coding scheme used in the experiment, Sripada and Konrath's model would predict a negative slope when anti-environmental attitudes are attributed to the chairman, no slope (i.e., a line with a slope of 0) when ambivalent environmental attitudes are attributed to the chairman, and a positive slope when pro-environmental attitudes are attributed to the chairman. See Figure 5 for the simple slope pattern predicted by Sripada's Deep Self Model.

A bidirectional deep-self model would make a very similar pattern of predictions. However, since a bidirectional deep-self model allows that normative judgments can have an effect on intentionality judgments above and beyond the effect of deep-self judgments, a bidirectional deep-self model would allow that, when ambivalent environmental attitudes are attributed to the chairman, intentionality judgments may remain higher in the harm case than the help case. Implied by this claim is that, under a bidirectional deep-self model, the crossover effect, if there is a crossover effect, may occur at some point in which people attribute pro-environmental attitudes to the chairman. If the strength of the effect of normative variables above and beyond the effect of the deep-self variables on intentionality judgments is strong, then the crossover would occur when the attributions of pro-environmental values and attitudes are high (or the crossover may not occur at all if the effect of the normative variables is strong enough). If

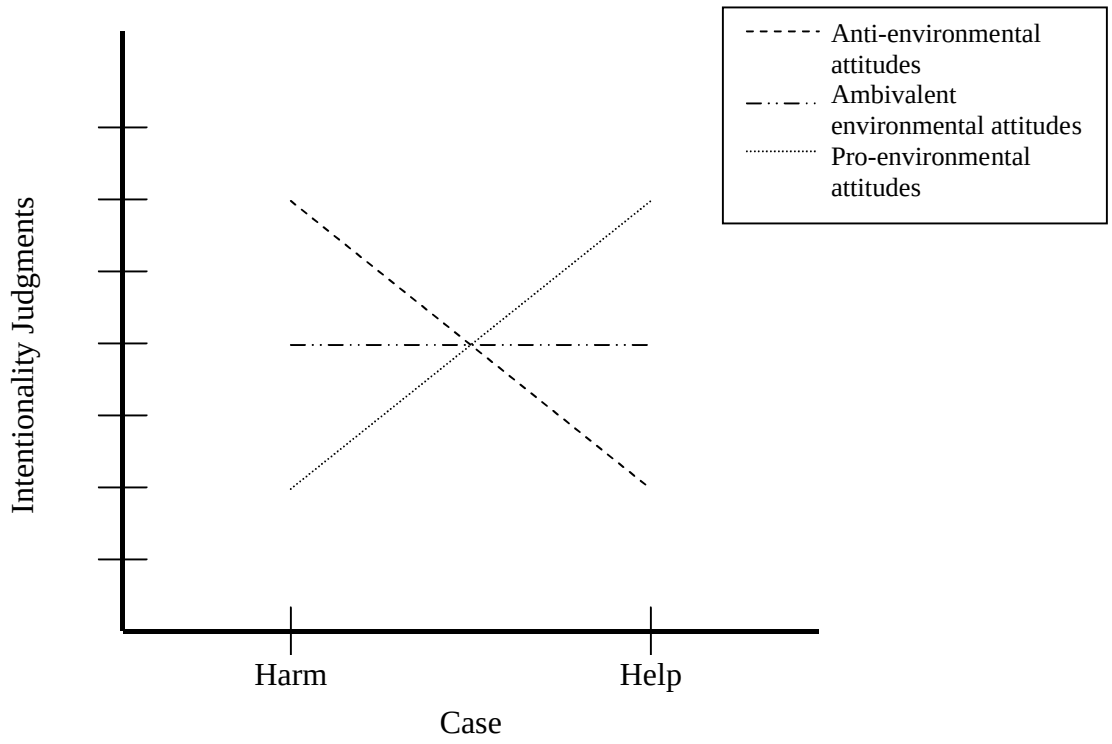

Figure 5. The Deep Self Model predicted simple slopes of intentionality judgments regressed on case at different levels of chairman values/attitudes judgments controlling for all the other variables in the model.

the strength of the effect of normative variables above and beyond the effect of the deep-self variables on intentionality judgments is weak, then the crossover would occur when the attributions of pro-environmental values and attitudes are low (but still on the pro-environmental side). If there is no effect of normative variables above and beyond the effect of deep-self variables on intentionality judgments, then the crossover would occur when ambivalent attitudes are attributed to the chairman (this is the prediction of Sripada's Deep Self Model).

