Philos Stud
DOI 10.1007/s11098-009-9423-5

The Deep Self Model and asymmetries in folk judgments about intentional action

Chandra Sekhar Sripada

Abstract Recent studies by experimental philosophers demonstrate puzzling asymmetries in people's judgments about intentional action, leading many philosophers to propose that normative factors are inappropriately influencing intentionality judgments. In this paper, I present and defend the Deep Self Model of judgments about intentional action that provides a quite different explanation for these judgment asymmetries. The Deep Self Model is based on the idea that people make an intuitive distinction between two parts of an agent's psychology, an Acting Self that contains the desires, means-end beliefs, and intentions that are the immediate causal source of an agent's actions, and a Deep Self, which contains an agent's stable and central psychological attitudes, including the agent's values, principles, life goals, and other more fundamental attitudes. The Deep Self Model proposes that when people are asked to make judgments about whether an agent brought about an outcome intentionally, in addition to standard criteria proposed in traditional models, people also assess an additional 'Concordance Criterion': Does the outcome concord with the psychological attitudes of the agent's Deep Self? I show that the Deep Self Model can explain a very complex pattern of judgment asymmetries documented in the experimental philosophy literature, and does so in a way that has significant advantages over competing models.

Keywords Experimental philosophy · Intentional action · Knobe effect · Folk psychology

C. S. Sripada (✉)
University of Michigan, 2215 Angell Hall, 435 South State Street, Ann Arbor, MI 48109-1003, USA
e-mail: sripada@umich.edu

Published online: 03 July 2009

1 Introduction

A man shoots a salesman walking up his driveway intending to steal the salesman's money. Another man shoots a salesman walking up his driveway under the mistaken belief that the salesman is an intruder. A third man shoots a salesman walking up his driveway in a fit of rage because he just found out the salesman is having an affair with his wife. When presented with cases such as these, subjects readily make sophisticated judgments about whether the agent (i.e., the man that did the shooting) intentionally brought about the outcome (the death of the salesman). Much recent work by philosophers and psychologists has sought to identify the factors that go into making these judgments (hereafter referred to as 'intentionality judgments').

According standard models of intentionality judgments, the question of whether an agent brings about an outcome intentionally depends on an inventory of factors relevant to assessing whether the agent *chooses* the outcome and *controls* its occurrence. This inventory typically includes the *desire* to bring about the outcome, the *means-end belief* that the action will bring about the outcome, the *intention* to bring about the outcome, the appropriate *skills* for bringing about he outcome, the absence of *duress*, and other factors such as these. Of course, different versions of these so-called 'Choice/Control Models' disagree with each other about which are the items that are required to be in the inventory (see Malle and Knobe 1997 for a discussion), but they uniformly agree that the items on the inventory consist *exclusively* of factors relevant to assessing the agent's choice of and control over the outcome.

Despite the plausibility of Choice/Control Models, many theorists now believe that such models are fundamentally flawed. To understand their critique, it is useful to begin by first exploring the question of what is the relationship between intentionality judgments and normative judgments. Most all philosophers and psychologists agree that judgments of intentionality serve as inputs for psychological processes that make *normative* judgments, such as whether the agent should be praised or blamed for bringing about the outcome. But a growing controversy has arisen on the *directionality* of the relationship that holds between intentionality and normative judgments.

According to the so-called *Unidirectional Thesis*, intentionality judgments are completed prior to and/or independently of normative judgments (see Fig. 1). According to the alternative *Bidirectional Thesis*, intentionality judgments and normative judgments are interdependent—intentionality judgments both influence and are influenced by normative evaluations (see Fig. 2). Theorists that endorse Choice/Control Models typically also endorse the Unidirectional Thesis. After all, Choice/Control Models propose a check-list of *descriptive* features of the agent that must be met (i.e., mental states that the agent must possess) for an

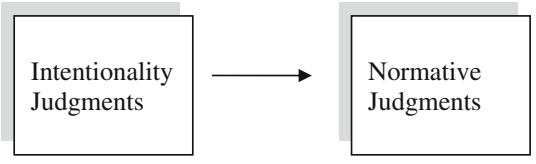
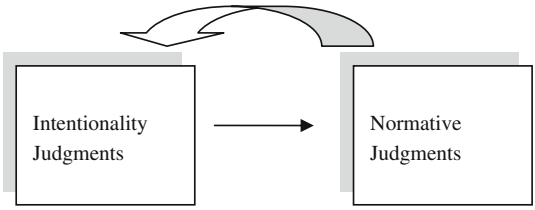


Fig. 1 Unidirectional Thesis

Fig. 2 Bidirectional Thesis



agent to have brought about an outcome intentionally. Normative judgments, in contrast, are fundamentally *evaluative*. It makes sense that descriptive features of a situation are processed prior to and/or independently of evaluative features of the situation, since to do otherwise would be tantamount to allowing one's evaluative reaction about what *ought* to be the case to inappropriately influence one's descriptive attitudes with regard to what *is* the case. But despite the initial plausibility of Choice/Control and the Unidirectional Thesis, there is now widespread belief among theorists that normative judgments do in fact play a significant role in influencing intentionality judgments.

A now classic study by to the philosopher Joshua Knobe illustrates the point vividly. In this experiment, subjects were presented with one of two conditions. In the Harm Condition, subjects read the following vignette:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'
>
> The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was harmed.

Subjects in the Help Condition read a vignette where 'harm' was replaced with 'help':

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.'
>
> The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was helped.

