A Causal Model of Intentionality Judgment

STEVEN A. SLOMAN, PHILIP M. FERNBACH
AND SCOTT EWING

Abstract: We propose a causal model theory to explain asymmetries in judgments of the intentionality of a foreseen side-effect that is either negative or positive (Knobe, 2003). The theory is implemented as a Bayesian network relating types of mental states, actions, and consequences that integrates previous hypotheses. It appeals to two inferential routes to judgment about the intentionality of someone else's action: bottom-up from action to desire and top-down from character and disposition. Support for the theory comes from three experiments that test the prediction that bottom-up inference should occur only when the actor's primary objective is known. The model fits intentionality judgments reasonably well with no free parameters.

1. Introduction

We judge the intentionality of others' actions rapidly and, for the most part, effortlessly (Malle, 2004). In this article, we propose that people use a mental model of the causes and effects of another person's actions to evaluate the intentionality of those actions for particular outcomes. Our theory is an attempt to identify the inferences that people use to determine other's attitudes (Heider, 1944, 1958). People do not take an agent's attitude at face value but rather infer it from the agent's actions and the situational pressures that could potentially coerce the attitude (Jones and Harris, 1967).

The framework for constructing the theory comes from the causal model literature (Pearl, 2000; Sloman, 2005; Waldmann and Holyoak, 1992) and the specific structure we propose is largely borrowed from Malle and Knobe (1997). The impetus for the theory was an attempt to explain some puzzling data reported by Knobe (2003). After describing Knobe's result, we will review extant theories of his effect before describing our model. It will become clear that our model is consistent with several accounts of the effect in the literature and represents them in terms of two basic types of inference: a diagnostic inference from action to desire and a top-down inference from character, dispositions, and attitudes to desire.

This work was supported by NSF Award 0518147 to Steven Sloman and by an American Psychological Association dissertation research award to Phil Fernbach. We thank Jonathan Bogard for help collecting data and Jonathan, Adam Darlow, Steve Guglielmo, Josh Knobe, Tania Lombrozo, and Bertram Malle for helpful discussion.

Address for correspondence: Steven Sloman, Cognitive, Linguistic, and Psychological Sciences, Brown University, Box 1821, Providence, RI 02912, USA.
Email: steven_sloman@brown.edu

Mind & Language, Vol. 27, No. 2 April 2012, pp. 154–180.

We will then report three experiments designed to show that both these types of inference contribute to Knobe's effect, and inform judgments of intentional action more generally.

1.1 The Side-Effect Effect

Knobe (2003) asked a group of people he found in a public park in Manhattan the following ('the chairman scenario'):

> 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 program.' They started the program. Sure enough, the environment was harmed. Did the chairman intentionally harm the environment?

The vast majority responded 'yes' (82%). He asked a second group a question that is almost identical differing only in that the program promised to help the environment rather than harm it. Accordingly, participants were asked, 'Did the chairman intentionally *help* the environment?' This time, the vast majority answered 'no' (23%). This finding of an asymmetry in intentionality judgments between a negative versus a positive foreseeable side-effect has been replicated many times and extended to different scenarios (e.g. Knobe, 2004; Nadelhoffer, 2005; Machery, 2008; Mallon, 2008; Uttich and Lombrozo, 2010), cultures (Knobe and Burra, 2006) and ages (Leslie, Knobe and Cohen, 2006).

1.2 Previous Accounts of the Side-Effect Effect

1.2.1 Morality Before Intentionality. Knobe's (2003) original position turned the standard view—that people first evaluate intentionality and use that judgment to evaluate the morality of an action—on its head. He claimed that people attribute intentionality to morally bad actions but not morally good ones. Later he softened this position and asserted that people attribute intentionality to actions that lead to bad outcomes but not actions that lead to good ones (Knobe, 2006).

1.2.2 Pragmatic Account. Contra Knobe (2003, 2004), Adams and Steadman (2004a, 2004b) argued that people believe that the agent in the vignette who acts in a way that causes the harmful side-effect is to blame, therefore labeling the action intentional is correct by virtue of pragmatically implying this blame. Because an agent is more responsible for intentional than unintentional actions, blame is implied by asserting that the harmful action is intentional. On this view, the side-effect asymmetry reflects participants' desire to communicate a moral appraisal.

1.2.3 Insensitive Dependent Measure. Like Adams and Steadman (2004a, 2004b), Guglielmo and Malle (2010) suggested the effect arises from respondents making an assertion they do not really believe. They claimed that both a desire to ascribe blame and the dichotomous response options provided in previous studies (typically *yes* or *no*) have induced participants to report that harmful side-effects were intentional when in fact what they believe is more nuanced, that it is more accurate to say that the harmful side-effect was brought about knowingly rather than intentionally.

1.2.4 Trade-off. A potential challenge to these accounts comes from studies showing the asymmetry in vignette pairs involving side-effects that participants deem to be blame neutral (Machery, 2008; Knobe and Mendlow, 2004). For instance, Machery asked participants to evaluate the following ('smoothie scenario', negative condition):

> Joe was feeling quite dehydrated, so he stopped by the local smoothie shop to buy the largest sized drink available. Before ordering, the cashier told him that the Mega-Sized Smoothies were now one dollar more than they used to be. Joe replied, 'I don't care if I have to pay one dollar more, I just want the biggest smoothie you have.' Sure enough, Joe received the Mega-Sized Smoothie and paid one dollar more for it.
>
> Did Joe intentionally pay one dollar more?

95% of students claimed that Joe did in fact intentionally pay an extra dollar for the smoothie. Machery contrasted this with a positive side-effect condition in which, rather than costing an extra dollar, the mega-sized smoothie came in a special commemorative cup that Joe also said he did not care about but received anyway. In this case, only 45% of students judged that Joe intentionally obtained the commemorative cup. An asymmetry arose even though neither the positive or negative side-effect was blame- or creditworthy.

To explain this result along with the original effect, Machery (2008) proposed that some side-effects (like harming the environment and paying an extra dollar) are perceived as costs that must be traded off with some benefit (like making a profit or obtaining a smoothie). But some side-effects (like helping the environment or obtaining a free cup) are not costs and do not require a trade-off. Side-effects are judged as intentionally brought about only if they are perceived as trade-offs.

1.2.5 Intentionality as Norm Violation. Jones and Davis (1965) argued that an action should influence an observer's judgment about an agent's attitude to the extent the action is unexpected, to the extent it departs from the norm. Uttich and Lombrozo (2010) applied this idea to explain the side-effect effect. They suggested that violations of social norms (e.g. harming the environment) provide more information about mental states than conformity to social norms (e.g. helping the environment). When a side-effect that violates a social norm is permitted to occur by intentionally taking an action that brings it about, this provides evidence that the person who took the action might have intended the side-effect. In contrast, when an action is taken that results in a side-effect that conforms to a social norm, but some other primary purpose for taking the action is clearly present, one need not appeal to the side-effect as an explanation for action. This *discounting* of one potential motive in the presence of another more certain motive may help explain why norm-conforming, helpful, or morally good side-effects are not seen as intentional.

Nanay (2010) proposes a related idea. Harm to the environment, he suggests, provides a reason *against* starting the program while help to the environment provides a reason *for* starting it. Thus in the negative case, according to Nanay, the chairman ignores a reason against acting while in the help case he ignores a reason *for* acting. If one were to consider the counterfactual possibility 'what if the chairman had not ignored the reason?' there is a counterfactual dependence between the action and the outcome in the negative but not the positive case. This applies to Machery's (2008) example as well.

1.2.6 Dispositional Account. Sripada (2009) argued that the side-effect effect is driven by beliefs about the character of the actor in the vignette. He found that people judged the chairman to be anti-environment in both the positive and negative conditions. One interpretation of this finding is that people assume that the chairman is willing to harm the environment, or not motivated to help it, even before learning about his decision.

1.2.7 Differing Implications of 'I Don't Care'. According to Guglielmo and Malle (2010), when the chairman says, 'I don't care at all about *harming* the environment', people understand this to mean that the chairman has set aside concerns, he is willing to harm the environment, and thus has a 'modest pro-attitude' toward this outcome (cf. Davidson, 1963); he mildly favors it. When the chairman says, 'I don't care at all about *helping* the environment', he is understood to mean that the opposite of helping the environment would be acceptable; he is not motivated. In other words, he has no pro-attitude toward helping the environment. This asymmetry in the interpretation of the 'I don't care' statements leads to the asymmetry in intentionality judgments.