To test these competing predictions, I will examine the exact nature of the case by chairman values/attitudes judgment moderation effect. That is, I will examine the relationship between case and intentionality judgments at various levels of chairman values/attitudes judgments. I predict that the results that will contrast with what should be predicted by Sripada's Deep Self Model, but will be consistent with the results predicted by a bidirectional deep-self model. More specifically, I predict that when people attribute ambivalent environmental attitudes to the chairman, people will be more likely to attribute intentionality to the chairman in the harm case than in the help case. Additionally, since the bidirectional deep-self model I have in mind predicts that the effect of normative judgments on intentionality judgments should not be attenuated in the chairman case, I also predict that, even as the attribution of pro-environmental attitudes to the chairman begin to become relatively strong, people will remain more likely to attribute intentionality to the chairman in the harm case than in the help case. Before continuing to the results, it should be kept in mind that the prediction that people will remain more likely to attribute intentionality to the chairman in the harm case than the help case when people attribute ambivalent environmental attitudes to the chairman is enough to provide a sufficient challenge to Sripada's Deep Self Model, as his model strictly implies that there should be no difference of intentionality attributions between the harm and help case when ambivalent environmental

attitudes are attributed. The second prediction concerning the pattern of intentionality attributions when people attribute pro-environmental attitudes to the chairman, if correct, will help make an even stronger case against Sripada's Deep Self Model, along with providing some initial plausibility for the type of bidirectional deep-self model outlined above.

6.2 Methods

The methods used here are the same as the methods mentioned in Section 4.2.

6.3 Results and Discussion*

See Table 3 for the descriptive statistics for intentionality judgments and chairman values/attitudes judgments.

In order to test the nature of the interaction effect of case and chairman values/attitudes judgments on intentionality judgments, a two-step hierarchical regression was conducted. In the first step, case (x_1) and chairman values/attitudes judgments (x_2) were entered as a predictor of intentionality judgments (y). Additionally, due to their theoretical importance for other models and in an effort to test the interaction effect in a manner that would be maximally consistent with Sripada and Konrath's model, behavioral generalization judgments (x_3) and moral status judgments (x_4) were included in the model. In an effort to facilitate accurate and meaningful interpretations, the variables chairman values/attitudes judgments, behavioral generalization judgments, and moral status judgments were all mean-centered (Cronbach, 1987). In the second step of the hierarchical regression, the interaction term case x (mean-centered) chairman values/attitudes judgments (x_5) was entered into the model as a moderator. The first step yielded the regression formula $-2.573x_1 - .116x_2 + .118x_3 - .080x_4$, $F(4, 294) = 62.629$, $p < .001$, $R^2 = .460$. The second step yielded the regression formula, $-2.638x_1 - .438x_2 + .098x_3 - .058x_4 + .616x_5$, $F(5, 293) = 56.256$, $p < .001$. The addition of the moderator improved overall model fit,

Table 3. Means for intentionality judgments and chairman values/attitudes judgments

	Harm Condition	Help Condition
Intentionality Judgments	4.93 (SD = 1.88)	1.95 (SD = 1.41)
Chairman Values/Attitudes Judgments	2.58 (SD = 1.24)	2.82 (SD = 1.28)

ΔR^2 = .030, p < .001, providing evidence of a significant moderation effect.

With the finding of the moderation effect, the next step is to test and plot simple slopes in order to get a better feel for the nature of the interaction effect (Aiken & West, 1991; Frazier, et al., 2004). In order to test the predictions of each model, I tested and plotted the simple slopes that represent the effect of case on intentionality judgments at different levels of chairman values/attitudes judgments. Specifically, I examined the effect of case on intentionality judgments at some value for which anti-environmental attitudes are attributed to the chairman, at some value for which ambivalent environmental attitudes are attributed to the chairman, and at some value for which pro-environmental attitudes are attributed to the chairman. Given that the average chairman values/attitudes judgments were anti-environmental (M = 2.70), I used average chairman values/attitudes judgments as the value chosen for anti-environmental. This choice may seem fairly arbitrary, but the exact value chosen for this slope really does not matter too much because neither model makes different predictions in the chairman cases when people attribute anti-environmental attitudes to the chairman.

Since the mid-point 4 on the 7-point scale used to measure chairman values/attitudes judgments represents *neither pro-environmental nor anti-environmental*, I used a rating of 4 as the value chosen for the attribution of ambivalent environmental attitudes.

For pro-environmental attitudes, I used the value 5.30. This choice was drive by two factors: a trivial, aesthetic reason and a substantive reason. The trivial reason why this value was chosen was because the chosen anti-environmental value was 1.30 units below the mid-point and the value of 5.30 is symmetrically 1.30 units above the midpoint. The substantive reason why 5.30 was chosen was because this value also represents attributions of moderately strong pro-environmental attitudes. Thus, this value allows us to test our second prediction, namely, that

people will remain more inclined to attribute intentionality in the harm case than in the help case even when people attribute moderately pro-environmental attitudes.

The tests reveal that when people attribute anti-environmental attitudes to the chairman, they are more inclined to attribute intentionality to the chairman in the harm case than in the help case, B = -2.638, $p < .001$, $r_{sp} = -.491$. Both models would predict this effect.

