Journal of Experimental Social Psychology xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Reports

Mental state attributions and the side-effect effect

Chandra Sekhar Sripada *

Department of Philosophy, University of Michigan, Ann Arbor, 435 South State Street, Ann Arbor, MI 48109-1003, United States

ARTICLE INFO

Article history:
Received 16 December 2010
Revised 17 July 2011
Available online xxxx

Keywords:
Intentional action
Intention
Moral judgment
Side-effect effect
Experimental philosophy

ABSTRACT

The side-effect effect, in which an agent who does not specifically intend an outcome is seen as having brought it about intentionally, is thought to show that moral factors inappropriately bias judgments of intentionality, and to challenge standard mental state models of intentionality judgments. This study used matched vignettes to dissociate a number of moral factors and mental states. Results support the view that mental states, and not moral factors, explain the side-effect effect. However, the critical mental states appear not to be desires as proposed in standard models, but rather 'deeper' evaluative states including values and core evaluative attitudes.

Introduction

Judgments that a person brought about an outcome intentionally ('intentionality judgments') play a pervasive role in social cognition. Intentionality judgments influence moral evaluations (Kleinke, Wallis, & Stalder, 1992), attributions of blame (Hermand, Mullet, Tomera, & Touzart, 2001; Lagnado & Channon, 2008), and punitive reactions (Horan & Kaplan, 1983). They are also critical for the law and jurisprudence, figuring centrally in legal notions such as mens rea (B. F. Malle & Nelson, 2003), and philosophical theories of just punishment (Hart, 1968).

Standard models of intentionality judgments

Models of intentionality judgments have been developed in philosophy and psychology, and have two key features. First, these models propose that intentionality attributions require that the target of the judgment must possess certain specific mental states (B. Malle & Knobe, 1997). Most models agree that these states include at least *desires* (i.e., the target wants the outcome to occur); and 2) *beliefs* (i.e., the target believes that the action performed will bring about the outcome) (Anscombe, 1957; Lowe, 1980; B. Malle & Knobe, 1997). Second, these models assume a unidirectional sequence of processing in which intentionality judgments occur prior to and serve as inputs for subsequent moral/evaluative judgments, such as judgments of blame and punishment (Hart, 1968; B. F. Malle, 2006), consistent with other stage models of social judgments (e.g., Weiner, 1995). Moreover, this unidirectional sequence of processing is thought by many to be rationally correct. The reverse sequence in which moral judgments influence descriptive judgments of intentionality seems problematic, as it is widely thought that moral evaluations of what ought to be the case should not influence factual/descriptive questions of what is in fact the case (M. D. Alicke, Davis, & Pezzo, 1994; Nadelhoffer, 2006a). Recently, however, these 'mental state models' and the unidirectional picture with which they are associated have been challenged by a body of findings regarding people's judgments of the intentionality of bringing about morally-infused side effects.

The side-effect effect

In a famous vignette that first drew attention to the side-effect effect, the chairman of a company is approached by his assistant and told about a new program they are thinking of starting that will help profits and harm the environment. The chairman replies, 'I don't care at all about harming the environment, I just want to make as much profit as I can'. The chairman starts the program, and the environment is indeed harmed. When asked if the chairman intentionally harmed the environment, most people (typically around 80%) say yes. A second group of subjects is given the same case except the word 'harm' is replaced by the word 'help'. That is, the program *helps* the environment, the chairman says 'I don't care at all about *helping* the environment', and when the program is started, the environment is indeed *helped*. When asked whether the chairman intentionally helped the environment, most people (again around 80%) say no (Knobe, 2003). This robust asymmetry in intentionality judgments, dubbed the side-effect effect, has been replicated and extended in a number of other studies in philosophy (Knobe, 2003, 2006; Nadelhoffer, 2004b) and psychology (Guglielmo & Malle, 2010; Leslie, Knobe, & Cohen, 2006; Uttich & Lombrozo, 2010).

* Fax: +1 734 763 8071.
E-mail address: sripada@umich.edu.

0022-1031/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jesp.2011.07.008

Moral factor explanations of the side-effect effect

A widely accepted explanation for the side-effect effect is that moral factors are influencing intentionality judgments (Knobe, 2006; Nadelhoffer, 2004b), though the precise way in which moral factors exert this influence is under debate. According to the 'Good/Bad' model, the key moral factor is the badness of the outcome; people are more likely to judge that the agent acted intentionally when the outcome is bad compared to when the outcome is good (Knobe, 2006). Others have argued the critical moral factor is the blameworthiness of the agent (M. Alicke, 2008; Nadelhoffer, 2004a). This view draws from Mark Alicke's influential Culpable Control Model (M. Alicke, 2000). According to this model, when an agent is perceived as blameworthy, people operate in a 'blame validation mode' in which they seek to justify their blame responses. They are therefore more likely to say the bad outcomes the agent brings about are intentional. For similar reasons, people will also be less likely to say the good outcomes he brings about are intentional (i.e., people will deny credit to a blameworthy agent for a good outcome) (M. Alicke, 1992, 2008; Nadelhoffer, 2004a).