1.2.8 Mutual Exclusivity of Accounts. A variety of compelling accounts of the side-effect effect have been proposed. The accounts are not all mutually exclusive both in the sense that more than one may be required to explained the full range of data and in that some of the hypotheses may describe different aspects of the same mechanism. We now propose a causal model intended to capture only those aspects of the existing hypotheses necessary to explain the data.

1.3 Model Structure

Malle and Knobe (1997) proposed a model of the components that people deem necessary to label an action intentional. The model has five hierarchically-related components. At the first level, an agent is judged to have an intention to perform an action if and only if he or she has a desire for the outcome of the action and a belief that the action would lead to the outcome. At the second level, an action is intentional if and only if the agent has the intention, awareness of the action as it is being performed, and sufficient skill to bring about the outcome as a consequence of the action.

Our model is inspired by the five-component model but differs from it in two critical ways. First, it specifies a different kind of relation among variables and thus obeys a different logic. It is a causal model and not a definitional model. Rather than specifying necessary and sufficient conditions for attributions of intentionality, it specifies the general form of a person's beliefs about the causes and effects of another's mental states and actions. Importing the logic of causality shapes the structure of the model (as we will see). It also provides a way to distinguish intervention from observation and gives a natural representation of temporal order (though we will not use these properties in this application).

A causal model also differs from a definitional one by allowing relations to be imperfectly specified due to both ignorance about all the factors influencing events and inherent uncertainty in the world. Thus, causes need not deterministically lead to effects and effects might sometimes have alternative causes. In order to represent this imperfect specification of knowledge, we use probability theory to formalize the model. However, we do not suppose that judgments conform closely to probability theory. Rather, probability theory offers a useful first approximation to people's reasoning in uncertain environments. Our strong claim is that reasoning occurs over the model's qualitative structure.

We express our model as a Causal Bayesian network (Pearl, 2000; Spirtes, Glymour and Scheines, 1993). Such a network is a graphical representation of a probability distribution that has two components: (1) qualitative structure in the form of an acyclic graph composed of nodes and links and (2) quantitative parameters in the form of probabilities and conditional probabilities. Nodes represent variables and directed links (arrows) represent probabilistic dependencies. Two variables are statistically independent if their nodes have no directed pathway connecting them and they do not share any parents (a parent is a node upstream on a directed path). The links do not merely represent statistical dependence but also represent causal power in the sense that they support intervention (see Sloman, 2005, for a simple introduction and Woodward, 2003, for a philosophical analysis).

Causal Bayesian networks are associated with a variety of theorems and algorithms that allow for correct probabilistic inference under very general conditions (Pearl, 2000). Our causal model theory of intentionality judgment is depicted in Figure 1. For the sake of simplicity, we treat all variables as binary (e.g. one either has the relevant desire or its opposite). We have included all five variables of the five-component model but, unlike Malle and Knobe's (1997) definitional formulation, the awareness node is independent of all other variables and therefore does no work in our theory. In all the cases we consider, we simply assume that awareness is present.

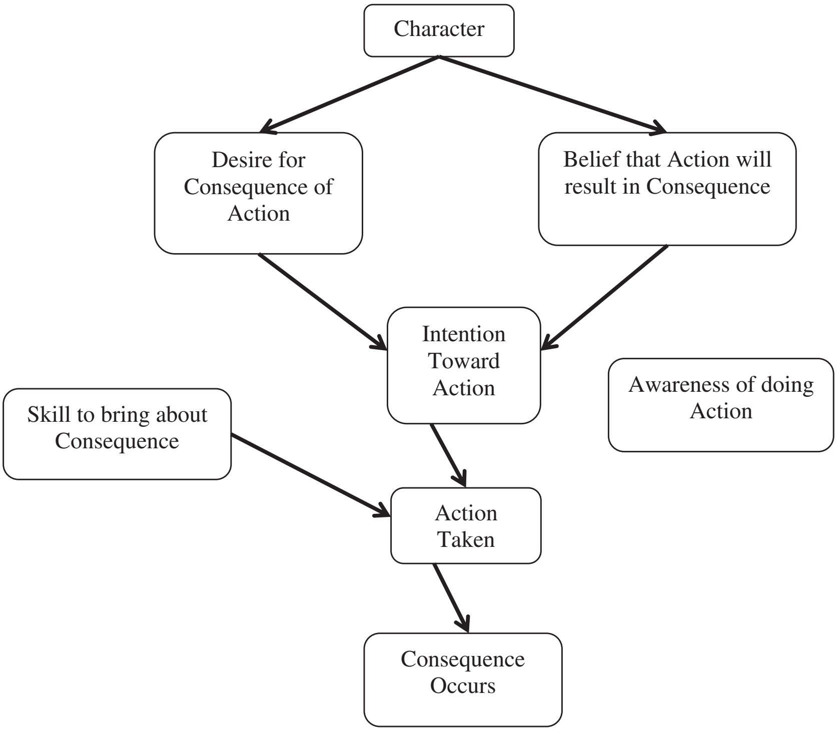
Character
Desire for Consequence of Action
Belief that Action will result in Consequence
Intention Toward Action
Awareness of doing Action
Skill to bring about Consequence
Action Taken
Consequence Occurs

Figure 1 *A causal model of intentionality judgment based on Malle and Knobe's (1997) five-component model*

We also add two variables not in Malle and Knobe's (1997) theory. The character node is intended to capture beliefs about the agent's disposition to have certain desires and beliefs. For instance, if an agent is a terrorist, he or she is likely to want to terrorize. The consequence node reflects beliefs about which outcome will result from the action.

To apply this model to the side-effect effect, the graph can be greatly simplified to represent only those variables whole values change in our account of the effect. Variables that remain constant (belief, skill, awareness, and consequence) retain important conceptual roles in the analysis of intentional action, but they have identical effects across experimental conditions and so their roles are implicit in the conditional probabilities that the model uses to make inferences. We can therefore omit them from the graph. The graph also needs to be complicated in one respect: the relevant scenarios always have two desires that correspond to two consequences, one for the primary consequence (e.g. profit) and the other for the side-effect (e.g. harming the environment). We therefore need two desire nodes, one for the desire for the primary consequence and the other for the side-effect. The resulting model is depicted in Figure 2.

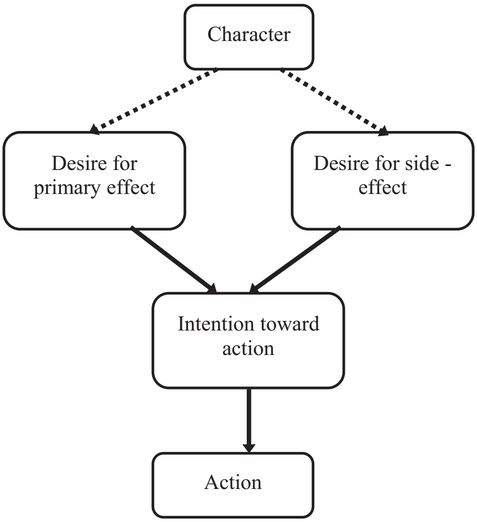

Figure 2 *A causal model theory of the side-effect effect. The solid arrows are involved in bottom-up inference and the dashed arrows in top-down inference. In the absence of top-down information, the two desires are independent*

1.4 Inferential Mechanisms in the Model

Inferring another's state of mind is an uncertain enterprise. In our model a person's attitude toward the side-effect is unobserved; it is inferred from available information. There are two inferential mechanisms:

1.4.1 Bottom-up Inference. First, the observation of action provides a bottom-up source of information through diagnostic inference from effect to cause. Learning that an action took place increases the probability of an intention to perform that action. However, when causes are independent and independently increase the probability of their common effect, they assume a characteristic pattern of inference called explaining away in the Bayes net literature (Pearl, 1988) and discounting in the social psychology literature (Kelley, 1972; Morris and Larrick, 1995). In a broad review of the psychological literature, McClure (1998) showed that discounting is widespread and sensitive to both how variables are related and how they contribute to an effect, though people do not always discount to an appropriate degree. When explaining away applies, diagnostic inference to an intention is mitigated in the presence of another known sufficient cause for the action. For instance, learning that the chairman engaged in an action that helped the environment increases the probability that he intended to help it. But if another sufficient cause is known (e.g. a desire for profit), his action is 'explained away' by that cause and the action is less informative about his intention to help the environment.