When people attribute ambivalent environmental attitudes to the chairman, they remain more inclined to attribute intentionality to the chairman in the harm case than in the help case, though to a slightly lesser degree than when they attribute anti-environmental attitudes to the chairman, B = -1.838, $p < .001$, $r_{sp} = -.235$.[10] This result is inconsistent with Sripada's Deep Self Model, as his model predicts no difference of intentionality judgments between cases when people attribute ambivalent environmental attitudes to the chairman. However, this result is perfectly consistent with a bidirectional deep-self model.

When people attribute pro-environmental attitudes to the chairman, they again remain more inclined to attribute intentionality to the chairman in the harm case than in the help case, though to a slightly lesser degree than when they attribute ambivalent environmental attitudes toward him, B -1.037, $p = .025$, $r_{sp} = -.130$. This result is clearly very problematic for Sripada's Deep Self Model, as his model predicts that when people attribute pro-environmental attitudes to the chairman, they should be more likely to attribute intentionality in the help case than in the harm case. However, this result are fully consistent with a bidirectional deep-self model that

[10] If for whatever reason one is not satisfied with the value of 5.40 being used to test the slope for the attributions pro-environmental attitudes and may want an even more pro-environmental value, I also tested the simple slope at a value of 6, which was the most pro-environmental rating any participant gave the chairman. Even when testing the slope using a value of 6, people remained more inclined to attribute intentionality in the harm case than in the help case.

holds that deep-self variables can attenuate (or intensify) the effects of normative variables on intentionality judgments. See Figure 6 for the simple slope pattern that was actually obtained.

7 THE CASE OF THE CARING CHAIRMAN

Almost all of the research on the side-effect effect has involved vignettes in which the actor expresses "not caring" about bringing about some outcome—an outcome that is meant to be recognized as a foreseeable side effect of some desired plan of action. The repeated use of the expression of indifference throughout the literature is understandable, as the expressed indifference serves as a cue that the relevant outcome is a side effect. However, being indifferent toward a foreseen outcome is not necessary for recognizing that an outcome is a side effect. Take for an example a patient who is planning to take prescription painkillers for extreme muscle soreness. Imagine the patient goes to the pharmacy to pick up her prescription, and the pharmacist informs her that the medicine will probably make her drowsy, so she should not drive or operate heavy machinery after taking the medicine. To this information, it seems perfectly natural for the patient to respond, "I am actually happy that a *side effect* of me taking this medicine is that the medicine might make me drowsy. I could definitely use the extra sleep." Here there is nothing inconsistent about the patient welcoming a side effect of her action's intended goal.

Furthermore, in the side-effect effect cases, the most likely cue through which the expressed indifference is meant to reinforce the idea that the relevant outcome is a side effect is through a failure to attribute to the actor desires to bring about the outcome. However, in Knobe's original chairman cases, people do tend to attribute a desire to harm the environment in the case where the chairman does harm the environment (Guglielmo & Malle, 2010). So, not only is an expression of indifference not necessary for a foreseeable outcome to be rightfully

