Philosophical Psychology, 2014
Vol. 28, No. 4, 510–528, http://dx.doi.org/10.1080/09515089.2013.867584

Routledge
Taylor & Francis Group

Pushing the Intuitions behind Moral Internalism

Derek Leben and Kristine Wilckens

Moral Internalism proposes a necessary link between judging that an action is right/wrong and being motivated to perform/avoid that action. Internalism is central to many arguments within ethics, including the claim that moral judgments are not beliefs, and the claim that certain types of moral skepticism are incoherent. However, most of the basis for accepting Internalism rests on intuitions that have recently been called into question by empirical work. This paper further investigates the intuitions behind Internalism. Three experiments show not only that these intuitions are not widespread, but that they are significantly influenced by normative evaluations of the situation in question. These results are taken to undermine Internalist intuitions, and contribute to the growing body of evidence showing that normative evaluations influence supposedly non-normative judgments.

Keywords: Folk Psychology; Internalism; Morality; Motivation; Non-Cognitivism

1. Introduction

Moral Internalism claims that there is a necessary connection between judging that some action is morally right/wrong and being motivated to perform/avoid that action. For instance, if I sincerely believe that it is morally wrong to eat animals, then I would be automatically motivated not to eat animals. If I sincerely believe that it is morally required for me to take care of my children, then I would be automatically motivated to take care of my children. This claim is called 'Internalism' (or more technically, 'Motivational Judgment Internalism') because in such cases, the motivation is *internal* to the evaluative judgment. There are different types of Moral

Internalism, but we will here be concerned with the conceptual variety advocated by Hare (1952), which claims that the link between moral judgments and motivation is an a priori conceptual truth.

Derek Leben is Assistant Professor in the Philosophy Department at the University of Pittsburgh at Johnstown. Kristine Wilckens is a post-doctoral scholar in the Department of Psychiatry at the University of Pittsburgh. Correspondence to: Derek Leben, University of Pittsburgh at Johnstown—Philosophy, 450 Schoolhouse Rd., Johnstown, PA 15904, USA. Email: leben@pitt.edu

The fact that Internalism appears intuitively to be true specifically for moral judgments has been extremely important to moral philosophers. In response to the skeptical question: "Why should I care about right and wrong?" some ethicists have argued that the question is nonsensical, since by making judgments about right and wrong, one is automatically motivated to care about these judgments. In response to the question: "What kind of judgments are moral judgments?" philosophers going back to Hume have argued that beliefs like "my car is black" or "today is Tuesday" can never in themselves motivate or direct anyone to perform some action, but only in conjunction with an emotion. If one adopts this Humean theory of motivation along with Moral Internalism, then, as Hume states, "it is impossible that the distinction betwixt moral good and evil can be made by reason; since that distinction has an influence on our actions, of which reason alone is incapable" (1739, p. 205). In other words, since beliefs are never inherently motivating, moral judgments cannot be normal beliefs about the world. This conclusion is known as (psychological) non-cognitivism, and has obvious consequences for how we engage in moral debate and consideration.

As a conceptual claim, debates about the truth of Internalism typically focus on the conceivability of counterexamples: characters without a necessary link between moral judgment and motivation. One such character is a person who sincerely believes it is morally required for her to do X, and yet has no motivation to do X. This person is called an 'amoralist'. The philosophical debate between Internalists and Externalists usually goes something like this:

Internalists propose the intuitive link between moral judgment and motivation. Externalists push back with the conceivability amoralists or other related characters. Internalists then stand their ground, or modify their position to make these characters outliers (see Björklund, Björnsson, Eriksson, Francen Olinder, & Strandberg, 2012 for a review).