These moral factor models challenge the unidirectional picture associated with standard mental state models because they propose an additional causal pathway in which moral judgments influence intentionality judgments. Moreover, it is widely thought that the effect of moral factors in producing the side-effect effect represents a distorting influence — moral factors inappropriately bias intentionality judgments (Nadelhoffer, 2006b). This picture is consistent with a large body of results in social psychology that show that people often engage in motivated cognition to reach desired conclusions (Ditto, Pizarro, & Tannenbaum, 2009; Kunda, 1990), and/or allow affectively loaded moral/evaluative factors to inappropriately influence descriptive judgments (M. Alicke, 2000; M. D. Alicke et al., 1994).

The Deep Self Concordance Model

An alternative approach to explaining the side-effect effect does not reject traditional mental state models of intentionality judgments altogether. This approach instead modifies these models by expanding the set of mental states that are relevant beyond just the agent's desires to include the agent's 'deeper' evaluative attitudes (Sripada, 2010). The difference between desires and deep attitudes reflects the traditional distinction in social psychology between imputing a mental state to a person to explain a specific action ['causal attributions' (B. F. Malle, 2004)] versus making a dispositional attribution to the person (Kruglanski, 1975; Reeder, 2009; Trope, 1986). Desires are typically temporary states directed towards an individual action at a determinate time, and hence typically figure in causal attributions. In contrast, deep attitudes, such as one's values and evaluative priorities, tend to be seen as more fundamental and stable, and thus figure in dispositional attributions. Deep attitudes also differ from desires in being more general. They often indicate only a broad evaluative orientation towards their objects, typically on a bipolar scale reflecting pro- or anti-orientation (Eagly & Chaiken, 1998). Of note, having deep attitudes directed at some object does not necessarily imply that the person has corresponding desires directed at that object. For example, a person might be anti-union, and yet not have any specific desire to harm a union, or a desire to perform any specific action at all that pertains to unions.

The Deep Self Concordance Account proposes that attributions of deep attitudes are relevant to intentionality judgments in the following way: people are more likely to judge that an agent intentionally brought about an outcome if the outcome concords with the agent's deeper evaluative attitudes, and less likely to judge an agent intentionally brought about an outcome if the outcome discords with the agent's deeper evaluative attitudes. The application of the model to the Chairman vignette proceeds as follows. First, people infer the chairman's deeper evaluative attitudes based on his statements as well as other evidence (e.g., information about characteristic attitudes of corporate chairmen). Specifically, it is plausible that a person declaring that he 'doesn't care at all about harming [helping] the environment' provides strong evidence that the person has an anti-environment orientation. The person, one might reasonably infer, has contempt for the environment, places a low value on the environment, or evidences a trait-like readiness to harm the environment across a range of situations and contexts. Having attributed deep attitudes to the chairman, people next check the concordance between these deep attitudes and the outcome brought about. Since the attributed anti-environment attitudes concord with the outcome in the harm condition but discord with the outcome in the help condition, people will be more likely to judge the agent intentionally brought about the outcome in the harm condition than the help condition, thus explaining the asymmetry that is actually found.

The Deep Self Concordance Model gains support from its application to a variety of cases in the intentionality judgment literature that do not pertain to the side-effect effect (Sripada, 2010). It is also supported by a recent study of the Chairman vignette using structural path analysis that showed that attributions to the chairman of anti-environment values/attitudes and cross-situational behavioral tendencies explained the majority of the asymmetry in intentionality judgments in this vignette (Sripada & Konrath, 2011). In addition, other recent studies have also found that in cases similar to the Chairman vignette, people make trait attributions to the agent (Uttich & Lombrozo, 2010), and are sensitive to the degree of concordance between imputed traits and the outcome brought about (Hughes & Trafimow, 2011).

Testing multiple models of intentionality judgment

A weakness of many existing studies of the side-effect effect is that they used either the Chairman vignette, or else vignettes very much like it, in which the factors relevant to traditional models (i.e., the agent's desires), moral factor models (i.e., moral assessments of the agent and/or outcome), and the Deep Self Concordance Model (i.e., the values and attitudes of the agent) are all highly correlated. In this study, a novel method for creating matched vignettes was deployed to dissociate key factors associated with existing models of intentionality judgments. Specifically, three versions of the Chairman vignette were developed in which the chairman either helps or hinders bringing about a morally good outcome, a morally bad outcome, and an outcome whose moral status varies across individuals. This study thus allowed direct comparisons of the predictions of multiple models including standard desire-based models, moral factor models, and the Deep Self Concordance Model. This study also allowed for testing of an additional model recently adopted by several theorists, which holds that participants' own moral views influence mental state attributions (e.g., Knobe, 2010).

Methods

Participants

Participants for this study were recruited through Amazon's Mechanical Turk (https://www.mturk.com), a validated method for collecting questionnaire data that has been shown to be comparable to pencil and paper methods (Buhrmester, Kwang, & Gosling, 2011). Participants were directed to a secure website to participate in an online study. A total of 450 participants took part.

Design

The study used a between-subject design. Three matched vignettes were used, each of which had a *harm* and *help* version (i.e., there were

Please cite this article as: Sripada, C.S., Mental state attributions and the side-effect effect, *Journal of Experimental Social Psychology* (2011), doi:10.1016/j.jesp.2011.07.008

6 total between-subject conditions in all with 75 participants in each condition).