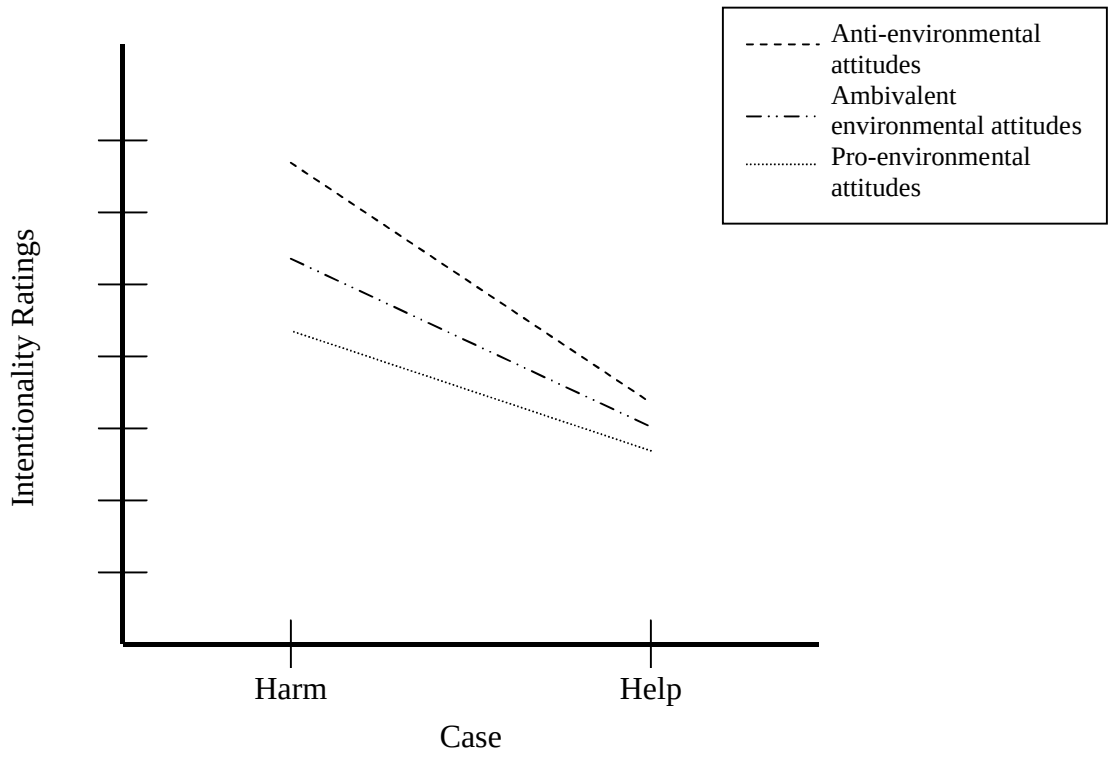

Figure 6. The observed simple slopes of intentionality judgments regressed on case at different levels of chairman values/attitudes judgments controlling for all the other variables in the model. Note: All simple slopes were statistically significant *p*s < .05.

considered a side effect, but the most likely cue through which expressed indifference might act to reinforce the idea that the relevant outcome is a side effect fails to occur in Knobe's chairman case.

These facts open up the potential to explore the side-effect effect with caring actors. Including manipulations that vary on a dimension of expressed concern could possibly shed more light on the current debate, as these types of manipulations could be used to experimentally manipulate the values and attitudes attributed to the actor. With this in mind, I would like to introduce the case of the caring chairman. The *caring, harm case* reads as follows:

> 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 have always been – and still am – deeply concerned about the environment. Our company is also facing bankruptcy. We must do something. With great regret, I say, 'let's start the new program.'"
>
> They started the new program. Sure enough, the environment was harmed.

The *caring, help case* reads as follows:

> 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 help the environment."
>
> The chairman of the board answered, "I have always been – and still am – deeply concerned about the environment. Our company is also facing bankruptcy. We must do something. With great satisfaction, I say, 'Let's start the new program.'"
>
> They started the new program. Sure enough, the environment was helped.

The predictions of Sripada's Deep Self Model should apply invariantly across manipulations of concern. That is, the model should again predict that to the extent that people attribute anti-environmental attitudes to the chairman, intentionality judgments will be higher in the cases where the chairman harms the environment (and lower in cases where the chairman helps the environment), but also, to the extent that people attribute pro-environmental attitudes to the chairman, intentionality judgments will be higher in cases where the chairman helps the environment (and lower in cases where the chairman harms the environment). Again, these predictions imply a crossover effect in which, at the point of indifference, there should be no effect of case on intentionality judgments.

On the other hand, if normative variables are having an effect above and beyond the effect of the deep-self variables, people should remain more likely to attribute intentionality in the harm case than the help case when they attribute ambivalent environmental attitudes to the chairman. Thus, if normative variables are having an effect above and beyond the effect of the

deep-self variables, any crossover effect, if there is a crossover effect, will occur at some point in which people attribute pro-environmental attitudes to the chairman. If the strength of the effect of normative judgments above and beyond the effect of the deep-self judgments on intentionality judgments is strong, then the crossover would occur when the attributions of pro-environmental values and attitudes are high (or the crossover may not occur at all if the effect of the normative variables is strong enough). If the strength of the effect of normative judgments above and beyond the effect of the deep-self variables on intentionality judgments is low, then the crossover would occur when the attributions of pro-environmental values and attitudes are low (but still on the pro-environmental side). If there are no effects of normative judgments above and beyond the effect of deep-self variables on intentionality judgments, then the crossover would occur when ambivalent attitudes are attributed to the chairman (this is the prediction of Sripada's Deep Self Model).

Again, the type of bidirectional deep-self model I outlined in Section 6 is committed to a specific prediction concerning the strength of the effect of the normative judgments on intentionality judgments above and beyond the effect of the deep-self variables. Since, in the caring chairman case, the chairman expresses great concern about the environment, the deep-self attributions should conflict with the norm violation of harming the environment. Thus, according to the type of bidirectional deep-self model outlined in Section 6, the predisposition to attribute greater intentionality in the harm case than in the help case should be attenuated, if not overridden, by the information provided by the deep-self judgments. This implies that, in the caring chairman case, the effect of normative variables on intentionality judgments above beyond the effect of deep-self variables will be weak at best. This further translates into the

specific claim that the crossover effect should occur when the attributions of environmental attitudes to the chairman are near ambivalence.