A consequence of the way we derive the parameters of our model from data is that explaining away applies only in cases in which norms (Uttich and Lombrozo,

2010) or reasons (Nanay, 2010) identify the side-effect as a reason for action, not when it opposes action. Explaining away requires that causes make independent contributions to their effect. In the negative cases, the side-effect is a reason not to act because willingness to allow a negative consequence is a necessary condition for acting, so the assumption of independent contributions is violated. The primary desire—a reason to act—does not explain away the desire for the side-effect because the observer must assume the agent is willing to allow the negative effect regardless of what is known about the primary desire. In contrast, in the positive case, a pro-attitude is a sufficient and not a necessary condition. For instance, a motivation to help the environment is sufficient for starting a program that does so regardless of one's disposition toward profit. Therefore the conditions for explaining away are satisfied. In our model, the differences between negative and positive arise in the functional forms that relate desire to intention and action. In the positive cases, there is a disjunctive relation between causes (either is individually sufficient). In the negative cases, the relation is conjunctive (both causes are necessary). In our tests of the model we do not assume these functions but rather obtain them empirically in the form of conditional probability judgments.

1.4.2 Top-Down Inference. The second inferential mechanism is top-down in that it involves reasoning from an actor's character to his or her likely desires. Character comprises a variety of potential causes of attitudes toward consequences including traits, dispositions, or current mood.

1.5 Pro Attitudes and Anti Attitudes

To model the side-effect effect, we need a way to answer the question 'did the actor intentionally cause the side-effect?' We take the answer to reside in our representation of the actor's attitude toward the side-effect, i.e. the probability that the agent desires it. Presence and absence of a desire do not reflect the full spectrum of potential attitudes, so to more fully express the range of possibilities, we expand on Davidson's (1963) concept of pro-attitude. We refer to a pro-attitude toward a side-effect as being *motivated* and its negation as *not motivated*. In contrast, we express opposition, or an 'anti-attitude', as *not willing* and the negation of an anti attitude as *willing*.

1.6 Relation to Previous Theories

Nanay (2010) pointed out that the chairman has two reasons to act in the help condition (increasing profit and helping the environment), but has only one reason to act and one reason not to act in the harm condition (increasing profit and harming the environment). The chairman is indifferent toward the side-effect in both cases but there is only a counterfactual dependence in the harm scenario; if he were not indifferent toward harming the environment, the outcome might change because he might be unwilling to start the program. If the chairman were not indifferent toward helping the environment however, the outcome would not change; he would still start the program. Our model can be interpreted as an implementation and elaboration of Nanay's hypothesis.

According to Nanay (2010), the asymmetry between negative and positive side-effects follows from an asymmetry in prescriptive norms, one ought to help the environment and refrain from harming it. Uttich and Lombrozo (2010), following Jones and Harris (1967), suggest that violating a social norm is more informative than conforming to one. Uttich and Lombrozo (2010) also note that statistical norms may play a similar role in setting expectations that are informative when violated. Our model (like Machery, 2008) allows all such norms: moral, conventional, or statistical.

The causal model theory formalizes these accounts using the bottom-up inferential mechanism. Both actual and counterfactual beliefs are encoded in the model's parameters that describe the probabilities of action given actual and counterfactual desires toward the primary objective (e.g. profit) and the side-effect. Our account shares with Machery (2008) the idea that the side-effect effect arises from a trade-off. It differs from Machery in what is being traded off. Machery proposes that intentionality judgments are high for costs traded off to obtain benefits. Our view is that the trade-off concerns information. Intentionality judgments are higher in the negative case because of stronger inferences about attitudes due to outcomes being governed by divergent norms.

Top-down inference in the model captures Sripada's (2009) notion that intentions must arise from a 'deep-self' to be considered intentional. But it is more general, as it allows for any inference from knowledge of a person's character or state of mind.

2. Mathematical Formulation

We model judgments of the intentionality of an action to achieve an outcome as the probability that the outcome was desired given what is known at the time of judgment.

2.1 Bottom-Up Inference

Consider a simple case with no top-down constraints. In such a case, we infer the actor's desire for the side-effect from the action, through an inference about intention, taking into account the actor's desire for the primary effect. Call the variable representing action **A** (this would be starting the program in the chairman scenario), the desire for the primary goal (profit) $\mathbf{D_p}$, and the desire for the side-effect (harming or helping the environment) $\mathbf{D_s}$. In general, we will write variable names in bold, high values in non-bold caps (e.g. A means the action was taken) and low values will be preceded with a tilde (e.g. ~A means the action was not taken). We orient variables such that high values indicate a pro-attitude toward the side-effect and low values represent an anti-attitude.

Assume that our knowledge base tells us the probability that the action will be taken depending on the actor's desires. So we know:

$$\mathrm{P(A\,|\,D_p, D_s), P(A\,|\,D_p, {\sim}D_s)},\ \mathit{etc.}$$

In the bottom-up only model, we want to know $\mathrm{P(D_s\,|\,D_p, A)}$, i.e. how much the agent desires the side-effect given that we know he has taken the action and that he desires the primary effect. An application of Bayes's rule reveals that:

$$\mathrm{P(D_s\,|\,D_p, A)} = \frac{\mathrm{P(A\,|\,D_p, D_s)\ P(D_s\,|\,D_p)}}{\mathrm{P(A\,|\,D_p, D_s)\ P(D_s\,|\,D_p) + P(A\,|\,D_p, {\sim}D_s)\ P({\sim}D_s\,|\,D_p)}}$$

In the absence of top-down constraints, Ds and Dp are independent. Therefore,

$$\mathrm{P(D_s\,|\,D_p) = P(D_s)\ \text{and}\ P({\sim}D_s\,|\,D_p) = P({\sim}D_s)}$$

with the result that:

$$\mathrm{P(D_s\,|\,D_p, A)} = \frac{\mathrm{P(A\,|\,D_p, D_s)P(D_s)}}{\mathrm{P(A\,|\,D_p, D_s)\ P(D_s) + P(A\,|\,D_p, {\sim}D_s)\ P({\sim}D_s)}}. \qquad (1)$$

Note that the quantity we are looking for, how much the agent desires the side-effect once we consider the individual's actions and other desires, depends not only on how likely the agent would be to take the action given those desires but also on how likely they would be to take the action *if they did not have the desire.* In this case, it is the absence of desire for the side-effect (in the denominator) that matters. In general, in Bayesian inference, updating belief requires considering not only what is the case but also what might have been the case if causes were different. This is one sense in which the model depends on counterfactuals.

2.2 Full Model: Adding Top-Down Inference

Let **C** be a variable that takes on values that reflect the agent's character or dispositions (e.g. a bad person, one likely to be selfish, etc.). Top-down inference refers to inferences about the individual's desires from this variable. We require an expression that states the probability of the desire for the side-effect given that we know the agent's character, desire for the primary effect, and the fact that the action was taken:

$$\mathrm{P(D_s\,|\,C, D_p,\ A).}$$

In the appendix, we show that this expression is equal to:

$$\frac{\text{Attitude contribution}}{\text{Attitude contribution} + \text{Opposite attitude contribution}} \tag{2}$$

where the Attitude contribution = $P(D_s \mid C)\,[P(I \mid D_p, D_s) + P(\sim I \mid D_p, D_s)P(A \mid \sim I)]$ and the Opposite attitude contribution = $P(\sim D_s \mid C)\,[P(I \mid D_p, \sim D_s) + P(\sim I \mid D_p, \sim D_s)P(A \mid \sim I)]$. I refers to the intention to act.