Measures and procedures

Participants completed a short (average length = 2 min 32 s) survey in which they first read a vignette and then answered a few questions about the vignette. In all vignettes, a chairman of the board either harms or helps an enterprise named 'Beta', but the vignettes differed in terms of the identity and moral status of Beta. In the *Charity* version of the vignette, Beta is a charity that cleans up polluted parks and streams, and the harm version of this vignette is as follows:

> Acme Software is located next to *Beta Charity*, which is *a charity that cleans up polluted parks and streams.* The vice-president of Acme Software went to the chairman of the board and said, "We are thinking of starting a new program to expand our company. It will help us increase profits, but it will also harm *Beta Charity*, forcing them to shut down permanently."
>
> The chairman of the board answered, "I don't care at all about harming *Beta Charity*. I just want to make as much profit as I can. Let's start the new program."
>
> They started the new program. Sure enough, *Beta Charity* was harmed.
>
> (Note: Italics were added for the purposes of exposition and were not present in the text shown to participants).

The help version of this vignette was generated by replacing the word 'harm' with the word 'help'. In addition to the Charity vignette, two other vignettes were utilized. These vignettes were generated by replacing the italicized name and description of the enterprise with alternatives. In the *Chemical* version of the vignette, the name of the enterprise was 'Beta Chemical' and the description was 'a chemical company that pollutes parks and streams'. In the *Clinic* version of the vignette, the name of the enterprise was 'Beta Clinic' and the description was 'an abortion clinic that does late-term abortions'.

After reading the vignette, participants answered the questions in Table 1 in randomized order. Question 2 ('chairman's values/attitudes judgments') and Question 3 ('desire judgments') asked about the chairman's mental states and tested the Deep Self Concordance Model and Desire Model respectively. Question 4 ('badness judgments') and Question 5 ('blameworthiness judgments') asked about moral factors and tested the Good/Bad Model and Culpable Control Model respectively. Question 6 asked people about their own values and attitudes towards Beta. This question was intended to test the Indirect Influence Model, which holds that people's own values and attitudes might influence intentionality judgments indirectly by first influencing mental state ascriptions.

Responses were recorded on a seven-point scale and were coded with values from −3 to 3. Of note, the Deep Self Concordance Model predicts opposite directions of influence between the chairman's values/attitudes variable and intentionality ratings in the harm versus help condition (values/attitudes that are more anti-environment predict higher intentionality ratings in the harm condition, but lower intentionality ratings in the help condition). The chairman's values/attitudes variable was thus reverse coded for all help conditions.

Table 1
Questions used in the study. '[Name]' refers to Charity, Chemical, or Clinic, depending on the version of the vignette.

Question #	Question wording	Anchors for 7-point scale
1	How much do you agree with the statement 'The Chairman intentionally harmed [helped] Beta [Name]?'	Strongly disagree, strongly agree
2	What are the Chairman's underlying values and core attitudes towards Beta [Name]?	Very anti-beta [name], very pro-beta [name]
3	How much do you agree with the statement: 'The Chairman desires that Beta [Name] is harmed [helped]?'	Strongly disagree, strongly agree
4	In your view, how good or bad is the outcome that Beta [Name] is harmed [helped]?	Very good, very bad
5	In your view, how praiseworthy or blameworthy is the Chairman?	Very praiseworthy, very blameworthy
6	What are your own values and core attitudes towards Beta [Name]?	Very anti-beta [name], very pro-beta [name]

Results

Results for all questions for the harm versus help conditions for each vignette are displayed in Fig. 1 and presented in Table 2.

Intentionality ratings

Intentionality ratings were examined with a 3×2 ANOVA with vignette type (Charity, Chemical, and Clinic) and case type (Harm and Help) as factors. Results showed a significant main effect of case type [$F(1444) = 191.96$, $p < 0.001$], as well as a significant vignette type × case

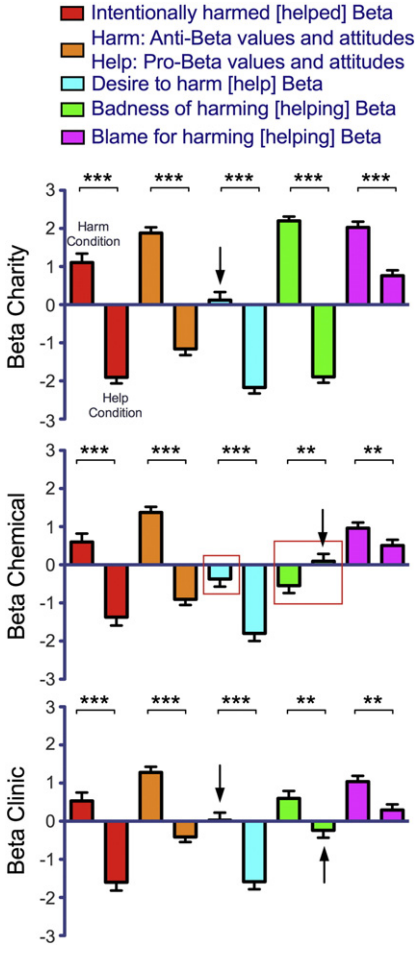

Fig. 1. Intentionality judgments, mental states judgments, and moral judgments. After reading a harm or help version of one of three vignettes, participants answered the questions (shown in Table 1) relevant to different mental state and moral factor models of intentionality judgment. Responses are shown by vignette and case condition. Red boxes indicate instances in which values for a variable are statistically significant in a direction opposite to that predicted by the associated model. Arrows indicate cases in which a statistically significant effect is predicted to occur by a model of intentionality judgment, but the predicted effect is found to be absent. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

Please cite this article as: Sripada, C.S., Mental state attributions and the side-effect effect, *Journal of Experimental Social Psychology* (2011), doi:10.1016/j.jesp.2011.07.008

Table 2
Mean and standard deviation for each variable by condition and vignette. The wording of questions and anchors for responses are given in Table 1.