Given the specific predictions made by the type of bidirectional deep-self model I outlined in Section 6, the caring chairman case should be able to provide a good test of this model. The caring chairman case also provides the potential for another opportunity to test Sripada's Deep Self Model. However, it should be noted that the attenuation prediction of the bidirectional deep-self model predicts that, in the caring chairman case, the results should look approximately like the results that would be predicted by Sripada's Deep Self Model. It is in this sense that the caring chairman case only provides *the potential* for another opportunity to test Sripada's Deep Self Model.

To test the predictions of the two models, I will examine the exact nature of the case *x* chairman values/attitudes judgment moderation effect. Given that the deep-self attributions conflict with the norm violation, I predict that there will not be a difference of intentionality judgments between the harm and help cases when the attributions of environmental attitudes to the chairman are slightly pro-environmental (but still near ambivalence). This prediction assumes that the deep-self judgments do not provide quite enough conflicting information to completely override the effect of normative judgments on intentionality judgments; rather the conflicting information simply attenuates the effect of intentionality judgments. If this exact prediction is correct, this implies that, when people attribute ambivalent environmental attitudes to the chairman, they should be more likely to attribute intentionality in the harm case than the help case. If this prediction is correct, this would again provide evidence against Sripada's Deep Self Model, as his model strictly implies that there should be no difference of intentionality

attributions between the harm and help case when ambivalent environmental attitudes are attributed.

7.2 Methods

The methods used here are the same methods mentioned in Section 4.2, except that the two vignettes used here were the caring chairman vignettes (n = 252) (see Section 7 for full description of these vignettes).

7.3 Results and Discussion*

See Table 4 for the descriptive statistics for intentionality judgments and chairman values/attitudes judgments.

In order to test the nature of the interaction effect of case and chairman values/attitudes judgments on intentionality judgments, a two-step hierarchical regression was conducted. In the first step, case (x_1) and chairman values/attitudes judgments (x_2) were entered as a predictor of intentionality judgments (y). Again, behavioral generalization judgments (x_3) and moral status judgments (x_4) were included in the model as covariates. In an effort to facilitate accurate and meaningful interpretations, the variables chairman values/attitudes judgments, behavioral generalization judgments, and moral status judgments were all mean-centered. In the second step of the hierarchical regression, the interaction term case by (mean-centered) chairman values/attitudes judgments (x_5) was entered into the model as a moderator. The first step yielded the regression formula $-.207x_1 + .137x_2 + .301x_3 + .025x_4$, $F(4, 247) = 5.805$, $p < .001$, $R^2 = .086$. The second step yielded the regression formula, $-.338x_1 - .174x_2 + .192x_3 + .023x_4 + .852x_5$, $F(5, 246) = 8.952$, $p < .001$. The addition of the moderator improved overall model fit, $\Delta R^2 = .068$, $p < .001$, providing evidence of a significant interaction effect.

In order to test the predictions of each model, I tested and plotted simple slopes that

Table 4. Means for intentionality judgments and chairman values/attitudes judgments in the caring chairman case

	Harm Condition	Help Condition
Intentionality Judgments	4.32 (SD = 1.94)	4.41 (SD = 1.65)
Chairman Values/Attitudes Judgments	4.02 (SD = 1.32)	5.36 (SD = 1.07)

represent the effect of case on intentionality judgments at some value for which anti-environmental attitudes are attributed to the chairman, at some value for which ambivalent environmental attitudes are attributed to the chairman, and at some value for which pro-environmental attitudes are attributed to the chairman. For the pro-environmental value, I used the value 4.69. This choice was drive by two factors: a semi-trivial reason and a substantive reason. The semi-trivial reason why this value was chosen was because this was the value of the mean chairman values/attitudes judgments, and this mean fell on the pro-environmental attitude side. The substantive reason why 4.69 was chosen was because this value represents attributions of environmental attitudes that are slightly pro-environmental. Thus, this value allows us to provide a test of the prediction that there will not be a difference of intentionality judgments between the harm and help cases when the attributions of environmental attitudes to the chairman are slightly pro-environmental.

The mid-point 4 on the 7- point scale used to measure chairman values/attitudes judgments represents *neither pro-environmental nor anti-environmental*, so I used a rating of 4 as the value chosen for the attribution of ambivalent environmental attitudes. For anti-environmental attitudes, I used the value 3.31. This value was chosen because the chosen pro-environmental value was .69 units above the mid-point and the value of 3.31 is symmetrically .69 units below the midpoint. This value may seem like a completely arbitrary choice, but the exact value chosen for this slope really does not matter too much because neither model makes different predictions in the chairman cases when people attribute anti-environmental attitudes to the chairman.

The tests reveal that when people attribute pro-environmental attitudes to the chairman, there is no evidence that intentionality attributions differ by case, $B = -.338$, $p = .20$, $r_{sp} = -.075$.