Subjects in the Harm Condition were asked if the Chairman intentionally harmed the environment, while subjects in the Help Condition were asked whether the Chairman intentionally helped the environment. Responses showed a marked asymmetry: Subjects tended to judge that the Chairman intentionally harmed the environment in the Harm Condition, but that he did not intentionally help the environment in the Help Condition (Knobe 2003). Notice that in the two vignettes,

the Chairman possesses the same degree of choice and control over the outcome. That is, in both vignettes, he desires more profit, he believes the new program will lead to more profit, he believes the new program will also have additional side-effects on the environment, he chooses to start the new program, and the program does in fact bring about just the outcomes he foresaw. Since the agent's degree of choice and control over the outcome are equivalent in the two cases, Choice/Control Models are hard-pressed to explain the dramatic difference in subjects' intentionality judgments between the cases.

The failure of Choice/Control Models has led many theorists to posit a role for backtracking influences from normative judgments to intentionality judgments, as captured by the Bidirectional Thesis. A Bidirectional interpretation of the Chairman case claims that in the Harm Condition, subjects make the normative judgment that the outcome the Chairman brought about (harming the environment) is morally bad, and this fact serves to inappropriately influence their judgment about that factual matter of whether he intentionally harmed the environment.[1] In the Help Condition, subjects *aren't* influenced by backtracking influences from the normative judgment that the Chairman brought about a morally bad outcome (since the Chairman *helps* the environment), and thus they reach the putatively factually correct judgment that the Chairman *did not* intentionally help the environment.

The preceding 'Chairman' case is among dozens of others in the experimental literature that claim to identify a fairly consistent pattern: People judge that an agent intentionally brought about an outcome *asymmetrically* based on evaluative or normative features of the situation. This body of asymmetry findings in the experimental literature has led many philosophers and psychologists to reject Choice/Control Models and the Unidirectional Thesis, and instead argue that the Bidirectional Thesis must be true.[2]

However, in this paper I argue that this conclusion is premature. I propose a new model called the *Deep Self Model* that offers an alternative account of how subjects make intentionality judgments. The model is based on an observation originally made by David Hume that notes the key relationship between judging that an agent is responsible for an action and locating the source of the action in the that agent's underlying *stable* and *enduring* psychological attitudes and traits. Hume writes:

> Actions are by their very nature temporary and perishing; and where they proceed not from some cause in the characters and disposition of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honour, if good, nor infamy, if evil, the action itself may be blameable; it may be contrary to all the rules of morality and religion: But the

[1] Knobe himself does argue for the existence of backtracking influences from normative judgments to intentionality judgments as postulated in the Bidirectional Model. But he doesn't view these backtracking influences as *inappropriate*. This intriguing point is defended in various papers, especially Knobe (2006). But since many theorists (myself included) find the backtracking influences postulated in the Bidirectional Thesis to be improper and epistemically unjustifiable, I continue to call these backtracking influences 'inappropriate' throughout the paper.

[2] An alternative approach to explaining the pattern of asymmetry findings in the literature is to argue that no single unified model can capture subjects' varied and inconsistent intuitions. See Doris et al. (2007), and the highly provocative arguments therein, for this kind of approach.

> person is not responsible for it: and as it proceeded from nothing in him, that is durable or constant, and leaves nothing of that nature behind it, 'tis impossible he can, upon its account, become the object of punishment or vengeance (Treatise, bk. 11, Pt. 111, sec. 2).

According to Hume, to find an agent responsible for an action, we must identify some basis for the action in the stable and constant features of the agent. If such an 'anchoring' is not found, the agent is not responsible for the action. In this paper, I propose a generalized (and significantly modified) version of Hume's model that applies to judgments of intentionality. The 'Deep Self Model' that I propose distinguishes two aspects of an agent's psychology: the agent's Acting Self and Deep Self.[3] I'll refine the distinction between the Acting Self and Deep Self a bit later, but as a rough first pass, we can understand the distinction roughly along Humean lines, i.e. the Acting Self contains the relatively 'temporary and perishing' beliefs and desires that are the immediate causal source of the action, while the Deep Self contains the agent's relatively stable values, principles and other more fundamental attitudes. The Deep Self Model agrees with Choice/Control Models that factors that determine whether and agent chooses an outcome and controls its occurrence are relevant for intentionality judgments. But in addition to these factors, the model proposes the following criterion:

Concordance criterion: does the outcome concord with the psychological attitudes of the agent's Deep Self?

According to the Deep Self Model, if the Concordance Criterion fails to be met, then subjects are less likely to judge that the agent brought about the outcome intentionally.

To apply the Deep Self Model to specific cases from the intentionality judgment literature, we must first step back a bit and recognize a somewhat subtle but important feature of the vignettes that are used in this literature. These vignettes often make explicit the factors relevant for assessing whether the agent chooses and has control over the outcome. But in addition to these explicit descriptions, it's highly plausible that subjects make *additional* inferences about the underlying psychological attitudes of the agent's Deep Self. We can call inferences of this sort '*deep attitude attributions*', and the role of such attributions has been highly underappreciated by most theorists. For example, in the Chairman case, the Chairman states 'I don't care at all about harming the environment. I just want to make as much profit as I can.' Based on this statement, it's plausible that subjects make attributions that the Chairman possesses deep attitudes with contents such as 'profit is more important than the environment', 'the environment is not worth helping or preserving', and even perhaps 'helping myself is more important than preventing harms to others', and other attitudes as well. These inferred deep attitudes then interact by means of the Concordance Criterion described above to yield judgments about intentionality (see Fig. 3).