	Beta Charity		Beta Chemical		Beta Clinic	
	Harm	Help	Harm	Help	Harm	Help
Intentionality	1.11	−1.91	0.6	−1.37	0.53	−1.6
	(2.02)	(1.37)	(1.9)	(1.9)	(1.93)	(1.72)
Chairman values/attitudes	1.85	1.16	1.4	0.91	1.28	0.41
	(1.33)	(1.41)	(1.24)	(1.52)	(1.26)	(1.14)
Desire	−0.12	−2.17	−0.37	−1.8	0.03	−1.59
	(1.85)	(1.34)	(1.71)	(1.74)	(1.78)	(1.73)
Badness	2.2	−1.89	−0.55	0.09	0.6	−0.24
	(0.97)	(1.33)	(1.67)	(2.07)	(1.82)	(1.92)
Blame-worthiness	2.03	0.76	0.96	0.51	1.04	0.29
	(1.31)	(1.23)	(1.29)	(1.38)	(1.54)	(1.16)
Self values/attitudes	−1.49	−1.71	1.29	1.76	−0.03	0.03
	(1.51)	(1.36)	(1.57)	(1.31)	(1.9)	(2.01)

type interaction [$F(2444) = 3.56$, $p = 0.03$], but did not show a significant main effect of vignette type [$F(2444) = 0.30$, $p = 0.74$].

Given the preceding pattern of main effects and interactions, follow-up t-tests were performed to clarify the nature and direction of differences. Probing of the main effect revealed that in all three vignettes, participants expressed significantly more agreement that the chairman intentionally harmed Beta in the harm condition than helped Beta in the help condition [Charity: harm/help difference = 3.01, $t(148) = 10.71$, $p < 0.001$; Chemical harm/help difference = 1.97, $t(148) = 6.36$, $p < 0.001$; Clinic harm/help difference = 2.13, $t(148) = 7.16$, $p < 0.001$] (Fig. 1). Probing of the vignette type × case type interaction revealed that it was driven by a larger difference in mean intentionality ratings between the harm versus help condition in the Charity vignette compared to the other two vignettes [Charity vs. Chemical: $t(148) = 2.31$, $p = 0.02$; Charity vs. Clinic: $t(148) = 1.85$, $p = 0.07$].

Candidate explanatory variables

Between condition analysis

Next 3 × 2 ANOVAs were performed separately on the candidate explanatory variables (chairman's values/attitudes, desire, badness, and blameworthiness). These variables also displayed a significant main effect of case type (all $ps < 0.001$) as well a significant vignette type × case type interaction (all $ps < 0.05$). The single exception was the desire variable for which the vignette type × case type interaction was not significant.

Next, simple effects of the case manipulation within each type of vignette were examined. Pairwise t-tests revealed that for all the candidate explanatory variables, there were significant differences across the harm and help conditions of all three vignettes (Fig. 1).

Given these significant differences, the directional pattern of each variable was next examined. The aim was to see if the direction of difference for each candidate explanatory variable across the harm and help conditions was appropriate for explaining the difference in intentionality ratings that were actually observed. All the candidate explanatory variables did indeed exhibit the appropriate directional pattern across harm versus help conditions to explain intentionality ratings (Fig. 1). The principal exception was the badness variable in the Beta Chemical vignette (shown with a red box in Fig. 1). Participants' badness ratings were significantly higher for helping Beta Chemical than harming Beta Chemical [$t(148) = 2.1$, $p = 0.04$], which is opposite to the pattern required by the Good/Bad Model to account for the asymmetry in intentionality judgments observed in this vignette. This finding is also at odds with the Culpable Control Model, a point that is taken up in the discussion.

Within condition analysis

Each model of intentionality judgment is associated with predictions about the size and direction of participants' responses (above or below the midline) within each condition of each vignette. For example, the Deep Self Concordance Model predicts that people will rate the chairman as significantly anti-environment in the Charity harm condition, thus accounting for observed strong agreement that the chairman intentionally harmed the environment in this condition. Were the chairman found to be significantly pro-environment in the Charity harm condition, then this would count as evidence against the model. Thus identifying discrepancies between the actual size and direction of responses for each variable versus the size and direction predicted by the respective models provides another method for testing the models of intentionality judgments against each other.