We can obtain most of the components to evaluate the expression empirically (see below). The parameter difficult to measure is $P(A \mid \sim I)$, the probability that the action would be taken in the absence of an intention to act. Clearly this should be a very low value. In fact, if it is set to 0, the model simplifies greatly to:

$$\frac{P(D_s \mid C)\,P(I \mid D_p, D_s)}{P(D_s \mid C)\,P(I \mid D_p, D_s) + P(\sim D_s \mid C)\,P(I \mid D_p, \sim D_s)} \tag{3}$$

3. Experiment 1

We test the model quantitatively using the following 4-step procedure: i. We replicate the side-effect effect and thereby obtain appropriate intentionality judgments to compare with model predictions (Experiment 1A). ii. To derive predictions, we need to know participants' judgments of the conditional probabilities that the agents in the scenarios would take the relevant action as a function of their attitudes toward the primary goal and the side-effect, $P(\mathbf{A} \mid \mathbf{D_p}, \mathbf{D_s})$. These judgments are obtained in Experiment 1B. iii. We obtained judgments of agents' attitudes based on their characters in Experiment 1C. iv. The data from Experiments 1B and 1C allow us to derive predictions from the models without any free parameters. The fits are then reported.

3.1 Experiment 1A

Our tests of the model will focus on three scenarios: the chairman (Knobe, 2003), the smoothie (Machery, 2008), and the terrorist (Mallon, 2008). Experiment 1A is designed to replicate the earlier findings using our participant population and to extend the results to a different dependent measure. The earlier studies all used binary (yes/no) questions. We used an 11-point rating scale labeled with 1 (*yes*), 6 (*maybe*), and 11 (*no*).

3.1.1 Method. 269 Brown University students spending time on the main green or passing through the post office volunteered for this study. Materials consisted of a single sheet of paper with the statement of consent on the upper half and below that the chairman, smoothie, or terrorist scenario in the negative or positive condition.

Scenario	Negative (SD)	Positive (SD)	t	df	p
Chairman	7.63 (3.21)	3.00 (2.41)	8.10	96	<.001
Smoothie	6.95 (3.62)	4.26 (2.66)	3.96	85	<.001
Terrorist	8.97 (2.69)	3.08 (2.50)	10.35	82	<.001

Table 1a *Mean intentionality judgments and standard deviations for each scenario in negative and positive conditions along with results of t-tests.*

The chairman and smoothie scenarios were presented earlier. The terrorist scenario (Mallon, 2008) read as follows in the negative condition (italics added):

> A member of a terrorist cell went to the leader and said, 'We are thinking of bombing a nightclub. It will kill many Americans, but *it will also harm the Australians since many Australians will be killed too*.' The leader answered, 'I admit it would be good to *harm the Australians*, but I don't really care about that. I just want to kill as many Americans as possible! Let's bomb the nightclub!' They did bomb the nightclub, and sure enough, *the Australians were harmed since many Australians were killed*. Did the terrorist leader intentionally harm the Australians?

In the positive condition, the phrases in italics were replaced, respectively, by the phrases 'it will also drive down property costs helping the nearby orphanage acquire the land it needs for the children', 'help the orphanage', and 'the orphanage was helped by falling property values'.

3.1.2 Results and Discussion. In order to make the data easier to compare to previous studies, the data were linearly transformed so that higher numbers reflect affirmative responses (see Table 1a). Replicating previous results, the patterns of responses in the negative and positive conditions were highly skewed in opposite directions for all three scenarios indicating that most people preferred to assert a strong yes or no response, rather than one that was equivocal. Differences were significant for each of the scenarios. T-values, degrees of freedom, and p-values are also shown in Table 1a.

In order to compare these results to those from the original experiments, we mapped judgments on the eleven-point scale into *Yes* (1, 2, 3, 4), *Maybe* (5, 6, 7), and *No* (8, 9, 10, 11) and found there was still a clear asymmetry between positive and negative conditions as shown in Table 1b. In summary, the asymmetry was replicated with a rating scale that allowed people to express uncertainty.

3.2 Experiment 1B

In Experiment 1B we collected estimates of the probability that the agents in the various scenarios would take the action described given various attitudes toward the side-effects. We use these estimates to derive predictions for the model.

%	Yes		Maybe		No	
Scenario	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.
Chairman	63	12	13	4	25	84
Smoothie	50	16	14	21	36	63
Terrorist	75	10	14	13	11	77

Table 1b *Percentages of responses coded as* Yes *(1,2,3,4)* Maybe *(5,6,7) and* No *(8,9,10,11) for each scenario in negative and positive conditions.*

3.2.1 Method. Methods were identical to Experiment 1A except that 214 students were tested and the original scenarios were truncated just before the agent responded and an explicit statement was then made about the agent's attitude toward the main and side-effects.

We describe the agent's attitude in the negative condition as opposed to the norm of not willing, i.e. *willing*. Likewise, we assumed that a positive action is one that people are normally motivated to do, and we expressed the opposite of this as *not motivated*. In all scenarios the agent was described as motivated to achieve the primary goal. Additional modifications were used where necessary to produce counterfactual scenarios that were internally consistent. Where the agent was not willing or not motivated to act, this was consistently made the second clause of the sentence. To illustrate, the beginning of the smoothie scenario in the negative condition was as follows:

> Joe was feeling quite dehydrated, so he stopped by the local smoothie shop. Before ordering, the cashier told him that the Mega-Sized Smoothies were now one dollar more than they used to be.

The actual and counterfactual endings to this scenario in the negative condition were:

> Joe *was motivated* to get the largest drink available *and was willing* to pay one dollar more.
>
> Joe *was motivated* to get the largest drink available, *but he was not willing* to pay one dollar more.

The terrorist scenarios presented a greater challenge. Mallon (2008) specifically designed these scenarios so that there would not be a trade-off in the sense of Machery (2008) whether the attitude was considered from either the reader's or terrorist's perspective. However, from the terrorist's perspective, 'motivated' and 'not motivated' seem to be the appropriate terms to use for the side-effect while, from the reader's perspective, 'willing' or 'not willing' are most appropriate. Rather than choose a perspective, we used the outer extremes (not willing or motivated) to construct negative condition statements. But as both the terrorist and the reader were presumed to think that helping an orphanage was good, 'motivated' and 'not motivated' were used for helping the orphanage.[1]

For each version of each scenario we asked each participant to predict the likelihood that the agent would behave in the manner described. For the chairman scenarios, the question was, 'Do you think the chairman will tell the vice-president to start the program?' For the smoothie scenarios, the question was, 'Do you think that Joe will buy the Mega-Sized smoothie?' And for the terrorist scenario the question was, 'Do you think the terrorist leader told the member of the terrorist cell to bomb the nightclub?' Participants were asked to circle a number between 1 (*No*) and 11 (*Yes*).

3.2.2 Results and Discussion. Responses were normalized so that 0 corresponded to *No* and 1 to *Yes*. Only scenarios where the agent had a pro attitude toward the main effect were used in this analysis. All three scenarios show a low probability of action when the agent is not willing to bring about the negative side-effect, and a high probability when the agent is willing or motivated to bring it about (see Figure 3). In contrast, all three scenarios show a high probability of action in the positive condition whether the agent is motivated to bring about the side-effect or not. These conclusions are supported by three 2 × 2 analyses of variance, one for each scenario. Each revealed a significant interaction ($p < .0001$) between negative/positive and high/low state: $F(1, 136) = 24.36$, for chairman; $F(1, 140) = 19.18$, for smoothies; and $F(1, 142) = 16.65$, for terrorist. In each case, the two independent variables also produced highly significant main effects.

3.3 Experiment 1C

To compute values from Equations (2) and (3), we require estimates of $P(D_s \mid C)$, the probability that the character in the scenario would desire the side-effect prior to learning about the action. These were obtained in this study along with estimates of $P(D_p \mid C)$, the probability that the character in the scenario would desire the primary effect although those are not required by the model.

3.3.1 Method. 62 Brown University students volunteered after a class. For each scenario there were three questions: One question asked for the probability that the actor would be motivated to obtain the side-effect in the positive condition, for example, 'How likely is it that a terrorist leader would be motivated to help an orphanage?' In the negative condition of the chairman and smoothie scenarios, a second question asked for the probability that the character would be willing to obtain the side-effect. However, we guessed that respondents would consider

[1] These choices of terms were validated by asking a separate group of participants to rate the appropriateness of four different terms (motivated, not motivated, willing, and not willing) for each attitude. The results were highly supportive.