[3] The only previous use of the term 'Deep Self' that I located in the philosophical literature is from Arpaly and Schroeder (1999). My notion of an agent's Deep Self appears to be broadly similar to theirs, though I am not at all sure whether my notion is *identical* to theirs.

The Deep Self Model differs from Choice/Control Models in that it claims that factors that determine whether an agent chooses an outcome and controls it's occurrence are not themselves sufficient for concluding whether an brought about the outcome intentionally. Rather, a crucial additional factor consists of the attitudes contained in the agent's Deep Self, and whether or not these attitudes and the outcome concord. The Deep Self Model also rejects the Bidirectional Thesis. According to the model, all of the information required to make an intentionality judgment (i.e., information about mental states, information about deep attitudes, assessment of concordance between deep attitudes and outcome) is fundamentally *descriptive*. Moreover, the model does not propose that normative judgments, i.e. judgments about whether the outcome brought about by the agent is morally good or bad and/or whether the agent deserves praise or blame for bringing about the outcome, exert backtracking influences on intentionality judgments. In this paper, I show that the Deep Self Model, despite being fully unidirectional, can nonetheless account for the full pattern of asymmetries in intentionality judgments found in the experimental literature.

This paper is divided into two main parts. In the following section (Sect. 2), I clarify the nature of the psychological attitudes that constitute an agent's Deep Self. In Sect. 3, I apply the Deep Self Model to cases that identify asymmetries in intentionality judgments. I conclude by comparing the empirical scorecard between the Deep Self Model and competing models that endorse the Bidirectional Thesis, highlighting strengths of the Deep Self Model.

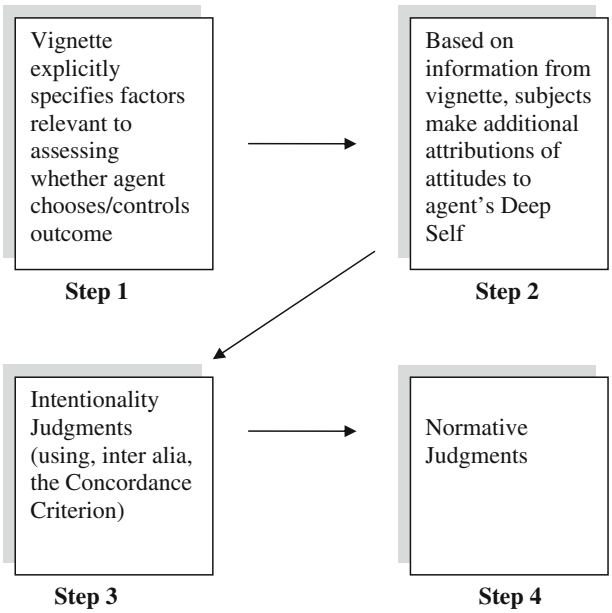


Fig. 3 The Deep Self Model

2 What is the Deep Self?

According to the Deep Self Model, people utilize a naive theory of the structure and contents of the mind and this theory guides judgments about intentionality. The key feature of this theory is that it posits that behind the agent's Acting Self, i.e., the narrow set of outcome-directed proximal desires, means-end beliefs, and intentions, that are the immediate causal source of the action, lies a much larger set of more stable, enduring and fundamental attitudes. These attitudes collectively constitute the agent's Deep Self. The notion of an agent's Deep Self is far from precise, which is not surprising because the idea is part of a rough and ready folk psychological theory. But nonetheless, we can point to a number of *characteristic features* that serve to clarify the Deep Self.

Psychological attitudes that belong to the Deep Self, or 'deep attitudes' as I'll frequently refer to them, tend to be stable and enduring features of a person's mental life. In addition, they are more central to the person's identity and self-conception. Examples of such attitudes might include a person's core moral beliefs and the core evaluative priorities that shape his or her life's goals and ambitions. Deep attitudes also tend to be more abstract. While the proximal attitudes that directly motivate any particular action tend to be narrow and specific (e.g., 'Leo's desire to sign up for Human Physiology this Fall Term'), the deep attitudes and values that lie behind these particular actions are invariably more general (e.g., 'Leo's desire to be a doctor', 'Leo's desire to devote his life to tending to the ill', and his values that endorse compassion and helping others). Another feature of deep attitudes is that they tend to be reflectively endorsed by the person. That is, in cases where a person has two conflicting attitudes towards an outcome, the person will on careful deliberation and reflection tend to endorse the member of the conflicting pair that belongs to his or her Deep Self.

It's worth emphasizing that the depth of an attitude is not the same thing as the attitude's strength. For example, suppose a man comes home to discover his wife in the arms of a lover. In a fit of rage, the man may develop an intensely strong desire to savagely pummel his wife's lover. However, suppose also that the man is a kind, gentle person, and beneficence is a value central to his identity. In this case, the desire to pummel his wife's lover may his strongest desire (in the sense that this is the desire that exerts the most force in guiding action), but it is not deep since the man's central values and ideals strongly reject this desire. Thus the man may pummel his wife's lover *despite* the fact that this action does not concord with his deep attitudes and values.

As noted in the introduction, the notion of an agent's Deep Self is implicitly found in Hume. But a number of contemporary theorists, such as Harry Frankfurt and Gary Watson, have also advanced broadly similar ideas. These theorists characterize the source and contents of the Deep Self in different ways, with Frankfurt emphasizing the importance of reflective endorsement by higher-order attitudes while Watson emphasizes the role of reason in shaping an agent's core

values.[4] Nonetheless, it is a virtue of the Deep Self Model that it posits a notion the general outlines of which are already well accepted by philosophers. Thus the model promises to help anchor the recent asymmetry findings from the experimental literature in an already familiar and established philosophical framework.