Recently, a number of researchers have begun empirically investigating Internalist intuitions. Nichols (2004) presented two scenarios: one where a psychopath claims to know that hurting others is wrong; and one where a harmless mathematician claims to know that hurting others is wrong. In both cases, it is emphasized that neither the psychopath nor the mathematician has any emotional reaction towards hurting other people:

> John is a psychopathic criminal. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. John has hurt and indeed killed other people when he has wanted to steal their money. He says that he knows that hurting others is wrong, but that he just doesn't care if he does things that are wrong. Does John really understand that hurting others is morally wrong?
>
> Bill is a mathematician. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. Nonetheless, Bill never hurts other people simply because he thinks that it is irrational to hurt others. He thinks that any rational person would be like him and not hurt other people. Does Bill really understand that hurting others is morally wrong?

Nichols found that 85% of people were inclined to attribute real moral understanding to the psychopath, but less than half attributed understanding to the mathematician. Nichols took this as evidence against Internalism being a conceptual "platitude."

In another study, Strandberg and Björklund (2012) give reason to think that many people find the possibility of amoralists to be quite conceivable. Participants in this study were presented with the following scenario:

> Anna is watching a TV program about a famine in Sudan. In the TV program, it is shown how the starving are suffering and desperately looking for food. At the same time, Anna is not motivated at all, not to any extent, to give any money to those who are starving.
>
> *Question*: Could it be the case that Anna thinks she is morally required to give some of her money to the starving even if she not motivated at all to do so?

Given the possible answers of "YES" or "NO," the authors found that significantly more participants answered "YES" rather than "NO." In additional studies, they found similar results when participants were told that Anna is normally functioning, depressed, apathetic, or a member of a society where nobody is ever motivated to give to famine relief. The only difference was when participants were told that Anna is a psychopath; in this case, significantly more people answered "NO" rather than "YES."

Both of the above studies present important results which undermine the claim that Internalism is a conceptual truth. Beyond whether Internalist intuitions are widespread, we are interested in what factors are driving some people to have them, as well as why the researchers discovered the specific effects they did. For instance, why are participants more inclined to attribute judgment to the psychopath than the mathematician? Why are Strandberg and Björklund's participant responses so consistent, even when provided with information that Anna is depressed, apathetic, or part of a generally unmotivated community? Also, despite being the minority, why are there many participants in each condition who do indeed endorse Internalism? Clearly participants are not unanimously externalist, nor are the responses completely random.

One hypothesis is that what is driving Internalist judgments is the normative evaluation of the participants. In other words, the extent which people find giving to famine relief to be required/permissible/forbidden is influencing their judgments of whether an agent could share that evaluation with no motivation. Similarly, the fact that the psychopath actually killed people is influencing judgments about his beliefs, compared to the mathematician who has never hurt anyone. Let us call this the 'Normative Force Hypothesis:'

> **Normative Force Hypothesis (NFH)**: What is driving judgments about Internalism is the force of the evaluator's own normative evaluation of a situation.

This hypothesis is inspired by findings from experimental philosophy suggesting that normative appraisals regularly influence non-normative judgments. In a series of well-known studies, Knobe (2003a, 2003b; Knobe & Burra, 2006) found that a person's

normative evaluation of a situation is likely to influence her judgment about a character's intention. In subsequent studies, Knobe and colleagues found that normative evaluations can influence not only attributions of intention but also causation (Knobe & Fraser, 2008), happiness (Phillips, Misenheimer, & Knobe, 2011), and "true selves" (Newman, Knobe, & Bloom, unpublished manuscript). It has long been known that normative evaluations are intertwined with judgments about causation, happiness, free will, and the self. However, the traditional assumption is that moral evaluation comes as a *result* of these considerations. What is surprising about these findings, as Knobe and Fraser note, is that "the relationship can sometimes go in the opposite direction" (2008, p. 1). Moral evaluations can sometimes have an influence on attributions of free will, causation, intention, and happiness; so it seems reasonable to hypothesize that they may also have an influence on attributions of motivation and belief.