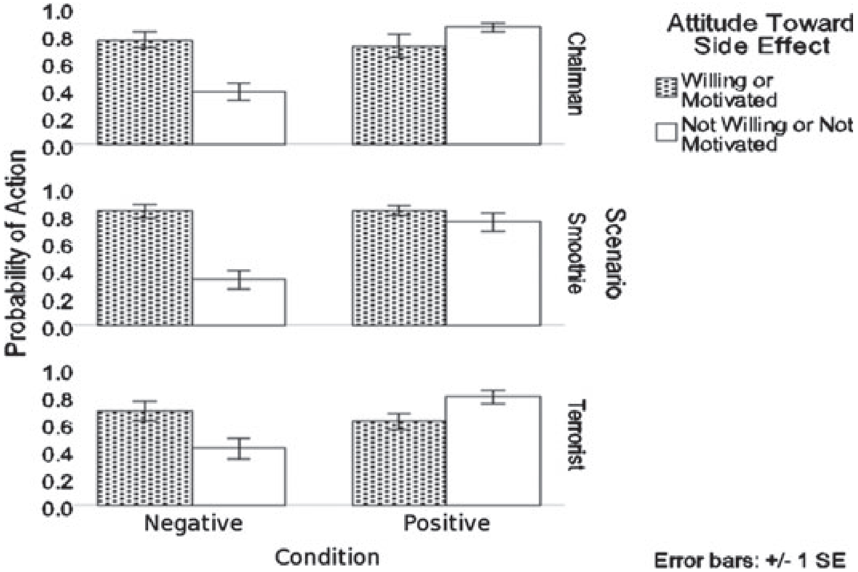

Figure 3 *The mean probability of action estimates shown for the actual and counterfactual attitudes toward the side-effects in each scenario for both negative and positive conditions. Actors were said to be either motivated or willing to obtain the primary objective. Data from Experiment 1B*

terrorists more likely to be motivated than willing to kill Australians and therefore we substituted 'motivated' for 'willing' in that scenario. Finally, one question asked about the primary effect (e.g. 'How likely is it that the chairman of the board of a company would be motivated to make profit?'). The nine questions were split into three questionnaires such that there was one question for each scenario per questionnaire. Participants therefore answered one question for each of the chairman, smoothie, or terrorist scenarios and again responded on an 11-point scale.

3.3.2 Results and Discussion. Means and standard errors for normalized responses are shown in Table 2. A significant difference between conditions obtained in the chairman ($M_{negative} = 0.63$, $M_{positive} = 0.41$; $t(40) = 4.2$, $p < 0.001$) and smoothie scenarios ($M_{negative} = 0.73$, $M_{positive} = 0.59$; $t(40) = 2.6$, $p < 0.05$) and a marginally significant one in the terrorist scenario ($M_{negative} = 0.50$, $M_{positive} = 0.39$; $t(40) = 1.9$, $p < 0.07$).

In all three scenarios, the agent was judged more likely to have a pro attitude toward the side-effect in the negative than positive case: The chairman was deemed more likely to be willing to harm the environment than motivated to help it; Joe more likely to be willing to pay a dollar than motivated to get a commemorative cup; and the terrorist leader marginally more likely to be motivated to kill Australians than help an orphanage.

Scenario	Effect	Type	Mean conditional probability	Standard Error
Chairman	Willing to harm the environment	Side-effect	.63	.04
Chairman	Motivated to help the environment	Side-effect	.41	.04
Chairman	Motivated to make profit	Primary objective	.89	.02
Smoothie	Willing to pay an extra dollar	Side-effect	.73	.03
Smoothie	Motivated to buy a commemorative cup	Side-effect	.59	.05
Smoothie	Motivated to buy the biggest drink available	Primary objective	.65	.05
Terrorist	Motivated to kill Australians	Side-effect	.50	.04
Terrorist	Motivated to help an orphanage	Side-effect	.39	.04
Terrorist	Motivated to kill Americans	Primary objective	.62	.04

Table 2 *Mean conditional probability judgments and standard errors that the agent would be willing or motivated to bring about the side-effect or the primary effect given the agent's identity in the negative and positive conditions for each of 3 scenarios. Data from Experiment 1C.*

3.4 Model Fits

We compare two causal models of the side-effect effect. The first uses bottom-up inference only (no character node) as expressed by Equation (1) above. In this model, we avoid biasing the results by assuming the prior probability of root nodes are distributed uniformly (an 'ignorance' prior). This means the prior probabilities that the agent has a pro-attitude or an anti-attitude were both set at .5.

To obtain model predictions, we used the estimates of $P(A \mid D_p,D_s)$ for values of $\mathbf{D_p}$, $\mathbf{D_s}$ from Experiment 1B. Because the scenarios assume that sufficient skill and awareness obtain, we can safely assume that $P(I \mid D_p,D_s) = P(A \mid D_p,D_s)$ for all D_p, D_s. Note that $P(\sim I \mid D_p,D_s) = 1—P(I \mid D_p,D_s)$ for all D_p, D_s.

We also derive predictions for the full model that includes the character variable, Equation (2). To do so, we used estimates of $P(D_s \mid C)$ from Experiment 1C. For reasons given earlier, we assume $P(A \mid I) = 1$. We evaluate the model with and without the simplifying assumption that $P(A \mid I) = 0$, Equation (3), by also testing Equation (2) with a high value of $P(A \mid \sim I) = .2$.

3.4.1 Results. Predictions of both the bottom-up and full models are shown in Figure 5 alongside the intentionality judgments. The bottom-up model is sufficient to reproduce the general pattern of the data, higher intentionality judgments in the

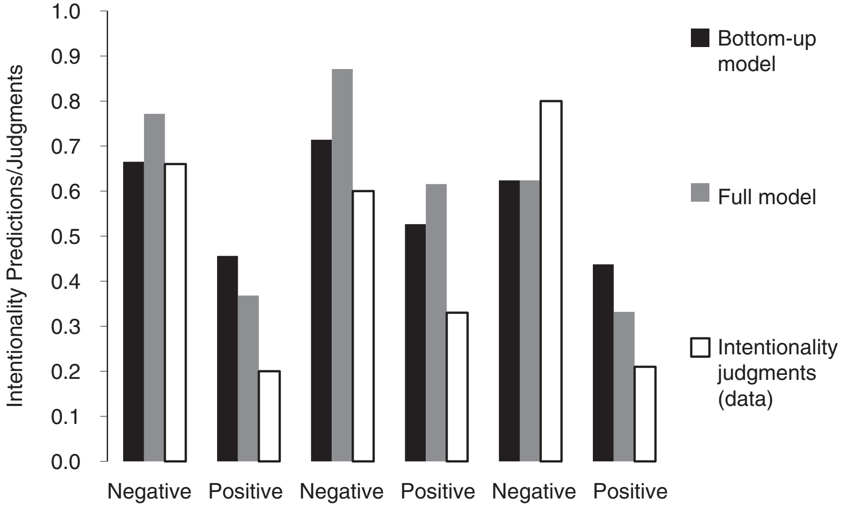

Figure 4 *Predictions for two models of intentionality judgment along with judgments themselves (data from Experiment 1A). Models have no free parameters*

negative than positive conditions in all 3 scenarios. However, it underestimates the differences. The mean actual difference is .44 but the model predicts it should only be .19. When we add top-down inference in the full model (Equation 3), the overall difference increases to .32. Both models tend to overestimate judgment, possibly because respondents had different biases in their treatments of the likelihood and intentionality scales. What is clear though is that the top-down portion of the model is not primarily responsible for the asymmetry in prediction. The bottom-up model correlates extremely well with judgments, r = .87, whereas for the top-down portion (the $P(D_s \mid C)$ judgments), r = .57. The full model actually has a lower correlation than the bottom-up model, r = .76. Note that these fits were obtained without any free parameters.

Using Equation (2) with a non-0 value for $P(A|{\sim}I)$ makes little difference. A value of $P(A|{\sim}I)$ as high as .2 changes the predicted values by an average of only 2%.

3.4.2 Discussion. The bottom-up model, encoding the explaining-away effect, predicted the side-effect asymmetry in all three scenarios using empirical estimates for all parameters other than the prior probabilities for the side-effect which were set to .5 to indicate ignorance. When the character node was added, the asymmetry was enhanced overall though the model did not fit the data quite as well. We believe that it is the combination of these two effects that explains the size and robustness of the side-effect asymmetry although the correlations between model and data suggest that the bottom-up portion of the model is primarily responsible.