Before moving on, it's worth asking the question of why people seem to naturally and intuitively understand the notion of an agent's Deep Self, and why they are so adept at recognizing whether or not an action is anchored in this part of the person's psychology. The answer, I believe, is bound up with our basic interest in predicting and controlling human behavior. The Deep Self contains the agent's stable, enduring, and most central psychological attitudes, and as a consequence, actions that emerge from an agent's Deep Self are likely to form part of a larger pattern in which actions of this same type regularly and reliably happen again. Actions that are not anchored in an agent's Deep Self are, in contrast, more fleeting and ephemeral, and relatively less likely to form a global, reoccurring pattern. Thus our capacity to predict long-term patterns of behavior requires an ability to distinguish actions that are rooted in the agent's Deep Self from those that aren't. Furthermore, our ability to potentially control and alter these global, long-term patterns of behavior is dependent on understanding and influencing the underlying deep psychological attitudes that produce these behavioral patterns.

If the preceding argument is correct, it seems natural to suppose that humans would have developed a host of concepts that perform the task of assessing the relationship between the outcomes that an agent brings about and the enduring and stable parts of the agent's psychology that might have potentially played a role in bringing about these outcomes. Concepts such as these would play an indispensable role in the larger task of predicting, understanding and potentially controlling long-term recurrent patterns of behavior. According to the Deep Self Model, the folk concept of intentionality performs precisely this role (that is, in addition to whatever *other* roles that it performs). In making judgments about intentionality, subjects are, inter alia, assessing the concordance between the outcomes an agent brings about and the relatively deep and enduring parts of the agent's underlying psychology, and this concept is applied only when such concordance obtains.

3 Explaining the asymmetry findings

In this part of the paper, I apply the Deep Self Model to a number of specific cases in the experimental literature that find asymmetries in intentionality judgments. These cases all exhibit a standard structure: A pair of cases is presented to subjects (typically each case is presented to a different group of subjects). The agents in the paired cases are explicitly described as possessing many of the same mental states (i.e., the same desire to bring about the outcome, the same means-end beliefs that his or her action will bring about the outcome, the same beliefs about whether the

[4] See Wolf (1990), especially chapter 2, and Arpaly and Schroeder (1999) for lucid general discussions of philosophical works that have drawn distinctions broadly similar to the distinction I've drawn between an agent's Deep Self versus other parts of an agent's psychology.

action will bring about any side-effects, etc.…). Nonetheless, judgments about intentionality differ between the cases, and the question is then posed, what creates the asymmetry?

The Deep Self Model proposes that the asymmetry between paired cases is explained by different levels of concordance between the agent's deep attitudes and the outcome in the two cases. There are basically three ways in which paired cases can exhibit different levels of concordance between deep attitudes and outcome, based on whether concordance differences between the cases are driven by difference in deep attitudes, differences in outcome, or both: (1) *Outcome-driven asymmetry*: the deep attitudes of the agents in the two cases are the same, but the outcomes in the two cases differ; (2) *Deep attitude-driven asymmetry*: the outcomes in the two cases are the same, but the deep attitudes of the agents in the two cases differ; and (3) *Dual asymmetry*: In the two cases, the deep attitudes of the agents and the outcomes *both* differ, and the agent's deep attitudes and outcome are concordant in one case but not the other. These three kinds of asymmetries are summarized in Table 1.

3.1 The Deep Self interpretation of the Chairman case

The Deep Self Model interprets the Chairman case, which was presented in full in the introduction to the paper, as an instance of an outcome-driven asymmetry (see Table 1). As noted earlier, in both the Harm Condition and Help Condition, it's plausible that subjects infer that the Chairman possesses deep attitudes that are hostile to the environment, since he explicitly declares that he cares only about making money and expresses contempt for the environment. However, the outcomes in the two cases are different. In the Harm condition, the agent brings about harm to the environment, which is *concordant* with his underlying deep attitudes, while in the Help Condition he helps the environment, which is *discordant* with his underlying deep attitudes. It follows from the Deep Self model that the chairman will be more likely to be judged as having intentionally brought about the outcome in the Harm condition versus the Help condition. This is precisely the asymmetry that was found.

To further test the Deep Self interpretation of the Chairman case, I conducted an additional study. The Chairman case was presented to 40 subjects (all University of

Table 1 Three kinds of asymmetries

Kind of asymmetry	Deep attitudes of agent in case 1 versus 2	Outcome in case 1 versus 2	Concordance between attitudes and outcome in case 1 versus 2
Outcome-driven asymmetry	Same	Different	Different
Deep attitude-driven asymmetry	Different	Same	Different
Dual asymmetry	Different	Different	Different

Three ways in which paired cases can exhibit differences in concordance between the agents' deep attitudes and the outcome, leading to asymmetries in intentionality judgments

Michigan undergraduates, 20 saw the Harm Condition and the other 20 saw the Help Condition). But this time, instead of making judgments about intentionality, subjects were asked to 'Rate the Chairman's values and attitudes with regard to the environment' on a 7 point scale (1 = Anti-environment and 7 = Pro-environment). Results showed that subjects in both the Harm and Help condition rated the Chairman as anti-environment (mean 1.9 for the Harm condition and 2.7 for the Help Condition).[5] Now, the Deep Self interpretation of the Chairman case requires that subjects attribute anti-environment attitudes to the Chairman in both the Harm and Help conditions. Indeed, this attribution is central to the model, and constitutes Step 2 of the model as outlined in Fig. 2. The preceding finding that subjects indeed do make the attributions that are required by the model provides the model with additional support.