Many ethical philosophers who are opposed to Internalism have provided a similar explanation as an "error theory" for why the claim seems so plausible. For instance, Svavarsdottir (1999, p. 183) claims that the Internalist is so committed to the moral requirement of an action (in both evaluation and motivation) that she could not even imagine anyone holding the same beliefs without sharing her motivation:

> I suspect that in many instances the [I]nternalist intuition reflects not a firm grip on moral concepts, but rather a deep moral commitment that makes it hard for the individual in question to imagine how anyone could be motivationally unaffected by his moral judgments.

Svavarsdottir is endorsing a version of NFH where there is a positive relationship between normative evaluation and Internalist judgment. In other words, as people view an action to be more required or forbidden, they also are more likely to see a necessary link between judging that action wrong/right and having a motivation:

> **NFH (+):** As normative evaluation increases, so do Internalist judgments.

Of course, it is also possible that the relationship is a negative (or inverse) one:

> **NFH (–):** As normative evaluation increases, Internalist judgments decrease.

The following studies will investigate these hypotheses. The ultimate goal is not only a unified account of the effects discovered by Nichols and Strandberg and Björklund, but also a new theory of what might be driving Internalist judgments.

The first study aims at testing people's intuitions about Internalism by comparing a protagonist who is apparently an amoralist in situations that differ in terms of perceived permissibility. For the present purposes, we can consider actions to be measured on a moral scale, from 'required' (i.e., taking care of one's children) to 'forbidden' (i.e., killing the innocent). In between are actions that are morally neutral, such as writing with pen or pencil. If NFH is correct, then situations at different points of the spectrum would produce significantly different responses about Internalism.

To show that these studies are really targeting the effect found by Strandberg and Björklund (2012), the authors wished to keep their story and add a similar story on a different point on the spectrum of moral evaluation. Inspiration for the other story was drawn from Peter Singer's famous article: "Famine, affluence, and morality" (1972).

Singer presents two scenarios; in the first, someone allows a child to die by not jumping in a shallow pond to save him, out of concern for destroying expensive shoes. In the second scenario, someone allows people in a distant country to die by not donating money to charity. Singer argues that there is no relevant difference in the two scenarios, and since the first is morally required, the second is also required. For decades, many students and commentators have rejected Singer's conclusion, often insisting that giving to charity is somewhere between neutral and required, a place on the scale sometimes called 'supererogatory'. These actions might be considered nice, but not required. This seems to set up an excellent comparison of Strandberg and Björklund's original story with a very similar one likely to be judged as morally required.[1]

Each of the two previous studies on Internalism presented very different target questions. In the Nichols study, the vignette specifies that the character believes (and, even stronger, knows) an action to be wrong, yet has no motivation. The question is then whether such a belief is genuine. However, in the Strandberg and Björklund study, it is not revealed what the actual belief of the protagonist is. Instead, participants are told that the protagonist has no motivation, and asked about her belief.[2] It is also possible to ask a third target question which may get at Internalism more directly, which provides the character's belief, but then asks about her motivation (the flip of Strandberg and Björklund's probe). To summarize, the three possible target questions are as follows:

a. Given no motivation, is it possible that someone has a belief?
b. Given belief, is it possible that someone has no motivation?
c. Given belief and no motivation, is it possible that someone *really* has the belief?

In this paper, these three target questions have been split into three separate studies. In the first study (following Strandberg and Björklund), participants are given information about the protagonist's motivation, but not the protagonist's judgment (A). In study 2 (following Nichols), participants are given information about both motivation and judgment and asked about the protagonist's "real" belief (C). Finally, study 3 will present information about motivation, but not judgment (B).

2. Study 1

2.1. Method

220 participants voluntarily agreed to take a study on Amazon Mechanical Turk. Each participant received $0.15 for their participation, which produced an average pay of $8.50/hour. The study was approved by the Institutional Review Board at the authors' institution. 100 participants were randomly assigned to one of the two stories described below. The remaining participants were randomly assigned to one of six control groups.