In the chairman scenarios, the chairman's statement that he does not care at all may give the impression that he is not a very nice guy. This is reflected in the character judgments which assign him a higher probability of willingness to harm than motivation to help the environment. Previous studies have shown that when the chairman is shown to express regret or an actor is shown to have a generally kind or callous character, the side-effect asymmetry can be attenuated or actually reversed (Guglielmo and Malle, 2010; Phelan and Sarkissian, 2009). We would appeal to the top-down portion of our model to explain such findings.

The smoothie scenario showed explaining away in the positive condition and not in the negative condition. To our surprise, the top-down portion of the model also contributes to the effect in this scenario by increasing the difference between the negative and positive conditions. Participants considered it more likely that Joe would be willing to pay a dollar than motivated to receive a commemorative cup. Clearly, the model's predictions in this case were independent of the moral character of Joe.

The terrorist scenario behaved much like the chairman scenario and we offer the same explanation for it. The twist in the terrorist scenario is that the actor's values are likely to oppose our own. Our account is based on the norms that our respondents attribute to the terrorist in the form of perceived probabilities of action.

4. Experiment 2

Experiment 2 tests the core property of the model we propose, explaining away. To minimize the availability of cues to the agent's character and culpability, we used abstract vignettes that do not suggest an actual positive or negative outcome. Accounts that depend on such cues, the pragmatic account (Adams and Steadman, 2004a, 2004b) and dispositional account (Sripada, 2009), predict no difference between negative and positive conditions in this experiment. We also removed the 'I don't care' statement to eliminate the possibility of a difference arising from differing interpretations of this phrase.

We manipulated the negative versus positive side-effect by varying people's normal attitude toward bringing about the side-effect, whether it is an action people are normally not willing to take or one they are normally motivated to take. Thus, the only feature that differentiates the negative versus positive conditions was the words 'not willing' in the negative condition and 'motivated' in the positive one.

To test the explaining away hypothesis, we manipulated the presence of an alternative cause. In scenarios used in previous studies, the agent's primary motivation has always been made explicit. For example, the chairman in the original vignette states that, 'I just want to make as much profit as I can.' This makes it virtually certain that the chairman is motivated to increase profits. In this experiment, we either included the primary motivation in the scenario explicitly or we did not mention it.

Our account appeals to explaining away in the positive but not the negative condition by virtue of the presence of an alternative explanation, the primary goal. We therefore predicted low intentionality judgments in the positive condition when the primary goal is stated explicitly (as in Experiment 1). In the absence of such an explicit statement, explaining away should be mitigated and intentionality judgments should be high. In the negative condition, there is no explaining away and the manipulation of the primary goal should have no effect; judgments should be consistently high.

4.1 Method

120 people recruited over the Internet participated in an online survey. Participants were asked before and after they answered experimental questions whether they had heard of the Knobe effect. Twenty people were excluded for answering that they had, which left 100 responses that were analyzed (56 women and 44 men; Mean Age = 31, *SD* = 12.22).

The survey consisted of a vignette followed by a question. The negative version read:

> *Z is something that people would normally not be willing to do. Person A went to Person B and said, 'We are thinking about doing Q. It will bring about Y, but it will also bring about Z.' Person B said, 'I want Y. Let's do Q.' They did Q. Sure enough, Z occurred. Did Person B intentionally Z?*

The *positive* condition differed only in that 'not be willing' was replaced by 'be motivated' in the first sentence. The only difference between the primary goal conditions was the presence or absence of the sentence 'I want Y.' Responses were given on the same 11-point rating scale as Experiment 1 (reversed for about half the participants).

4.2 Results

The rating scale reversal had no effect so we collapsed the data across scale directions. Mean intentionality judgments and standard errors are shown in Figure 5. As predicted, judgments were high in all conditions other than the positive condition with the primary goal stated. A two-way analysis of variance showed significant effects for positive versus negative, $F(1, 97) = 9.55$; $p = .003$, primary goal $F(1, 97) = 6.79$; $p = .011$, and the interaction $F(1, 97) = 4.12$; $p = .044$. Post hoc t-tests showed non-significant differences in the negative scenarios with and without the primary goal, $t(52) < 1$, as well as between judgments when the primary goal was absent in positive and negative scenarios, $t(35) < 1$. However, when the primary goal was present, the normal asymmetry appeared, $t(62) = 3.82$, $p < .001$.

4.3 Discussion

In Experiment 2 no information was presented about Person B's attitude, character, or dispositions yet the side-effect asymmetry appeared when the primary goal was

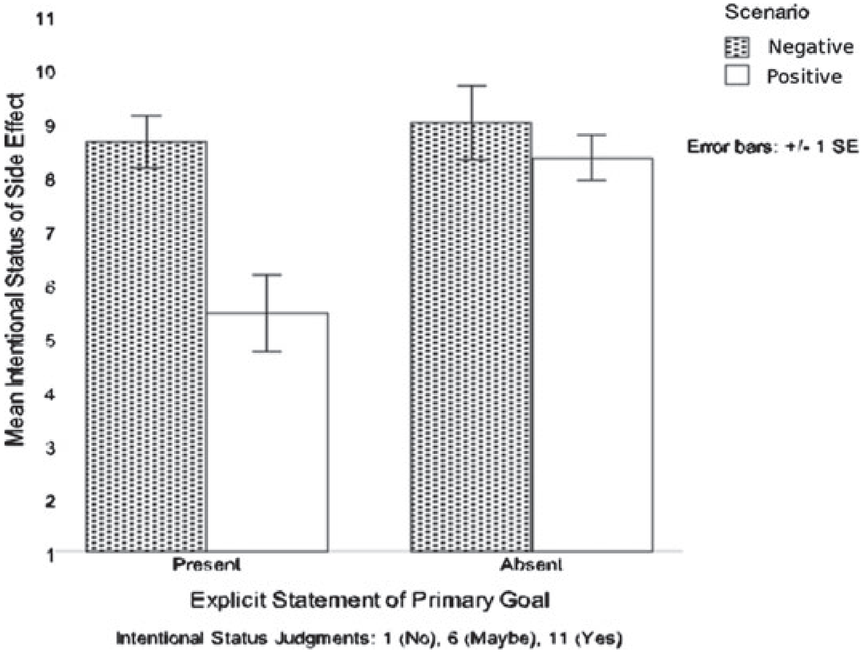

Figure 5 *Mean intentional status judgments with and without an explicit statement of the primary goal in positive and negative scenarios*

stated. When the primary goal was not stated, there was no asymmetry as predicted by the explaining away hypothesis. No other hypothesis except Nanay (2010) predicts these data. High intentionality judgments cannot be attributed to the desire to ascribe blame (Adams and Steadman, 2004a, 2004b) or morally bad outcomes (Knobe, 2003) because the side-effect in the negative condition was not necessarily blameworthy. It was described as something people are normally not willing to do but this could be because it is unconventional, costly, or because it does not offer benefits. Further, the results cannot be due to limited response options. We used a scaled response that permitted the expression of uncertainty and yet participants still indicated that the side-effect was brought about intentionally in three of the four conditions.

5. Experiment 3

The goal of this experiment was parallel to that of Experiment 2: to test the model's prediction that explaining away obtains in positive but not negative conditions, this time using the three scenarios of Experiment 1. As in Experiment 2, we manipulated the presence versus absence of an explicit statement of a primary goal, such as 'I just want to make as much profit as I can.' Unlike in the abstract case, taking out this statement does not necessarily eliminate awareness of the goal. The vice president still states that the program will increase profits and it is common knowledge that CEOs value making money. Nevertheless, the actor's attitude toward the primary goal may be less clear when it is not stated explicitly. Therefore the model predicts that we may see more evidence of explaining away in the positive condition when the primary goal is stated explicitly.

This bottom-up inference should only occur when there is sufficient uncertainty regarding the agent's attitude toward the side-effect. In Experiment 2, the agent never made an explicit statement about the side-effect. But in all the scenarios used in Experiment 1, the agent does make such a statement. In the chairman and smoothie scenarios, the agent states that he does not care at all about the side-effect. In the terrorist scenario the agent explicitly states first that he believes the side-effect to be good and then that he does not care about it. This statement provides some direct information about the actor's attitude toward the side-effect (Guglielmo and Malle, 2010). We therefore also manipulated the presence of this statement. We predicted that explaining away would have a more noticeable effect on intentionality judgments when the 'I don't care' phrase was absent, because in this case the agent's attitudes must be inferred solely from the top-down and bottom-up mechanisms.