3.2 Is the Deep Self Model a version of the bidirectional model?

Let's pause for a moment to address the following worry. It might be argued that when subjects infer that the Chairman has anti-environment attitudes as part of his Deep Self, they are in fact also making a normative assessment, since virtually all subjects are likely to believe that holding anti-environment attitudes is morally wrong. Thus one might worry that the Deep Self Model is a version of the Bidirectional Model, since subjects do in fact make a normative judgment prior to making judgments about intentionality.

I believe that this worry is mistaken. In addressing this worry, let's begin by making a careful distinction between j*udging that an agent possesses an evaluative attitude* and *making an evaluative judgment*. For example, *abortion is wrong* is an evaluative judgment. Judging that *George W. Bush believes that abortion is wrong* is a factual judgment. In the Chairman case, subjects attribute to the Chairman anti-environment deep attitudes. The ascription of this evaluative attitude is what allows subjects to subsequently judge that the outcome of harming the environment concords with this underlying evaluative attitude in the Harm Condition and the outcome of helping the environment does not concord with this underlying evaluative attitude in the Help Condition.

Now, in addition to assigning evaluative attitudes to the Chairman, subjects, no doubt, *also* make a normative judgment that holding these anti-environment evaluative attitudes is morally right or wrong. But for the Deep Self Model to count as a version of the Bidirectional Model, it must be shown that these normative judgments actually influence intentionality judgments. However, as I argue in the remainder of the paper, in each of the asymmetry cases discussed in this paper, subjects' normative judgments are essentially *epiphenomenal*. That is, these

[5] In statistical testing, subjects' ratings were compared with '4', the middle point on the scale, since the extent to which ratings fall below '4' represents the degree to which the Chairman is rated as being anti-environment. By this measure, the Chairman was found to significantly anti-environment in both conditions (Harm Condition: $T(19) = 11.02$, $p < 0.001$, Help Condition: $T(19) = 6.30$, $p < 0.001$). Subjects' rating were also compared between the two conditions, which showed that the Chairman is judged as being more anti-environment in the Harm Condition than in the Help Condition ($T(38) = 2.85$, $p = 0.007$).

normative judgments don't influence intentionality judgments, and instead it is concordance between the agent's imputed deep attitudes and the outcome that actually drives asymmetric intentionality judgments.

Another objection, which also holds that the Deep Self Model is a version of the Bidirectional Model, is also worth discussing. According to this objection, people's judgments about an agent's underlying deep attitudes might themselves be influenced by normative considerations. For example, in the Chairman case, it might be that people determine whether an agent truly counts as 'pro-environment' or 'anti-environment' by asking where the agent's attitudes towards the environment lie relative to some normatively acceptable standard. If this is the case then, normative standards regarding how a person ought to behave with respect to the environment are *implicitly* in fact at work when people attribute to an agent that he or she is pro or anti-environment.[6]

In addressing this objection, its worth first stepping back a bit and recognizing that in a large number of *other* cases in which we attribute to a person pro- or anti-attitudes, these attitudes *aren't* attributed relative to some normative standard. Rather, they are attributed based on factual considerations such as the person's verbal utterances, actions, or other kind of evidence, and these attributed attitudes are intended to play a purely *descriptive role* in helping to explain and predict that person's subsequent behavior. For example, if when Paul is presented with a plan to go to new Italian restaurant, he says 'I don't care at all for Italian food', this statement serves as a factual consideration that supports the inference that Paul is in some sense anti-Italian food (perhaps he can only handle bland foods). And once this attribution is made, the attributed attitudes help us explain why Paul is disinclined towards certain behaviors (such as driving all the way across town to go to the new Italian restaurant). In this case, the attribution that Paul is anti-Italian food is *not* made relative to a normative standard. In particular, there is no normative obligation that a person *ought* to like Italian food. Rather, this attribution is based purely on the factual evidence (e.g., Paul's verbal utterances).

But much like in the case of Paul, in the Chairman case the Chairman says he doesn't care at all about harming [helping] the environment, and this statement allows subjects to infer that the Chairman is in some sense anti-environment. This attribution in turn helps subjects explain why the Chairman is disinclined to certain behaviors (such as putting a lot of effort into bringing about an outcome in which the environment is helped). It seems then that the case of Paul and the Chairman case are exactly parallel. But if the attitude of being anti-Italian food isn't attributed to Paul based on normative considerations, but rather is attributed based on factual considerations for the purpose of helping to predict and explain Paul's subsequent behavior, then it follows that the highly similar pattern of attributions made in the Chairman case should also be seen as based on factual rather than normative considerations.

[6] I thank an anonymous reviewer at Philosophical Studies for helping me see and formulate this objection.

3.3 Knobe's bad beats good principle and the Rifle cases

Knobe's own interpretation of the Chairman case utilizes a model that accepts the Bidirectional Thesis. In particular, he argues for the view that people are considerably more willing to say that an agent brought about an outcome intentionally when they regard the outcome as morally bad than when they regard the outcome as morally good (Knobe 2003). This is because morally bad outcomes facilitate backtracking influences from the normative judgment that the outcome is morally bad to the intentionality judgment. Morally good outcomes, in contrast, don't generate this backtracking effect to the same degree. We can call this view the Bad Beats Good Principle, and it does in fact predict the asymmetry found in the Chairman case.[7] But as we shall see in the following series of cases, again due to the philosopher Joshua Knobe, the Bad Beats Good Principle incorrectly predicts peoples' intuitions in certain key cases, while the Deep Self Model captures these intuitions.