Discrepancies between predictions of individual models and observed results were observed in several cases (these are noted with red boxes and black arrows in Fig. 1). With regard to desire ratings, in the Chemical harm condition participants significantly disagreed that the chairman desired to harm Beta Chemical ($M = 0.37$, $t(74) = 1.90$, $p = 0.06$), which is discrepant with their strongly agreeing that the chairman intentionally harmed Beta Chemical. In the Charity and Clinic harm conditions, desire ratings were not statistically different than the midpoint of the scale, which is not consistent with observed strong agreement that the chairman intentionally harmed Beta in both conditions.

Turning to badness ratings, to a significant degree, participants judged harming Beta Chemical to be good rather than bad ($M = -0.55$, $t(74) = 2.83$, $p = 0.006$), which is discrepant with their strongly agreeing that the chairman intentionally harmed Beta Chemical. Additionally, badness judgments in the Beta Chemical and Beta Clinic help conditions were not statistically different from the midpoint of the scale, which is not consistent with the observed strong agreement that the chairman intentionally helped Beta in both conditions.

Mediation analysis

In order to help identify which factor(s) explain the relationship between the harm/help manipulation and intentionality judgments, a mediation analysis (Baron & Kenny, 1986; Shrout & Bolger, 2002) was performed. Mental state variables (i.e., chairman's values/attitudes and chairman's desires) and moral factors (badness of the outcome) were tested using separate mediation models for each variable and each vignette. Results showed that both mental state variables were highly significant mediators of the relationship between the harm/help manipulation and intentionality judgments for all three vignettes (Table 3). The badness variable, however, did not mediate this relationship for any of the three vignettes. It was not possible to test the Culpable Control Model in mediation analysis because this model proposes a complex interaction between moral judgments of the agent (i.e., blameworthiness assessments) and moral judgments of the outcome (i.e., badness judgments), and more details on how these

Table 3
Mediation analysis results. For all three vignettes, mental states (chairman's values/attitudes and desires) mediated the relationship between the harm/help case manipulation and intentionality judgments. In contrast, moral judgments (not shown in table) did not significantly explain differences in intentionality judgments.

	Candidate mediating variable	% of variance in intentionality judgments explained by mediation pathway	Statistical significance of mediation pathway (a*b)	Statistical significance of direct pathway (c′)
Beta Charity	Values and attitudes	40.8	<0.001	<0.001
	Desire	38.5	<0.001	<0.001
Beta Chemical	Values and attitudes	42.6	<0.001	<0.001
	Desire	41.3	<0.001	<0.001
Beta Clinic	Values and attitudes	61.8	<0.001	<0.05
	Desire	51.7	<0.001	<0.001

Please cite this article as: Sripada, C.S., Mental state attributions and the side-effect effect, *Journal of Experimental Social Psychology* (2011), doi:10.1016/j.jesp.2011.07.008

variables are proposed to interact will be required before the model can be tested in mediation analysis.[1]

Testing the Indirect Influence Model

Another possibility is that moral factors might influence intentionality judgments indirectly by first influencing mental state ascriptions. For example, in response to the Chairman vignette (see the Introduction), a person's judgments of whether the chairman is pro- or anti-environment might be significantly influenced by the person's own values and attitudes towards the environment. In particular, people who are more pro-environment might tend to see the chairman's behavior as morally inadequate, thus rating the chairman as more anti-environment (Knobe, 2010; Sripada, 2010, esp. Section 3.3).

In order to test the Indirect Influence Model, participants' own attitudes towards Beta were measured for each vignette. As shown in Fig. 2, participants' attitudes towards Beta varied substantially across the three vignettes, but this variability did not predict their ascriptions of attitudes to the chairman. For example, participants in the Charity and Chemical harm conditions had opposing attitudes towards Beta with the former significantly pro-Beta and the latter significantly anti-Beta (Fig. 2). Yet both groups viewed the chairman as highly anti-Beta (Fig. 1). In further testing, correlations were calculated for each vignette between participants' own values and attitudes towards Beta and their ratings of the chairman's values and attitudes towards Beta. There was a small but statistically significant correlation in the Beta Chemical vignette (Chemical: $r = -0.16$, $p = 0.05$). In the other two vignettes (the Charity and Clinic vignettes), participants' own values and attitudes towards Beta did not predict their ascriptions of values and attitudes to the chairman (Charity: $r = -0.03$, $p = 0.71$; Clinic: $r = -0.08$, $p = 0.32$).

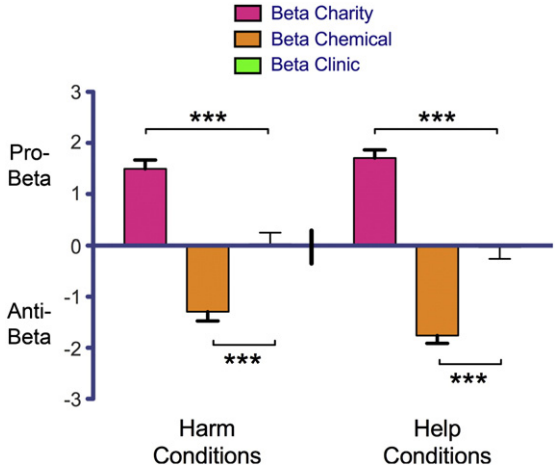

Fig. 2. Participants' attitudes towards Beta by vignette and condition. Participants were pro-Beta in the Beta Charity vignette and anti-Beta in the Beta Chemical vignette. Attitudes were highly variable in the Beta Clinic vignette, with the mean close to the midpoint. $^{*} = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$.