At first glance, this result appears to support the predictions made by the type of bidirectinal deep-self model outlined in Section 6, while also providing some evidence against Sripada's Deep Self Model. However, it may be claimed that this result is somewhat ambiguous as to whether it should count as evidence for or against either model. A proponent of the Deep Self Model may want to claim that a value of 4.69 is close enough to the mid-point that it should be interpreted as an attribution of ambivalent environmental attitudes.

When people attribute ambivalent environmental attitudes to the chairman, they are more likely to attribute intentionality to the chairman in the harm case than in the help case, B = -926, p = .003, r_{sp} = -.177. This result is a bit less ambiguous, disfavoring Sripada's Deep Self Model, as his model predicts that when people attribute ambivalent environmental attitudes to the chairman, there should be no effect of case. Sripada's model predicts that attributions of intentionality will be more likely in the harm case than the help case *only* when anti-environmental attitudes are attributed to the chairman. Whereas, at the tested pro-environmental level, a proponent of the Deep Self Model could argue that the chosen value is so close to ambivalence that, for the purposes of testing their model, it should be treated as ambivalence, there does not appear to be an available response by the proponent of the Deep Self Model to claim that the tested ambivalent level should be taken to be representative of anti-environmental attitudes. On the other hand, this result is consistent with a bidirectional deep-self model, and is fact predicted by the type of bidirectional deep-self model outlined in Section 6.

When people attribute anti-environmental attitudes to the chairman, they remain more inclined to attribute intentionality to the chairman in the harm case than in the help case, and to an even greater degree than when they attribute ambivalent environmental attitudes toward him,

B -1.514, $p < .001$, $r_{sp} = -.226$. Both models would predict this result. See Figure 7 for the simple slope pattern obtained for the caring chairman cases.

Taken together, the results of the original chairman case and the caring chairman case, if correct, provide strong evidence against Sripada's Deep Self Model, while providing evidence for a bidirectional deep-self model (along with providing initial plausibility for a bidirectional deep-self model, like the one I outlined in Section 6).

The original chairman cases provided fairly strong evidence against Sripada's Deep Self Model, as the model's predictions concerning the relationships between case and intentionality judgments were clearly contradicted when ambivalent attitudes were attributed to the chairman and also when moderately strong pro-environmental attitudes were attributed to the chairman. I would also argue that the caring chairman case provided additional evidence against Sripada's Deep Self Model, as the model's predictions concerning the relationship between case and intentionality judgments were not upheld when ambivalent attitudes were attributed to the chairman, and arguably, when slightly pro-environmental attitudes were attributed to the chairman. However, a proponent of Sripada's Deep Self Model may want to lean on the claim that the Deep Self Model can allow some minimal role for normative variables while legitimately remaining a unidirectional model. After all, there is probably some minimal level of influence at which one could legitimately claim that the influence was irrelevantly small. However, even if I were to make this concession to the proponent of the Deep Self Model, the results based on the original chairman case reported in this paper provide strong evidence that normative factors are playing a large role in the original Knobe case. This result, if correct, is without a doubt problematic for the Deep Self Model. After all, his model was explicitly designed to be able to handle this very case.