Consider the following three somewhat similar cases:

Rifle Contest

Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bulls-eye. He raises the rifle, gets the bull's eye in the sights, and presses the trigger. But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…Nonetheless, the bullet lands directly on the bull's-eye. Jake wins the contest.

Aunt Killer

Jake desperately wants to have more money. He knows that he will inherit a lot of money when his aunt dies. One day, he sees his aunt walking by the window. He raises his rifle, gets her in the sights, and presses the trigger. But John isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…Nonetheless, the bullet hits her directly in the heart. She dies instantly.

Selfless Soldier

Klaus is a soldier in the German army during World War II. His regiment has been sent on a mission that he believes to be deeply immoral. He knows that many innocent people will die unless he can somehow stop the mission before it is completed. One day, it occurs to him that the best way to sabotage the mission would be to shoot a bullet into his own regiment's communication device.

He knows that, if he gets caught shooting the device, he may be imprisoned, tortured or even killed. He could try to pretend that he was simply making a mistake – that he just got confused and thought the device belonged to the enemy – but he is almost certain that no one will believe him. With that

[7] Knobe actually argues that the *moral* badness of the outcome is not needed. Any kind of badness, moral or not, is sufficient to generate the backtracking influence from the normative judgment that the outcome is bad to the intentionality judgment. I ignore this complication as it is not relevant to my overall argument.

thought in mind, he raises his rifle, gets the device in his sights, and presses the trigger. But Klaus isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…Nonetheless, the bullet lands directly in the communications device. The mission is foiled, and many innocent lives are saved.

Knobe found that just 23% of subjects say the agent intentionally hit the target in the Rifle Contest vignette. However, 91% of subjects say the agent intentionally hit the target in the Aunt Killer vignette (Knobe 2006). This result seems to bear out Knobe's Bad Beats Good Principle, since people were considerably more willing to say that an agent brought about an outcome intentionally when they regard the outcome as morally bad than in the case where they don't. However, 92% of subjects also said the agent intentionally hit the target in the Selfless Soldier vignette (Knobe 2006). This result is highly problematic for Knobe because subjects were equally willing to say that the Selfless Soldier and Aunt Killer brought about their respective outcomes, despite that fact that the outcome brought about by the Aunt Killer is bad while the outcome brought about by Selfless Soldier is good.

The Deep Self Model, however, correctly predicts the results of these cases based on the degree of concordance between attributed deep attitudes and the outcome in each case. In the Aunt Killer case, subjects are led to make deep attributions such as that the agent 'values money over human life', 'desires his own wellbeing over that of others', and 'cares little about human suffering', all of which strongly concord with the outcome (i.e., killing the Aunt). The Selfish Soldier case also licenses a host of deep attitude attributions, including that the agent 'values human life', 'prioritizes the protection of innocents', and 'cares more about helping others than saving himself', where these attitudes also strongly concord with the outcome (i.e. saving innocents from attack). In the Rifle Contest case, in contrast, it is not possible to make robust deep attitude attributions. Indeed, the agent in the Rifle Contest Case somewhat bafflingly enters a rifle contest without having any skill at all at using a rifle, leaving one to tentatively infer that the agent must not have any *deep* commitment to winning the contest, since a person who genuinely values winning the contest would have presumably learned how to shoot a rifle. Since there is strong concordance between deep attitudes and outcome in the Aunt Killer and Selfless Soldier cases, but this concordance is lacking in the Rifle Contest case, the Deep Self Model predicts that subjects are more likely to judge that the agent brought about the outcome intentionally in the Aunt Killer and Selfless Soldier cases compared to the Rifle Contest case, which is precisely the asymmetry that was found.

To underscore the point that it is differential concordance with deep attitudes, and not the differential moral valence of the outcome, that drives the asymmetry in these three cases, consider the following case that is similar to the original Rifle Contest case in it's moral neutrality, except in this case, the outcome does in fact strongly concord with deep attitudes of the agent:

Policeman Rifle Contest

Ever since he was a little boy, Jake has wanted to be a police officer more than anything else, and he would be devastated if his dream never came true. So

> Jake decides to go to register at the local police academy. However, the police academy has too many recruits for its entering class. In order to decide who will stay and who will be cut, the recruits have a competition to see who can hit a target with a rifle.
>
> Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bulls-eye. He raises the rifle, gets the bull's eye in the sights, and presses the trigger. But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…Nonetheless, the bullet lands directly on the bull's-eye. Jake wins the contest.

Notice that in this vignette, the second paragraph is identical to Knobe's original Rifle Contest vignette. However, unlike the original Rifle Contest vignette, the first paragraph makes it clear that the agent does possess central and stable desires and values that concord with the outcome. I presented the Policeman Rifle Contest vignette to 40 University of Michigan undergraduates. Subjects were asked to rate their agreement with the statement 'Jake intentionally hit the target' on a 6 point scale (1 = Agree and 6 = Disagree). The mean response was 2.0 with a considerable majority of subjects, 90%, providing responses that were to the left of the midline indicating a preponderance of agreement with the statement that the agent intentionally hit the target. Of note, this percentage is highly similar to that obtained in Knobe's Aunt Killer and Selfish Soldier cases.

Notice that the Bad Beats Good Principle cannot predict the pattern of results in the preceding four cases. One problem for the Bad Beats Good Principle is that subjects were equally willing to say that the agent in the Aunt Killer, Selfless Soldier, and Police Rifle Contest cases brought about their respective outcomes intentionally, despite that fact that the outcome brought about in Aunt Killer is morally bad, the outcome brought in Selfish Soldier is morally good and the outcome brought about in Police Rifle Contest is morally neutral. But a second, less obvious problem is that the Bad Beats Good Principle cannot *itself* provide any explanation for why subjects are *unwilling* to say that the agent in the Rifle Contest case hit the target intentionally (it can only predict that judgments that the agent brought about the outcome intentionally will be *relatively* higher for the Aunt Killer, who brings about a morally bad outcome, than the Rifle Contest winner, who doesn't).