Participants were told they were participating in a study on moral motivation. They were also instructed that for our purposes "being motivated" means being inclined to do it (again following the design of Strandberg and Björklund). Participants in the

main analysis were presented with one of two scenarios. The first story was nearly identical to the story in Strandberg and Björklund (2012), and expected to be judged by participants as supererogatory:

> **Motivation (Supererogatory)**[3]
> Anna is watching a TV program about a famine in Sudan. In the TV program, it is shown how the starving are suffering and desperately looking for food. At the same time, Anna is not motivated at all, not to any extent, to give any money to those who are starving.

There were two reading comprehension questions asked immediately afterwards to make sure that participants had carefully read the story and understood the important points: "What is the name of the country that the TV program is about?" and "Is Anna motivated at all to give money to help the starving in this country?"

Participants were then presented with the target question: "Is it possible that Anna thinks she is morally required to give some of her money to the starving, even though she isn't motivated at all to do so?" A Likert scale was presented from 1 (No) to 7 (Yes), with 4 marked "Maybe." The next question asked: "Do you think that Anna is morally required to give money to help the starving?" and an identical Likert scale was presented for response. There were then demographic questions about gender, age, and education level. The order of the questions above was carefully chosen to avoid priming. The Internalist judgment was presented prior to the moral evaluation, and demographic information was presented at the end.

Another group of participants were presented with a story which may be viewed by participants as morally required rather than supererogatory. The story is as follows:

> **Motivation (Required):**
> John and Tim work in the same office. John works in sales and Tim works in accounting.
> They sometimes have lunch together, and have always gotten along well. Tim has talked many times about how he used to be a lifeguard in college. One morning, Tim is walking to work with plenty of time to get there. He sees that John's car has crashed into a nearby lake and John is in the water shouting for help. Nobody else appears to be around. At the same time, Tim is not motivated at all, not to any extent, to jump in the lake and save John.

There were then two reading comprehension questions asked: "What department does John work in?" and "Who used to be a lifeguard in college?" Participants were then presented with the target question: "Is it possible that Tim thinks he is morally required to save John, even though he is not motivated at all to do so?" A Likert scale was presented from 1 (No) to 7 (Yes), with 4 marked "Maybe." The next question asked: "Do you think that Tim is morally required to save John?" and an identical Likert scale was presented for response. There were then demographic questions about gender, age, and education level.

The alternate wording groups were run as separate studies. In the alternate wording groups, 120 participants received the same instructions and stories as above, but with slightly different wording in the questions. Instead of 'morally required', these questions used the terms: 'must'; 'should'; or 'ought' (i.e., "Could Tim think that he must/should/ought to save John, even though he is not motivated to do so?"). The

reason for this is to ensure that any effects found in the main analysis group carry over to alternate wording, since one might think that the term 'morally required' is confusing or vague to the layperson. 60 of the participants received the Motivation (Supererogatory) story, and 60 received the Motivation (Required) story. In each condition, 20 participants received 'must', 20 participants received 'should', and 20 participants received 'ought'. There were then identical demographic questions about gender, age, and education level.

2.2. Results and Discussion

The first analysis was simply aimed at confirming the expectation that participants would view the supererogatory condition to be less morally required than the required condition. This hypothesis was confirmed. Participants viewed the required prompt as significantly more morally required, $t(88) = 11.38$, $p < 0.001$. $M_{supererogatory} = 2.16$, $SD = 1.53$. $M_{required} = 6.0$, $SD = 1.664$.

The next analyses were aimed at determining whether there was a difference between evaluative conditions in terms of Internalist judgments. There was indeed a significant difference in Internalist judgments between the two conditions, $t(88) = 2.605$, $p = 0.011$. Subjects who received the required prompt were significantly more likely to attribute possible belief to the character (Figure 1).