Discussion

In this study of the side-effect effect, participants were randomly assigned to read one of six vignettes in which a person either harms or helps an enterprise with high moral status (charitable organization), low moral status (chemical company), or variable moral status (abortion clinic). This design allowed an effective dissociation of the factors relevant to a number of models of intentionality judgments. Two of these models require the agent to possess certain mental states (Desire Model, Deep Self Concordance Model) while two models focus on moral judgments of the agent or outcome (Good/Bad Model, Culpable Control Model). Consistent with the Deep Self Concordance Model (Sripada, 2010), participants' judgments regarding the agent's values and core attitudes were found to correctly predict the direction of intentionality judgments across all of the vignettes. In addition, these values/attitude judgments were found to mediate the difference in intentionality judgments between the harm and help conditions for each of the vignettes. Variables relevant to other models of intentionality judgments including judgments of desire (Desire Model), the badness of the outcome (Good/Bad Model), and the blameworthiness of the agent (Culpable Control Model), either failed to exhibit the correct directional pattern across vignettes, failed to mediate the difference in intentionality judgments between the harm and help conditions, or both. Overall, these results provide support for three conclusions: 1) Attributions of mental states may help to explain a sizable portion of the side-effect effect; 2) The mental state that appears best suited to explaining the side-effect effect consists of deeper evaluative states, including the agent's values and other core evaluative attitudes; and 3) two moral factors, the badness of the outcome and the blameworthiness of the agent, are unlikely to explain the side-effect effect.

The prevailing view is that the side-effect effect arises due to the biasing influence of moral factors on intentionality judgments, consistent with a large body of findings that support a pervasive role for motivated cognition and affective biases on social judgments (M. Alicke, 2000; Kunda, 1990). Results of the present study identified problems for two influential moral factor models, the Good/Bad Model (Knobe, 2006) and the Culpable Control Model (M. Alicke, 2008; Nadelhoffer, 2004a). For example, in the Chemical vignette, people's badness ratings were significantly lower in the harm condition versus the help condition, but their intentionality ratings were significantly higher in the former than the latter — the reverse of what is predicted by the Good/Bad model. This pattern also runs counter to the predictions of the Culpable Control Model. This model holds that when a person is judged to be blameworthy, intentionality for bad outcomes will be amplified, while intentionality for good outcomes will be reduced (people deny credit to a blameworthy person for good outcomes). However, in the Beta Chemical vignette, the chairman is judged blameworthy in both harm and help conditions, yet intentionality for harming Beta Chemical (an outcome judged to be good) is greater than intentionality for helping Beta Chemical (an outcome judged to be significantly less good). Finally, badness ratings also failed to mediate the relationship between the harm/help manipulation and intentionality judgments for all three vignettes, consistent with a previous study (Sripada & Konrath, 2011) using structural path modeling that produced similar findings.

Results showed that attributions of desires as well as values/attitudes each mediated the relationship between the harm/help case manipulation

[1] The Culpable Control Model predicts that intentionality judgments will be higher when blameworthy agents produce a bad outcome, and lower when they produce a good outcome (people will deny them credit for a good outcome). The model might be tested by calculating a discrepancy score between blameworthiness judgments and badness judgments (i.e., calculate the absolute value of the difference) and entering it as a mediator in the mediation model (low discrepancy scores should predict higher intentionality). The problem with this approach is that an agent judged to be neutral (neither blameworthy or praiseworthy) who brings about a neutral outcome will receive a low discrepancy score, even though the Culpable Control Model would not predict that intentionality ratings would be enhanced. Alternatively, one might bin blameworthiness and badness ratings so that the Culpable Control Model is tested only with clear cases in which the chairman is judged very blameworthy and the outcome is judged either very bad or very good. This method requires specification of the cut-offs for the bins. Finally, it is not clear whether the model is intended to apply to cases in which a praiseworthy agent brings about either a bad or good outcome (Does the model propose that people also go into a 'praise' validation mode when the agent is praiseworthy?). If the model is only supposed to be applied to blameworthy agents, then the preceding analyses would need to be restricted accordingly.

Please cite this article as: Sripada, C.S., Mental state attributions and the side-effect effect, *Journal of Experimental Social Psychology* (2011), doi:10.1016/j.jesp.2011.07.008

and intentionality judgments in all three vignettes, thus supporting the Desire Model and Deep Self Concordance Model respectively. Other features of the results from this study, however, provide a slight edge for the Deep Self Concordance Model over the Desire Model. In particular, the Desire Model appears to encounter problems in explaining the harm conditions of all three vignettes. In the Chemical harm condition, participants significantly disagreed that the chairman desired to harm Beta Chemical, yet strongly agreed that the chairman intentionally harmed Beta Chemical. In the Charity and Clinic harm conditions, participants failed to agree that the chairman desired to harm Beta (desire ratings were not statistically different from the midpoint of the scale), yet again participants strongly agreed that the chairman intentionally harmed Beta. These results are inconsistent with the Desire Model, which requires that a person must have a desire in favor of the outcome for the person to have brought about the outcome intentionally. In contrast, participants judged that the chairman's values and core attitudes were strongly anti-Beta in all three harm conditions, consistent with the predictions of the Deep Self Concordance Model. In sum, the fact that this study, as well as two prior studies (Guglielmo & Malle, 2010; Sripada & Konrath, 2011), found that mental states mediated the side-effect effect provides strong support for mental state models. In addition, the present study provides some evidence that it is specifically attributions of deeper mental states such as values and core attitudes, rather than surface mental states such as desires, that drive the side-effect effect.