The Deep Self Model, in contrast, explains all four cases using the same strategy: The fact that that the Rifle Contest winner lacks any enduring aspect of his Deep Self that concords with the outcome explains why he isn't judged as bringing about his outcome intentionally, while the fact that the Aunt Killer, Selfless Soldier, and Police Rifle Contest winner *do* possess aspects of their respective Deep Selves that concord with their respective outcomes explains why they *are* judged as having brought about their outcomes intentionally. The fact that the Deep Self Model explains all four cases in the same unified manner provides additional support for the model.

3.4 Emotion versus deliberation cases

Let's turn now to another class of asymmetry findings from the experimental literature that probes how subjects differentially regard actions that arise from

emotion versus deliberation. Pizarro et al. (2003) constructed a series of vignettes about agents performing actions resulting from potent emotion. In some vignettes, the agent brings about a positive consequence:

a. Because of his overwhelming and uncontrollable sympathy, Jack impulsively gave the homeless man his only jacket even though it was freezing outside.

In other vignettes, the agent brings about a negative consequence:

b. Because of his overwhelming and uncontrollable anger, Jack impulsively smashed the window of the car parked in front of him because it was parked too close to his.

For each of these vignettes, Pizarro et al. constructed analogous cases that differed only in that the agent acted calmly and deliberately, rather than from potent emotion. For example:

c. Jack calmly and deliberately gave the homeless man his only jacket even though it was freezing outside.

d. Jack calmly and deliberately smashed the window of the car parked in front of him because it was parked too close to his.

Subjects were asked to rate 'How much praise or blame does Jack deserve?' on a 9 point scale (0 = Extreme Praise, 5 = Neither, 9 = Extreme Blame). The results demonstrated a striking asymmetry. Subjects gave the agent considerably less blame for bringing about a negative outcome when those behaviors were the result of overwhelming emotion. But when the agent brought about a positive consequence, there was no corresponding effect; subjects assigned just as much praise when the agent acted on overwhelming emotion as when the agent acted after calm deliberation.[8]

To explore the role of an agent's having acted due to emotion versus calm deliberation on judgments of intentionality (as opposed to judgments of praise and blame), I re-ran the four vignettes listed above with 80 University of Michigan undergraduates (20 subjects saw each vignette).[9] This time, subjects were asked to rate their agreement with the statement 'Jack intentionally gave the homeless man his jacket [smashed the window of the car]' on a six point scale (1 = Agree, 6 = Disagree). Results followed a pattern highly similar to the original Pizarro and colleagues study. When the agent brought about a negative consequence, subject's ratings of intentionality were considerably less for the case where he acted from strong emotion compared to the case where he acted from calm deliberation. When

[8] For the Pizarro et al. study, ratings were as follows: Case *a* 1.4, *b* 6.1, *c* 1.6, *d* 7.9, and there was a significant outcome valence by decision process (i.e. emotion versus deliberation) interaction driven by the fact that case *b* differed from case *d* ($p < 0.05$).

[9] Since 'deliberately' can be regarded as being inappropriately close in meaning to 'intentionally', vignettes *c* and *d* were modified slightly to read: *c*. After calm deliberation, Jack gave the homeless… and *d*. After calm deliberation, Jack smashed the window….

the agent brought about a positive consequence, however, subject's ratings of intentionality did not correspondingly differ.[10]

Notice that standard versions of the Bidirectional Model, for example, Knobe's Bad Beats Good Principle, can not explain this pattern of results. Indeed, in the preceding series of vignettes, 'good' beats 'bad' in that subjects were more likely to judge that Jack intentionally brought about the outcome when he brings about a positive outcome due to strong emotion relative versus when he brings about a negative outcome due to strong emotion.

The Deep Self Model, however, provides a ready explanation for the preceding pattern of judgments. The key to the explanation is to recognize that there are differences in how actions performed in the context of strong emotions versus calm deliberation impact attributive inferences about the contents of an agent's Deep Self. Other things being equal, when someone acts based on calm deliberation, the results provide a good basis to infer underlying aspects of the agent's Deep Self. So when Jack calmly and deliberately smashes a car window, subjects infer that he possesses pro-destruction and pro-flouting of the law attitudes as part of his Deep Self. When Jack calmly and deliberately helps a homeless person, subjects infer he possesses pro-charity attitudes as part of his Deep Self.

However, cases involving strong emotions are quite different—these cases *asymmetrically* license inferences about underlying aspects of an agent's Deep Self with positive outcomes, but not negative ones. The reason is that it appears that folk psychology contains an important tacit principle: A person that holds deeply to values that prohibit harm can nevertheless perform a harm-producing action under the influence of strong emotion. In other words, other things being equal, harm-producing actions performed under strong emotion don't necessarily reflect the agent's Deep Self. This principle underlies 'heat of passion' defenses used to explain how an otherwise decent person could end up committing an act of violence or other kinds of hurtful actions. It follows then that when Jack helps a homeless person based on strong emotion, subjects infer that he possesses pro-charity attitudes as part of his Deep Self. But when he smashes the car window based on strong emotion, they don't infer that he possesses pro-smashing attitudes as part of Deep Self. Indeed, if they are more charitable, they may even infer that his Deep Self actively *repudiates* these motives.[11]

Having clarified the inferences that subjects plausibly make about underlying aspects of the agent's Deep Self in the preceding vignettes, the Deep Self Model makes a straightforward predictions about subjects' judgments. It predicts that in cases *a*, *c* and *d*, subjects will be more likely to judge that Jack intentionally brought

[10] For my study, ratings were as follows: Case *a* 1.7, *b* 3.2, *c* 1.65, *d* 1.45. Like Pizarro and colleagues, my study also found a significant valence by decision process (i.e. emotion versus deliberation) interaction driven by the fact that case *b* differed from the others ($F(1,76) = 7.79$, $p = 0.007$).