There are many differences between the two conditions; for example, the agents have different backgrounds (one used to be a lifeguard, the other has no obvious experience in charity work), the patients have different levels of familiarity to the agent (strangers versus friend), and the participants can relate to the supererogatory situation more easily than to the required situation. These factors can arguably be grouped as factors which jointly determine moral evaluation, yet in order to more convincingly demonstrate that moral evaluation is driving the differences in responses between the two conditions, we performed a Pearson's correlation. This analysis

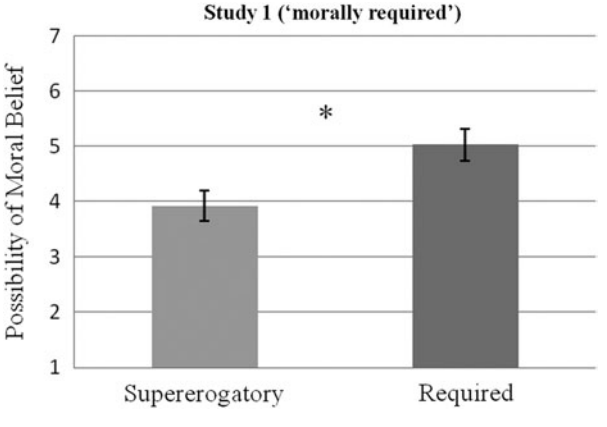

Figure 1. Differences in ratings of the possibility of amoralism by evaluative condition.

revealed a significant correlation between normative evaluation and Internalist judgment, $r = 0.339$, $p = 0.001$.

Given this correlation, it is plausible that the difference between groups is driven by moral evaluation. To directly show that it is indeed moral evaluation which is responsible for the difference between groups, we performed a mediation analysis to assess the difference in Internalist judgments while controlling for evaluation using ANCOVA. After controlling for evaluation the difference between groups in internalist judgment was indeed eliminated, $F(1, 87) = 0.010$, $p = 0.92$, suggesting that group differences were driven by evaluation. Mediation analysis using the bootstrapping method (Preacher & Hayes, 2008) confirmed that evaluation was a significant mediator of this effect. The mediating effect of evaluation on the relationship between condition and Internalist judgment was 0.9938, confidence interval $= 0.0987: 2.0931$. Because the confidence interval did not contain zero, mediation was established and the mediating effect of evaluation was significant. These results suggest that the group difference in Internalist judgment is at least partially mediated by evaluation.

The next analyses were performed on responses from the six control groups. A 2×3 ANOVA with evaluation group (supererogatory, required) and word group (must, should, ought) as independent variables and Internalist judgment as the dependent variable revealed significant main effects of evaluation, $F(1, 89) = 25.219$, $p < 0.00001$, but no main effect of word group, $F(2,89) = 1.567$, $p = 0.214$ (Figure 2). The main effect of evaluation group reflected that Internalist judgments were significantly stronger in the supererogatory compared with the required condition. The main effect of evaluation demonstrates that the effect of normative evaluations on Internalist judgments observed in the first two experiments is consistent across different wordings of the probes.

These results do indeed support NFH. Normative evaluations of the situation significantly affected judgments about the conceptual possibility of amoralism in that situation. Yet surprisingly, the results support a negative relationship between evaluation and Internalism, rather than the positive one predicated by Svavarsdottir. If participants

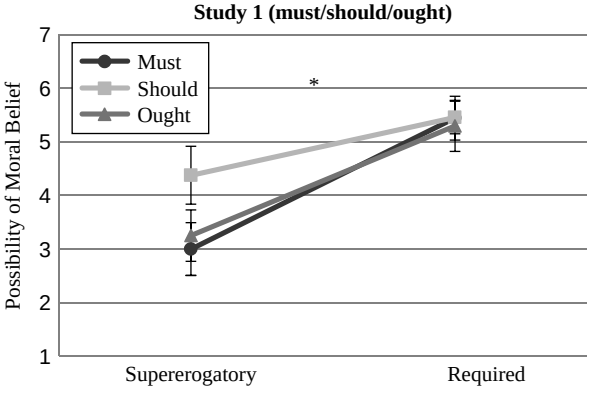

Figure 2. Differences in ratings of the possibility of amoralism by evaluative condition across word type controls.

view a decision as morally required, they are more likely to think that an agent could agree with their judgment without needing any corresponding motivation.