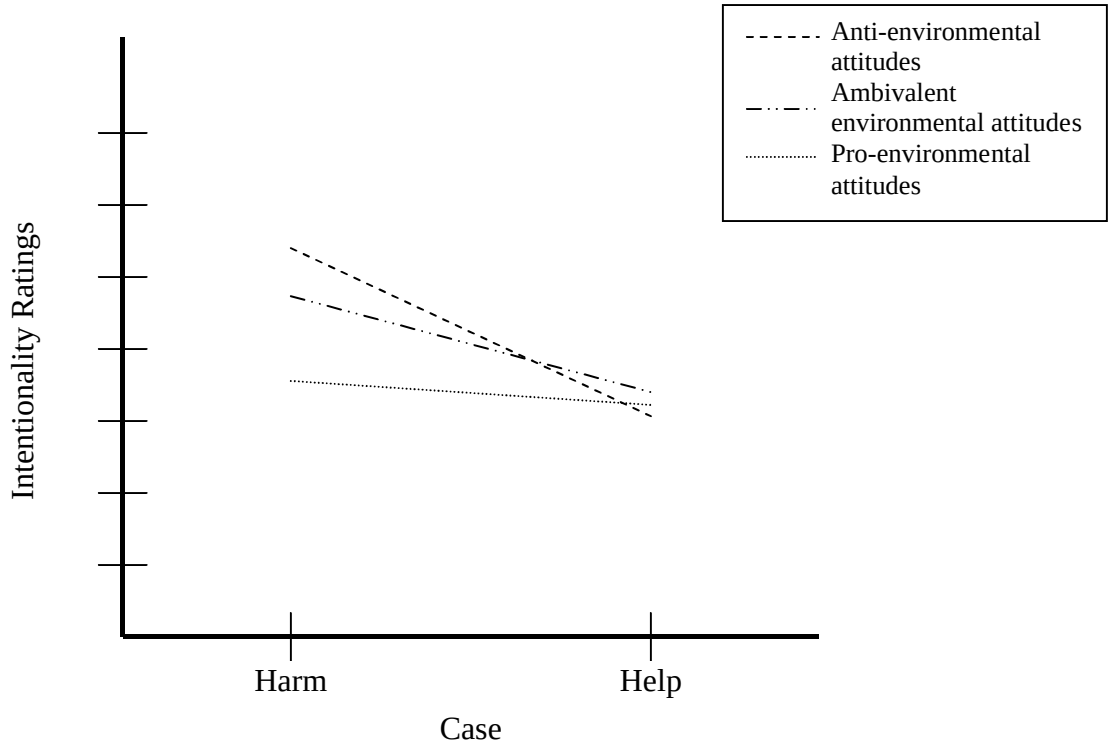

Figure 7. The observed simple slopes of intentionality judgments regressed on case at different levels of chairman values/attitudes judgments controlling for all the other variables in the model for the caring chairman cases.
Note: Simple slopes for anti-environmental attitudes and ambivalent environmental attitudes were significant, *p*s < .05.

On the other hand, both the results of the original chairman cases and the results of the caring chairman cases provided confirmatory evidence for the type of bidirectional model outlined in Section 6. In the original chairman case, since the deep-self judgments do not conflict with the attribution of intentionality in the norm violating case, the effect of normative judgments on intentionality judgments above and beyond the effects of the deep-self variables were strong. The claim that normative judgments have an effect on judgments of intentional action was evidenced by the fact that people were more likely to attribute intentionality in the harm case than in the help case even when ambivalent attitudes were attributed to the chairman. The claim that the effect of normative judgments on judgments of intentional action was strong was evidenced by the fact that people remained more likely to attribute intentionality in the harm case than in the help case even when moderately strong pro-environmental attitudes were attributed to the chairman. In the caring chairman case, since the deep-self judgments do conflict with the attribution of intentionality in the norm violating case, the effect of normative judgments on judgments of intentional action were weak (but still present). Again, the claim that normative judgments have an effect on judgments of intentional action was evidenced by the fact that people were more likely to attribute intentionality in the harm case than in the help case even when ambivalent attitudes were attributed to the chairman. The claim that the effect of normative judgments on judgments of intentional action was weak was evidenced by the fact that there was no effect of case when slightly pro-environmental attitudes were attributed to the chairman. All of these predictions are exactly the predictions the type of bidirectional model I outlined in Section 6 would predict.

The combination of all these results suggest that normative judgments do in fact have an effect on judgments of intentional action, at least some cases. Going forward, I think the debate

concerning the role of normative judgments when making judgments of intentional action would best be served, not by debating whether or not normative judgments play a role in the production of judgments of intentional action, but rather by trying to figure out *in what context and to what extent* normative judgments play a role in the production of judgments of intentional action. In the process of trying to convince the reader that normative judgments are having an effect on judgments of intentional action above and beyond the effects of deep-self judgments, I have provided a rough outline of a model that attempts to give an account of the context and the extent that normative judgments play a role in the production of judgments of intentional action. One potential project going forward may be trying to fill in more details of this (or some similar) model and rigorously test the commitments of the model.

8. A DISAPPEARING ACT?

It is my hope that as we come to the end of this thesis the reader has become convinced—or at least moved to accept—that, even though deep-self-type judgments may play an important role in folk judgments of intentional action, there seems to be no denying the fact that normative judgments retain a significant influence on folk judgments of intentional action, above and beyond the influence of deep-self-type judgments, at least in some contexts. A plausible bidirectional account of the data—both in terms of accounting for the data presented here and the data that has been amassed throughout the literature—might be one that admits that deep-self-type judgments are judgments that can attenuate (or, perhaps intensify) the effect of normative judgments on judgments of intentional action.

If these main points are (roughly) correct, then there are some consequences for the philosophically relevant debates I raised in the introduction of this paper. The first general conclusion that can be drawn from the arguments and evidence presented in this paper is that the