[11] Pizarro and colleagues did in fact ask subjects the degree to which the agent in each vignette possessed second-order desires that are consistent with impulses to perform the behaviors specified in the vignette. They found no differences across vignettes in judgments that the agent possessed second-order desires consistent with the relevant behaviors, *except for* the vignette in which Jack smashed the car window due to strong emotion. In this vignette, subjects' judgments of consistency with second-order desires is significantly reduced, suggesting that Jack is fact imputed by subjects with higher-order desires that repudiate his smashing the window.

about the outcome, because the outcome concords with his inferred deep attitudes. In case *b*, however, they will be less likely to judge that Jack intentionally brought about the outcome, because the outcome fails to concord, or even actively discords, with his inferred deep attitudes. And indeed, these predictions of the Deep Self Model are born out in subjects' actual pattern of responses.

The manipulation used in the preceding four vignettes, i.e. manipulating whether an action arises from calm deliberation versus strong emotion, provides a powerful method to influence how subjects attribute underlying attitudes to the agent's Deep Self. It is a unique feature of the Deep Self Model that it predicts that this type of manipulation will in turn influence judgments about intentionality. Thus the results of the preceding four cases, which, as noted earlier, are an anomaly for Knobe's Bad Beats Good Principle, are not just explained, but in fact *antecedently predicted*, by the Deep Self Model.

Table 2 Summary of Cases

Case	Attitudes of agent's Deep Self	Nature of outcome brought about	Are Deep Self and outcome concordant or discordant?	Result—Subjects *more* or *less* likely to judge agent intentionally brought about outcome
Chairman—harm condition	Devalues environment	Harms environment	Concordant	More
Chairman—help condition	Devalues environment	Helps environment	Discordant	Less
Rifle Contest	Not genuinely committed to winning contest	Wins contest	Discordant	Less
Aunt Killer	Values money/ devalues human life	Kills Aunt	Concordant	More
Selfless Soldier	Values saving innocents	Saves innocents	Concordant	More
Police Rifle Contest	Values being a police officer	Joins the police academy	Concordant	More
Jack—a	Values charity	Charitable	Concordant	More
Jack—b	None	Bad	Discordant	Less
Jack—c	Values charity	Charitable	Concordant	More
Jack—d	Devalues other's property	Destructive	Concordant	More

Findings from studies identifying asymmetric patterns of intentionality judgments. In each case, the Deep Self Model correctly predicts the asymmetry that was actually found

4 Conclusion

In this paper, I presented and defended the Deep Self Model of intentionality judgments. This model claims that people make an intuitive distinction between two parts of an agent's psychology, an Acting Self that contains the desires, means-end beliefs and intentions that are the immediate causal source of an agent's actions, and a Deep Self, which contains an agent's stable and central psychological attitudes, including the agent's values, principles, life goals, and other more fundamental attitudes. According to the model, an agent is more likely to be judged to have brought about an outcome intentionally if that outcome is concordant with the agent's underlying Deep Self. The Deep Self Model agrees with more traditional models that subjects assess a host of other factors in making judgments about intentionality (in particular, factors relevant to assessing whether the agent chose the outcome and controls its occurrence). But the model highlights that inferences about the attitudes contained in the agent's Deep Self play a pervasive and heretofore underappreciated role in influencing intentionality judgments.

The specific cases discussed in this paper evidence the power of the Deep Self Model to predict the very complex pattern of asymmetric judgments documented in the experimental literature (see Table 2). Many theorists have attempted to capture the pattern of subjects' judgments in cases such as these by accepting the so-called Bidirectional Thesis. According to this thesis, in making judgments about intentionality, subjects allow normative judgments about the agent's behavior to exert inappropriate backtracking influences that shape judgments about intentionality. But as I've argued in this paper, standard approaches to explaining the asymmetry findings that accept the Bidirectional Thesis, such as the Bad Beats Good Principle due to Joshua Knobe, incorrectly predict the results of many of these cases, and leave other cases largely unexplained. Thus the Deep Self Model explains the complex pattern of judgment asymmetries found in the experimental literature, and does so in a way that has advantages over competing models.

Acknowledgements Thanks to Joshua Knobe, Thomas Nadelhoffer, Nina Strohminger, Erica Roedder, Sven Nyholm, Angela Mendelovici, and audiences at the Moral Psychology Research Group and Princeton Moral Psychology Conference for invaluable feedback.

References

Arpaly, N., & Schroeder, T. (1999). Praise, blame and the whole self. *Philosophical Studies, 93*, 161–188.

Doris, J. M., Knobe, J., & Woolfolk, R. L. (2007). Variantism about responsibility. *Philosophical Perspectives, 21*, 183–214.

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.

Malle, B., & Knobe, J. (1997). The Folk concept of intentional action. *Journal of Experimental Social Psychology, 33*, 101–121.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*, 267–272.

Wolf, S. (1990). *Freedom within reason*. New York: Oxford University Press.