Could it be the case that the responses in supererogatory and required conditions are only reflecting base-rate responses about how likely a character is to have the relevant moral belief? These certainly seem like the kinds of responses you would expect from just giving people the stories and the target questions, without any information about motivation. However, the information about (lack of) motivation *was* provided, and subjects even confirmed that they acknowledged the information in comprehension questions. Given that the subjects are aware of the lack of motivation, it is a very surprising finding that "base rate" assumptions would influence judgments about Internalism. Our hypothesis is that the specific information in base-rate assumptions is normative evaluation, and our correlational analyses seem to support this idea.

3. Study 2

The results of study 1 have not only replicated the findings of and Strandberg and Björklund (2012), but also supported NFH (–). If this hypothesis is true, it should also carry over to the effect found by Nichols (2004). In his study, participants were given information about both the motivation and the belief of the amoralist, and asked whether the amoralist "really" understood their belief. According to NFH, people are more willing to attribute "real" understanding (and thus amoralism) in the case of the psychopath because he actually harmed people, while the mathematician did not, and real harm is judged to be worse than mere approval of harm. Study 2 was intended to determine whether the effect found by Nichols (2004) could be explained by differences in normative evaluation. To keep consistency with the previous studies, the same stories were used in study 2. However, in addition to providing the (lack of) motivation of the character, the stories also specified the belief of the protagonist.

3.1. Method

100 participants voluntarily agreed to take a study on Amazon Mechanical Turk. This study was approved by the Internal Review Board of the authors' institution. Each participant received $0.15 for their participation, which produced an average pay of $7.30/hour. Participants received the same instructions as the previous studies, and were given one of the following two stories specifying both the character's belief and lack of motivation:

> **Belief and Motivation (Supererogatory):**
> Anna is watching a TV program about a famine in Sudan. In the TV program, it is shown how the starving are suffering and desperately looking for food. Nonetheless, Anna does not call the charity number given by the TV program to donate any money.
>
> When asked about the incident afterwards, Anna says she did think she was morally required to give money to those who are starving, but she just had no motivation to do so.

Belief and Motivation (Required):
John and Tim work in the same office. John works in sales and Tim works in accounting. They sometimes have lunch together, and have always gotten along well. Tim has talked many times about how he used to be a lifeguard in college. One morning, Tim is walking to work with plenty of time to get there. He sees that John's car has crashed into a nearby lake and John is in the water shouting for help. Nobody else appears to be around. Nevertheless, Tim continues walking to work and does not jump in the water to save John.

When asked about the incident afterwards, Tim says he did think he was morally required to save John, but he just had no motivation to do so.

There were then the same reading comprehension questions and target question: "Is it possible that [Anna/Tim] really thought [s/he] was morally required to [give money to the starving/save John]?" The evaluation question and demographic questions followed.

3.2. Results and Discussion

The first analysis again confirmed that the supererogatory condition was viewed as less required: $t(76.818) = 13.793$, $p < 0.001$. $M_{supererogatory} = 2.7$, $SD = 1.618$. $M_{required} = 6.4$, $SD = 0.851$. The next analysis showed a difference between evaluative conditions in terms of Internalist judgments. There was indeed a significant difference in Internalist judgments between the two conditions, $t(91) = 2.406$, $p = 0.018$ (Figure 3).

These results suggest that NFH (–) can also explain the effect found by Nichols, who reported that 85% of participants attributed real understanding to the psychopathic amoralist, while less than half attributed real understanding to the mathematician amoralist. Once again, the reason for this would be that participants view merely approving of harm (like deciding not to donate to charity) as having less normative force than directly causing harm (like deciding not to save the drowning coworker). People think that Tim and the psychopath *should* hold certain moral beliefs, so they are