debates mentioned in the introduction of this paper are simply not going to disappear. More to the point, these debates will retain the content of the "bidirectional problem." Remember, one of the advantages of a unidirectional model is that, if a unidirectional model is true, then the puzzles associated with bidirectionalism disappear. So, for instance, if unidirectionalism is right, conservatives—that is, philosophers who think philosophical accounts of various concepts are constrained by folk belief and application—no longer have to wrestle the uneasy choice of either (a) admitting that a philosophically appropriate account of intentional action must account for the influence of normative judgments on intentional action, (b) convincingly argue that the folk are pervasively and systematically confused when they allow normative judgments to influence their judgments of intentional action and then further provide some kind of error theory that would allow the conservative to remain conservative while rejecting the folk's application of intentional action when these applications are (wrongly) influenced by normative judgments, or (c) rejecting conservatism when it comes to the influence of normative judgments on judgments of intentional action. Similarly, if unidirectionalism is correct, those who wish to argue that folk psychology has a fundamentally normative component that is completely independent of explanation and prediction would lose their ability to lean on bidirectional(-seeming) results. In particular, if unidirectionalism is correct, this fact would be devastating to Knobe's argument for the claim that folk psychology has a fundamentally normative component, as the primary motivating premise for his argument relies heavily on bidirectional-seeming results actually being bidirectional. Additionally, if unidirectionalism is correct, the concern that bidirectionality calls into question whether judgments of intentional action can function as the sorts of judgments that can, independently and impartially, guide judgments of responsibility (or wrongness of act or etc.) would no longer be a problem. However, if the arguments and evidence presented in the

paper are (roughly) correct, then these problems do not go away. Bidirectionalism is here to stay, as are the potential problems that bidirectionalism raises.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Alicke, M. (2008). Blaming Badly. *Journal of Cognition and Culture*, 1-2(8), 179-186.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Buckwalter, W. & Beebe, J. (2010). The epistemic side-effect effect. *Mind and Language*, 25 (4), 474-498.

Churchland, Paul. (1989). Folk psychology and the explanation of human behavior. *Philosophical Perspectives*, 3, 225-241.

Cronbach, L. (1987). Statistical Tests for Moderator Variables: Flaws in Analyses Recently Proposed. *Quantitative Methods in Psychology*, 102(3), 414-417.

Doris, J., Knobe, J., Woolfolk, R. (2007). Variantism about responsibility. *Philosophical Perspectives*, 21, 183-214.

Frazier, P., Tix, A., & Barron, K. (2004) Testing Moderator and Mediator Effects in Counseling Psychology Research. *Journal of Counseling Psychology*, 51(1), 115-134.

Goldman, A (1989) Interpretation psychologized. *Mind and Language*, 4, 104-119.

Gopnik, A. & Meltzoff, A. (1997). *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

Gordon, R. (1986) Folk psychology as simulation. *Mind and Language*, 1, 158-171.

Hitchcock, C. & Knobe, J. (2009). Cause and Norm. *Journal of Philosophy*, 106 (11), 587-612.

Hoyle, R. H., & Kenny, D. A. (1999). Statistical power and tests of mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research*. Newbury Park: Sage.

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. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Science*, 9, 357-359.

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

Knobe, J. (2010). Person as a moralists account and its alternative. *Behavioral and Brain Sciences.*

Knobe, J. & Burra, A. (2006). Intentional and Intentional Action: A Cross-Cultural Study. *Journal of Culture and Cognition*, 6, 113-132.

Knobe, J. & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (ed.), *Moral Psychology*, Cambridge, MA: MIT Press. 441-448.

Knobe, J. & Roedder, E. (2009). The ordinary concept of valuing. *Philsophical Issues*, 19 (1), 131-147.

Leslie, A., Knobe, J. & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, 17, 421-427.

Mallon, R. (2008). Knobe vs. Machery: Testing the trade-off hypothesis. Mind and Language, 23, 247-255.

Mele, A. (2001). Acting Intentionally: Probing Folk Notions. In B. Malle, L. Moses, & D. Baldwin (Eds), *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: MIT Press. 27-44.

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

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

Nadelhoffer, T. (2004b). The Butler problem revisited. *Analysis*, 64, 277-284.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions.

Nado, J. 2008: Effects of moral cognition on judgments of intentionality. *British Journal for the Philosophy of Science*, 59, 709-731.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality, *Philosophical Explorations*, 9(2), 203-219.

Nisbett, R. and Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259.

Pettit, D. and Knobe, J. (2009). The pervasive impact of moral judgment. *Mind and Language*, 24 (5), 586-604.

Pizarro, D & Tannenbaum, D. (forthcoming). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. For inclusion in M. Mikulincer & Shaver, P. (Eds) *The Social Psychology of Morality: Exploring the Causes of Good and Evil*. APA Press.

Rose, D., Livengood, J., Sytsma, J., Machery, E. (manuscript). Deep Trouble for the Deep Self. Carnegie Mellon University.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7,* 422-445.

Sigall, H. & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary?. *Personality and Social Psychology Review*, 2(3), 218-226.

Sripada, C. (2010). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies,* 151(2), 159-176.

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

Ulatowski, J. (manuscript). Action under a description. Unpublished manuscript. University of Neveda Las Vegas.

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

Williams, E. J. (1949): *Experimental designs balanced for the estimation of residual effects of treatments*. Australian Journal of Scientific Research, Ser. A 2, 149-168.

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

Young, L., Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture,* 6, 291-304.