Removing statements is likely to have different effects in each condition because each has its own background norms. For instance, if we are confident that Joe in the smoothie scenario 'does not care at all' about paying an extra dollar, then we can also be confident that he is not unwilling to pay it. But if Joe does *not* say that he 'does not care at all' about paying the extra dollar, we do not know what his attitude is. Only a clear expression of indifference toward a good outcome, if believed, precludes the possibility that the person has a pro-attitude toward the outcome. In general, assessments of another's intentions require considering the person's actions, our own prior beliefs, as well as what the person says.

5.1 Method

694 people spending time at a state beach in Rhode Island volunteered. All manipulations were between-participants. Materials were presented as in Experiment 1, one per questionnaire. The questionnaires differed in the presence versus absence of an explicit statement of the primary goal, the presence versus absence of the 'I don't care...' phrase, positive versus negative side-effect, and 3 different scenarios. This yielded 24 questionnaires.

The terrorist scenario differs from the others in that the agent states that the side-effect (killing Australians or helping the orphanage) is good. We left this statement in every condition. He also states that he does not care about these things, and then that he 'just want[s] to kill as many Americans as possible.' These last two statements about the side-effect and the primary goal were manipulated as in the other scenarios. The terrorist therefore expressed general approval of the side-effect accompanied by indifference in the moment versus just general approval of the side-effect.

5.2 Results

5.2.1 Chairman. The asymmetry was present in all four conditions (see Table 5). Analysis of variance ($2 \times 2 \times 2$ factorial) showed a significant main effect for whether the scenario was positive or negative, F (1, 222) = 280.75, $p < .001$. The main effects for primary goal and indifference phrase were not significant. There were significant interactions between the indifference phrase and positive/negative, $F(1, 222) = 4.60$, $p = .03$, and among all three factors, $F(1, 222) = 5.03$, $p = .03$. Other interactions were not significant.

Goal	Indif-ference	Negative (*SD*)	Positive (*SD*)	Dif-ference	*df*	*t*	*p*
Yes	Yes	9.60 (2.54)	3.76 (3.15)	5.84	56	7.74	<.001
Yes	No	9.64 (2.67)	3.72 (2.78)	5.92	55	8.20	<.001
No	Yes	10.00 (1.91)	2.97 (2.65)	7.03	55	11.47	<.001
No	No	9.83 (2.47)	5.79 (2.27)	4.03	56	6.48	<.001

Table 5 *Mean intentionality judgments and standard deviations, differences, and t-tests from Experiment 3 chairman scenario as a function of presence or absence of statement of the primary goal (Goal) and attitude toward the side-effect (Indifference) for both negative and positive conditions.*

Independent t-tests between positive and negative conditions are provided in Table 5. The characteristic asymmetry was present in all four conditions, with the greatest asymmetry arising when the primary goal was absent but the indifference phrase was present. The highest intentionality judgment among the positive scenarios was found in the absence of the indifference phrase and the presence of the primary goal.[2]

5.2.2 Smoothie. An analysis of variance ($2 \times 2\times 2$ factorial) showed a main effect for positive/negative, $F(1, 226) = 10.02$, $p = .002$. There were also significant interactions between indifference and positive/negative, $F(1, 226) = 5.48$, $p = .02$, and between primary goal and positive/negative, $F(1, 226) = 6.52$, $p = .01$. No other interactions were significant. The asymmetry was only significant in one condition, when both the primary goal and the indifference phrase were present (see Table 6). When the indifference phrase was absent, the mean intentionality judgment was higher in the positive condition when the statement of the primary goal was removed. This difference was significant, $t(54) = -2.48$, $p = .016$.

5.2.3 Terrorist. The only significant effect in the terrorist scenario was positive versus negative condition, $F(1, 222) = 544.26$, $p < .001$. Independent t-tests confirmed that all four conditions showed a significant asymmetry (see Table 7).

[2] As a reviewer of an earlier draft pointed out, the robustness of the effect with the chairman scenarios could reflect that they were cleverly designed (by Knobe) to be overdetermined.

Goal	Indifference	Negative (*SD*)	Positive (*SD*)	Difference	*df*	*t*	*p*
Yes	Yes	8.03 (3.79)	3.97 (3.51)	4.07	57	4.23	<.001
Yes	No	5.90 (4.29)	4.31 (3.98)	1.59	57	1.48	>.05
No	Yes	6.03 (4.00)	4.66 (3.64)	1.38	57	1.38	>.05
No	No	5.97 (3.62)	6.74 (3.28)	−.77	55	−.84	>.05

Table 6 *Mean intentionality judgments and standard deviations, differences, and t-tests from Experiment 3 smoothie scenario as a function of presence or absence of statement of the primary goal (Goal) and attitude toward the side-effect (Indifference) for both negative and positive conditions.*

Goal	Indifference	Negative (*SD*)	Positive (*SD*)	Difference	*df*	*t*	*p*
Yes	Yes	10.04 (1.99)	2.52 (2.41)	7.52	55	12.80	<.001
Yes	No	9.89 (2.08)	2.66 (2.53)	7.24	55	11.80	<.001
No	Yes	9.07 (2.28)	2.63 (2.22)	6.44	56	10.90	<.001
No	No	9.70 (2.60)	2.58 (2.23)	7.12	56	11.23	<.001

Table 7 *Mean intentionality judgments and standard deviations, differences, and t-tests from Experiment 3 terrorist scenario as a function of presence or absence of statement of the primary goal (Goal) and attitude toward the side-effect (Indifference) for both negative and positive conditions.*

5.3 Discussion

In the absence of the indifference phrase, removing the primary goal decreased the size of the asymmetry in the smoothie and chairman scenarios by raising intentionality judgments in the positive but not the negative condition, as predicted. But when the indifference phrase was present, the manipulation of the explicit statement of the primary goal had no systematic effect. This may be because the effect of the 'I don't care' phrase is to fix belief that the agent is indifferent to the side-effect to such an extent that explaining away has no influence. A second possibility is that the presence of the 'I don't care' phrase somehow suggests the primary motive. For instance, hearing a chairman say 'I don't care about harming the environment' may make a listener wonder what he does care about. An obvious answer is profit.

Explicit statements about the side-effect or the primary goal had no measurable effect on judgments in the terrorist scenarios. One explanation is that character judgments dominate. People may simply be unwilling to concede that the terrorist would intentionally help an orphanage, even when he states that doing so would be good.

6. General Discussion

We propose that two distinct inferential sources drive judgments of intentionality, a bottom-up diagnostic inference from action to attitudes and a top-down one from character and disposition to attitude. These sources are informed by guesses about other mental states including those we have focused on: motivation, opposition, and indifference. The evidence provided by these mental states depends not only on what we estimate an agent's current state to be, but also on whether that state makes a difference, whether our intentions and actions would be different were it to change. That calculation depends on what state the agent would or should otherwise be in and that requires an assessment of norms. We echo Nanay (2010) in arguing that the causal link between our desires and our intentions reflects norms about what serves as a satisfactory reason for an action.

This model integrates several previous proposals. Like Machery (2008) trade-off account, bottom-up inference involves a trade-off but a trade-off of uncertainty rather than costs and benefits Uttich and Lombrozo's (2009) hypothesis is also closely related to our bottom-up mechanism. Sripada's (2009) Deep Self model is related to top-down causal inference in positing that inferences about character determine likely dispositions toward outcomes. Like Knobe (2003), we offer a place for moral concerns in intentionality judgment, but their role is restricted. Intentionality judgments do not depend on a complete moral analysis; rather they depend on norms about what serves as a typical reason for action. Sometimes those reasons are moral but sometimes they are not.