Results from this study did not support the Indirect Influence Model, which proposes that participants' own moral attitudes operate indirectly to influence attributions of mental states to the agent (e.g., Knobe, 2010). For example, correlations between participants' own values and attitudes towards Beta and their attributions of values and attitudes to the chairman were either small ($r = 0.16$ for the Beta Chemical vignette), or else not statistically significant. The absence of a correlation for the Beta Clinic condition is particularly interesting. Participants' attitudes towards Beta Charity and Beta Chemical were fairly polarized (Fig. 2), raising a concern that ceiling and floor effects might have prevented detection of the relevant correlation. However, participants' attitudes towards Beta Clinic were quite variable (standard deviation was 2.0) and balanced around the midpoint of the scale (mean rating was 0), mitigating concerns for ceiling and floor effects. Thus, the absence of a correlation in the Clinic vignette between participants' own attitudes towards Beta and Beta-directed attitudes ascribed to the chairman provides strong evidence that these two variables are not related in the way proposed by the Indirect Influence Model.

Another line of evidence, however, might provide some support for a relatively new model that is somewhat related to the Indirect Influence Model. As noted earlier, the difference in intentionality judgments between the harm and help conditions dropped from the Charity vignette (where the difference was 3.0 points) to the Chemical and Clinic vignettes (where the mean difference was 2.1 points). One explanation for this attenuation is that in the Charity vignette, there is a clear social norm against harming Beta (which in this vignette cleans up polluted parks and streams), while in the Chemical and Clinic vignettes, people are presumably less likely to perceive that a norm of this sort exists. If the presence of a clear social norm provides information about the agent's values and attitudes (in particular, when there is a clear social norm that protects X, a person who violates the norm can be inferred to have stronger anti-X values and attitudes than in the case where no norm prevails), then this might explain the attenuation of the harm/help difference observed in the Chemical and Clinic vignettes. Kevin Uttich and Tania Lombrozo have proposed a model along just these lines (Uttich & Lombrozo, 2010). Their 'Rational Scientist' model is fully consistent with the Deep Self Concordance Model in that both models hold that mental states mediate the side-effect effect. A key difference lies in the importance that Uttich and Lombrozo attach to perceived social norms in helping to inform mental state ascriptions. The Deep Self Concordance Model assumes a relatively 'direct' information pathway in which people make attributions of values and core attitudes to the agent based on the agent's statements and behaviors described in the vignette. Uttich and Lombrozo's model proposes an 'indirect' information pathway in which an agent's violation of perceived social norms provides critical information about the agent's mental states. The attenuation of the difference in intentionality judgments (between harm/help conditions) from the Charity vignette to the Chemical/Clinic vignettes provides some support for the operation of Uttich and Lombrozo's indirect information pathway.

The overall importance of the indirect information pathway for explaining the full pattern of results from this study, however, may be somewhat limited. For example, consider the Chemical vignette, in which the chairman harms or helps a company that pollutes parks and streams. In this vignette, it is not clear what information the indirect information pathway would convey, because there is no clear social norm pertaining to a person who harms or helps a company that pollutes parks and streams. If there is a norm against one or both of these actions, it is presumably quite weak, or alternatively there may even be a norm in favor of one or both of these actions. Yet, in response to this vignette, people still judge that the chairman's values and core attitudes are strongly anti-Beta Chemical and the asymmetry in intentionality judgments between the harm and help conditions of this vignette is still robustly present (it is 66% the size of the asymmetry in the Charity condition). The most plausible explanation for this result is that people use the direct information pathway to determine the chairman's evaluative attitudes towards the company (i.e. that he is anti-Beta Chemical), which concord with the outcome in the harm condition and discord with the outcome in the help condition, yielding the observed pattern of intentionality judgments. Similar remarks apply to the Clinic vignette in which there is likely to be much variability in people's views about whether there is a social norm in favor of or against harming or helping an abortion clinic that performs late-term abortions. Yet here again, people view the chairman as strongly anti-Beta and the asymmetry in intentionality judgments between the harm and help conditions of this vignette is still robustly present (71% the size of the asymmetry in the Charity condition). Thus while the present study provides some evidence for the operation of the indirect information pathway proposed by Uttich and Lombrozo, the direct information pathway assumed by the Deep Self Concordance Model plausibly plays the more central role in explaining the overall pattern of mental state attributions and intentionality judgments across the harm and help conditions of the three vignettes.