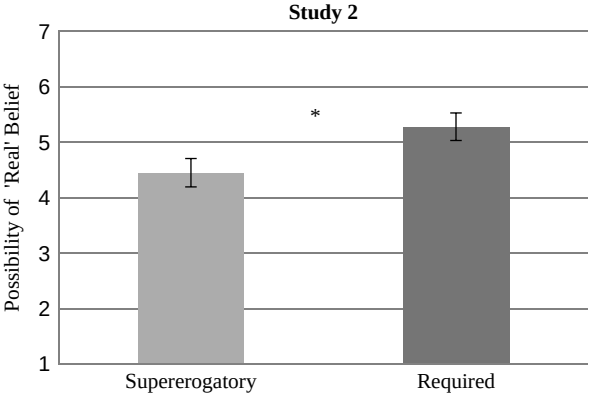

Figure 3. Differences in judgments about "real" belief by evaluative condition.

more likely to attribute moral beliefs to the characters. The results of study 2 suggest that the effect found by Nichols (2004) is attributable to differences in normative evaluation.

4. Study 3

So far the stories given to participants in both conditions have provided information about the character's motivation alone (study 1), as well as both motivation and belief (study 2). Yet it is unclear whether the same effect applies when the character's belief alone is provided. It would seem that NFH (–) predicts the same result as studies 1 and 2 above, since normative evaluation is held constant. Study 3 sought to address this question by including the character's belief but not their motivation.

4.1. Method

100 participants voluntarily agreed to take a study on Amazon Mechanical Turk. Each participant received \$0.15 for their participation, which produced an average pay of \$8.50/hour. The study was approved by the Internal Review Board at authors' institution. Five participants were excluded because they took multiple studies or failed to correctly answer all of the qualifying questions. Participants received the same instructions as the previous studies, and were given one of the following stories:

> **Belief (Supererogatory):**
> Anna is watching a TV program about a famine in Sudan. In the TV program, it is shown how the starving are suffering and desperately looking for food. At the same time, Anna thinks she is morally required to give money to those who are starving.
>
> **Motivation (Required):**
> John and Tim work in the same office. John works in sales and Tim works in accounting.
> They sometimes have lunch together, and have always gotten along well. Tim has talked many times about how he used to be a lifeguard in college. One morning, Tim is walking to work with plenty of time to get there. He sees that John's car has crashed into a nearby lake and John is in the water shouting for help. Nobody else appears to be around. At the same time, Tim thinks he is morally required to jump in the lake and save John.

There were then the same reading comprehension questions, except the second was changed to ask about the belief of the character rather than his/her motivation. The target question read: "Is it possible that [Anna/Tim] is not motivated at all to [give money to the starving/save John], even though [s/he] thinks he is morally required to do so?" The evaluation question and demographic questions followed.

4.2. Results and Discussion

The first analysis again confirmed that the supererogatory condition was viewed as less required: $t(89.698) = 12.640$, $p < 0.001$. $M_{supererogatory} = 2.69$, $SD = 1.544$.

$M_{required} = 6.23$, $SD = 1.173$. The next analysis showed a difference between evaluative conditions in terms of Internalist judgments. Unlike the previous studies, there was no significant difference in Internalist judgments between the two conditions, $t(94) = 0.816$, $p = 0.416$ (Figure 4).

Once more, participants showed no signs of being Internalists, since judgments about the characters' motivation varied widely within the two groups with both means centering on the midpoint. However, this result apparently conflicts with NFH (−), which might predict a significant difference in Internalist judgments regardless of whether the belief was included or not. One difference that might be important between study 3 and the previous two is that in studies 1–2, the Tim character had beliefs and/or motivation that were inconsistent with those of the participants. On the other hand, in study 3, the attitudes of both characters are consistent with those of the participants, who generally agree that Anna is not morally required to help the starving, but Tim is morally required to save his coworker. So it seems that the NFH hypothesis that people are more Internalist when evaluating situations of greater normative force is too simplistic. We will next turn to more sophisticated ways of explaining these results, subsuming the influence of moral evaluation, in the general discussion.