Evidence for our proposal comes from three studies. In the first we replicated the side-effect effect with an undergraduate population and then showed that empirically obtained parameters yielded model fits that closely approximated the judgments without any free parameters. In the second, we showed the necessity of the bottom-up part of our model in an abstract scenario by using the principle of explaining away to predict that the effect would disappear in the absence of a statement about the agent's primary goal. Finally, in Experiment 3, we derived predictions for intentionality judgments using the scenarios from Experiment 1 as a function of the presence or absence of statements about the primary goal and the side-effect. In the absence of an explicit statement about the actor's attitude toward the side-effect, explaining away was demonstrated in two out of three scenarios. In the presence of the 'I don't care about [the side-effect]' phrase, the manipulation of the presence of an explicit statement of the primary objective had no effect. This suggests either that explaining away matters only in the absence of this phrase, or that the phrase implies the presence of the primary objective. The fact that the explaining away model fit the data so well in Experiment 1 lends some support to the latter interpretation.

Our analysis depends on how mental states that drive behavior are represented. We have followed Davidson (1963) in representing them in terms of norms that serve as satisfactory explanations for action. Both moral norms and conventions can satisfy this criterion of explanation. One can answer the question 'why did you act?' by answering, 'because it was good' or 'because that's what my group does.' However, merely statistical norms often do not satisfy this criterion. It is less satisfactory to explain a behavior by saying 'because that's what I usually do.' That said, statistical norms are sufficient when they have a causal basis. A good explanation could have the form 'because this set of conditions usually compels me to behave this way.'

The side-effect effect is surprising because in the negative condition it appears to be a case of action that is not intended and yet intentional. McCann (2005) and Knobe (2004) have shown that people judge both that the chairman did not intend the side-effect and that he intentionally brought it about. We agree with those authors that this implies that phrases like 'to intend an outcome' and 'to obtain an outcome intentionally' have different meanings. Our analysis is consistent with the proposal that 'to intend an outcome' implicates that the attitude toward the outcome was the single most important reason for taking the action whereas 'to obtain an outcome intentionally' suggests that the attitude toward the outcome was necessary for the action (it would not have been taken otherwise, relative to some norm). For instance, the chairman did not intend to harm the environment because that was not his chief reason for taking the action. But he acted intentionally because if he had not been willing to harm the environment, he would not have taken the action.

An important lesson from this work (especially Experiment 3) is that judges use all information available to make inferences about the intentionality of an outcome. In the cases we have been studying, the agent's beliefs are fixed, so an inference about intentionality comes down to an inference about desires. But even in these contrived cases, the agent's statements, the consideration of counterfactuals, the violation of norms, and the agent's character conspire to suggest whether or not the side-effect was intentional.

Cognitive, Linguistic, and Psychological Sciences
Brown University

Appendix

Here is the derivation of the full model:

$$P(D_s \mid C, D_p, A) = P(D_s, C, D_p, A)/P(C, D_p, A)$$

$$= \frac{\sum_i P(C)\, P(D_p \mid C)\, P(D_s \mid C)\, P(I_i \mid D_p, D_s)\, P(A \mid I_i)}{\sum_i \sum_j P(C)\, P(D_p \mid C)\, P(D_{s,j} \mid C)\, P(I_i \mid D_p, D_{s,j})\, P(A \mid I_i)}$$

where i and j sum over the two states of I and D_s, respectively.

$$= \frac{P(D_s \mid C) \sum_i P(I_i \mid D_p, D_s)\, P(A \mid I_i)}{\sum_j P(D_{s,j} \mid C) \sum_i P(I_i \mid D_p, D_{s,j})\, P(A \mid I_i)}$$

We assume that all the other necessary conditions for intentional action are satisfied and therefore that $P(A \mid I) = 1$. Therefore, the numerator equals:

$$P(D_s \mid C)\ [P(I \mid D_p, D_s) + P(\sim I \mid D_p, D_s) P(A \mid \sim I)].$$

Call that the Attitude contribution. The denominator equals

$$P(D_s \mid C)[P(I \mid D_p,D_s) + P({\sim}I \mid D_p,D_s)P(A \mid {\sim}I)] + P({\sim}D_s \mid C)[P(I \mid D_p,{\sim}D_s) + P({\sim}I \mid D_p,{\sim}D_s)P(A \mid {\sim}I)].$$

Call the second two summands the Opposite attitude contribution. It follows that:

$$P(D_s \mid C, D_p, A) = \frac{\text{Attitude contribution}}{\text{Attitude contribution} + \text{Opposite attitude contribution}}.$$

References

Adams, F. and Steadman, A. 2004a: Intentional action in ordinary language: core concept or pragmatic understanding? *Analysis*, 64, 173–81

Adams, F. and Steadman, A. 2004b: Intentional actions and moral considerations: Still pragmatic. *Analysis*, 64, 268–76.

Davidson, D. 1963: Actions, reasons and causes. *The Journal of Philosophy*, 60, 685–700.

Guglielmo, S. and Malle, B. F. 2010: Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, 36, 1635–47.

Heider, F. 1944: Social perception and phenomenal causality. *Psychological Review*, 51, 358–74.

Heider, F. 1958: *The Psychology of Interpersonal Relations*. New York: Wiley.

Jones, E. E. and Davis, K. E. 1965: From acts to dispositions. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*. New York: Academic Press, 219–66.

Jones, E. E. and Harris, V. A. 1967: The attribution of attitudes. *Journal of Experimental Social Psychology*, 3, 1–24.

Kelley, H. H. 1972: Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. S. Nisbett, S. Valins and B. Weiner (eds), *Attribution: Perceiving the Causes of Behavior*. Morristown, NJ: General Learning Press, 11–26.

Knobe, J. 2003: Intentional action and side effects in ordinary language. *Analysis,* 63, 190–93.

Knobe, J. 2004: Intention, intentional action and moral considerations. *Analysis,* 64, 181–87.

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

Knobe, J. and Burra, A. 2006: Intention and intentional action: a cross-cultural study. *Journal of Culture and Cognition*, 6, 113–32.

Knobe, J. and Mendlow, G. 2004: The good, the bad and the blameworthy: understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology*, 24, 252–58.

Leslie, A., Knobe, J. and Cohen, A. 2006: Acting intentionally and the side-effect effect: Theory of mind' and moral judgment. *Psychological Science*, 17, 421–27.

Machery, E. 2008: Understanding the folk concept of intentional action: philosophical and experimental issues. *Mind & Language*, 23, 101–21.

Malle, B. F. 2004: *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction.* Cambridge, MA: MIT Press.

Malle, B. F., and Knobe, J. 1997: The folk concept of intentionality. *Journal of Experimental Social Psychology*, 33, 101–21.

Mallon, R. 2008: Knobe versus Machery: testing the trade-off hypothesis. *Mind & Language*, 23, 247–55.

McCann, H. J. 2005: Intentional action and intending: recent empirical studies. *Philosophical Psychology*, 18, 737–48.

McClure, J. 1998: Discounting of causes of behavior: are two reasons better than one? *Journal of Personality and Social Psychology*, 74, 7–20.

Morris, M. W., and Larrick, R. P. 1995: When one cause casts doubt on another: a normative analysis of discounting in causal attribution. *Psychological Review*, 102, 331–55.

Nadelhoffer, T. 2005: Skill, luck, control, and intentional action. *Philosophical Psychology* 18(3), 343–54.

Nanay, B. 2010: Morality of modality? What does the attribution of intentionality depend on? *Canadian Journal of Philosophy,* 40, 28–40.

Pearl, J. 2000: *Causality*. Cambridge: Cambridge University Press.

Pearl, J. 1988: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann.

Phelan, M. and Sarkissian, H. 2009: Is the 'trade-off hypothesis' worth trading for? *Mind & Language*, 24, 164–80.

Sloman, S. A. 2005: *Causal Models; How People Think About the World and its Alternatives.* New York: Oxford University Press.

Spirtes, P., Glymour, C. and Scheines R. 1993: *Causation, Prediction and Search.* New York: Springer-Verlag.

Sripada, C. S. 2009: The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, 151(2), 159–76.

Uttich, K. and Lombrozo T. 2010: Norms inform mental state ascriptions: a rational explanation for the side-effect effect. *Cognition*, 116, 87–100.

Waldmann, M. R. and Holyoak, K. J. 1992: Predictive and diagnostic learning within causal models: asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121, 222–36.

Woodward, J. 2003: *Making Things Happen: A Theory of Causal Explanation.* New York: Oxford University Press.