This study raises further questions and suggests additional avenues of research. First, in the present study, the chairman's deep attitudes were generally selfish and harmful to others' interests. It would be useful to examine the effects of a wider range of deep attitudes, including unselfish and pro-social attitudes, on intentionality judgments to ensure that the effect is not specific to just certain attitude contents. Second, it is not currently known whether attributions of deep attitudes in the Chairman vignette occur explicitly, or whether they represent spontaneous, implicit ascriptions that occur largely without conscious awareness (Uleman, Saribay, & Gonzalez, 2008). In a previous study (Sripada & Konrath, 2011), participants were presented with both versions of the Chairman case and asked to provide explanations of why judgments of intentionality differ in the two versions of the case. Participants overwhelmingly cited moral factors (such as the badness of the outcome or blameworthiness of the chairman) as explaining the differences in intentionality judgments. This result tentatively suggests that deep attitude ascriptions in the Chairman case occur largely implicitly, and/or their role in influencing intentionality judgments is not readily accessible to awareness. These hypotheses warrant further investigation.

In conclusion, this study of the side-effect effect used three pairs of matched vignettes to dissociate the factors relevant to a number of models of intentionality judgments. Results challenge the prevailing view that moral factors drive the side-effect effect, and instead

Please cite this article as: Sripada, C.S., Mental state attributions and the side-effect effect, *Journal of Experimental Social Psychology* (2011), doi:10.1016/j.jesp.2011.07.008

support the idea that mental state attributions, and in particular attributions of values and other core evaluative attitudes, are primarily responsible for the effect.

Acknowledgments

Special thanks to Sara Konrath for detailed comments that greatly improved the manuscript.

References

Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*, 368–378.
Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574.
Alicke, M. (2008). Blaming badly. *Journal of Cognition and Culture*, *1–2*(8), 179–186.
Alicke, M. D., Davis, T. L., & Pezzo, M. V. (1994). A posteriori adjustment of a priori decision criteria. *Social Cognition*, *12*(4), 281–308.
Anscombe, E. (1957). *Intention.* Ithaca, NY: Cornell University Press.
Baron, R., & Kenny, D. (1986). The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.
Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.
Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Chapter 10 motivated moral reasoning. In H. R. Brian (Ed.), *Psychology of learning and motivation* (pp. 307–338). : Academic Press.
Eagly, A. H., & Chaiken, S. (1998). Attitude structure and function. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), (4th ed.). *The handbook of social psychology*, *Vols. 1 and 2*. (pp. 269–322)New York, NY US: McGraw-Hill.
Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and social psychology bulletin*, *36*(12), 1635–1647.
Hart, H. L. A. (1968). Intention and punishment. In H. L. A. Hart (Ed.), *Punishment and responsibility* (pp. 113–135). Oxford: Oxford University Press.
Hermand, D., Mullet, E., Tomera, P., & Touzart, V. (2001). The relationship between intent, consequences, the dangerousness of the victim, and blame: The case of self-defense. *Psychology, Crime & Law*, *7*(1), 57–69.
Horan, H. D., & Kaplan, M. F. (1983). Criminal intent and consequence severity: Effects of moral reasoning on punishment. *Personality and Social Psychology Bulletin*, *9*(4), 638–645.
Hughes, J. S., & Trafimow, D. (2011). Inferences about character and motive influence intentionality attributions about side effects. *British Journal of Social Psychology*, doi:10.1111/j.2044-8309.2011.02031.x.
Kleinke, C. L., Wallis, R., & Stalder, K. (1992). Evaluation of a rapist as a function of expressed intent and remorse. *The Journal of Social Psychology*, *132*(4), 525–537.
Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190–193.
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 scientist, person as moralist. *The Behavioral and Brain Sciences*, *33*, 315–365.
Kruglanski, A. W. (1975). The endogenous–exogenous partition in attribution theory. *Psychological Review*, *82*(6), 387–406.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.
Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.
Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, *17*, 421–427.
Lowe, E. J. (1980). An analysis of intentionality. *The Philosophical Quarterly*, *30*(121), 294–304.
Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* Cambridge, MA US: MIT Press.
Malle, B. F. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture*, *6*(1–2), 87–112.
Malle, B., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101–121.
Malle, B. F., & Nelson, S. E. (2003). Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences & the Law*, *21*(5), 563–580.
Nadelhoffer, T. (2004). Blame, badness, and intentional action: A reply to Knobe and Mendlow. *The Journal of Theoretical and Philosophical Psychology*, *24*, 259–269.
Nadelhoffer, T. (2004). Praise, side effects, and intentional action. *Journal of Theoretical and Philosophical Psychology*, *24*, 196–213.
Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations*, *9*(2), 203–219.
Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality. *Philosophical Explorations*, *9*, 203–220.
Reeder, G. D. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological Inquiry*, *20*(1), 1–18.
Shrout, P., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods*, *7*(4), 422–445.
Sripada, C. S. (2010). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*(2), 159–176.
Sripada, C. S., & Konrath, S. (2011). Telling more than we know about intentional action. *Mind & Language*, *26*(3), 353–380.
Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, *93*(3), 239–257.
Uleman, J., Saribay, S. A., & Gonzalez, C. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*, 329–360.
Uttich, K., & Lombrozo, T. (2010). Moral norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*, 87–100.
Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct.* New York, NY US: Guilford Press.

Please cite this article as: Sripada, C.S., Mental state attributions and the side-effect effect, *Journal of Experimental Social Psychology* (2011), doi:10.1016/j.jesp.2011.07.008