5. General Discussion

The aim of these studies has been to expand on research about Moral Internalism done by Nichols (2004) and Strandberg and Björklund (2012), and investigate a hypothesis which promises a unifying explanation of their data. This hypothesis, NFH, claims that people's judgments about Internalism are driven by their normative evaluation of a particular scenario, rather than any necessary links between a character's moral beliefs and his/her motivation.

The first thing to note is that the findings of Nichols (2004) and Strandberg and Björklund (2012) are clearly widespread and robust. It is clear that the intuitions

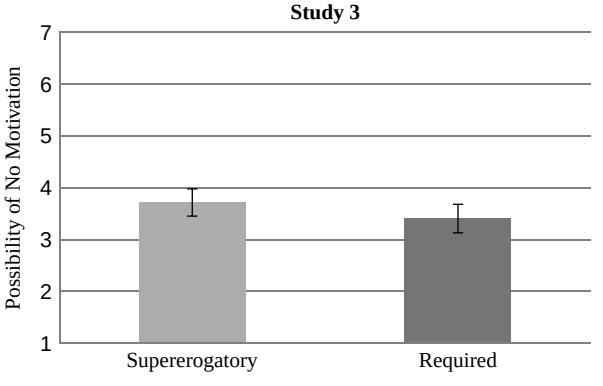

Figure 4. No difference in ratings of motivation by evaluative condition.

behind Internalism are not shared by all members of the community. In fact, in some situations, such as "Motivation (Required)" it is possible to generate almost universal agreement *against* Internalism! Thus, any philosophers inclined to say that the link between moral judgment and motivation is a conceptual necessity are committed to the odd claim that a massive portion of the population is not competent in their own language, or in moral concepts.[4]

As for investigating the Normative Force Hypothesis, all three studies used a "supererogatory" story and a "required" story to compare judgments about Internalism. Study 1 used a method similar to that of Strandberg and Björklund (2012), who asked whether a character could have a moral belief, given no motivation. Study 2 used a method similar to that of Nichols (2004), who asked whether a character could "really" understand/hold a belief, given no motivation. In both studies, participants were significantly more willing to attribute a genuine moral belief to the character in the required condition than in the supererogatory condition. This supports the Normative Force Hypothesis, and perhaps surprisingly, shows that the relationship between evaluation and Internalist judgment is an inverse one. As normative evaluation increases, Internalist judgments appear to decrease. Study 3 carried the first two studies through to their logical complement, using the same two stories and asking whether a character could have no motivation, given a moral belief. The results showed no difference of Internalist judgment between the two conditions, suggesting that the Normative Force Hypothesis is too simplistic a story. Both the conditions in study 3 appear to have the same normative force as in studies 1–2, yet Internalist judgments are the same. Thus, a more sophisticated explanation is needed to show why normative force appears to influence attribution of beliefs, but not attribution of motivation.

We will consider three hypotheses which go beyond the Normative Force Hypothesis to explain the available data. The first is an "affective" hypothesis, where normative force does not cause greater attribution of belief, but both are jointly caused by affective arousal. This hypothesis is inspired by a group of theories which propose that affective arousal causes increases in normative force (Greene, 2007; Prinz, 2007), along with a study by Nichols and Knobe (2007) suggesting that affective arousal also causes increases in attributions of responsibility. If this is true, and attributions of responsibility entail attributions of belief, then it might seem like normative evaluations are really driving Internalist judgments, when the effect is epiphenomenal (Figure 5). Essentially, the idea is that when people are emotionally aroused by a situation they are more likely to hold agents responsible, and this requires saying: "s/he knows what's right and wrong!"

Although affect was not measured in our experiments, it is highly plausible that Anna's story does not lead to increased affective arousal, while Tim's story does. According to the affective hypothesis, this would cause the observed differences in moral evaluation observed across all three studies, as well as the increased tendency to attribute a moral belief to Tim rather than Anna in studies 1 and 2. A major difference between the first two studies and the third is that in studies 1 and 2 it is either implied or directly stated that Tim does not save his coworker, which would arouse people's

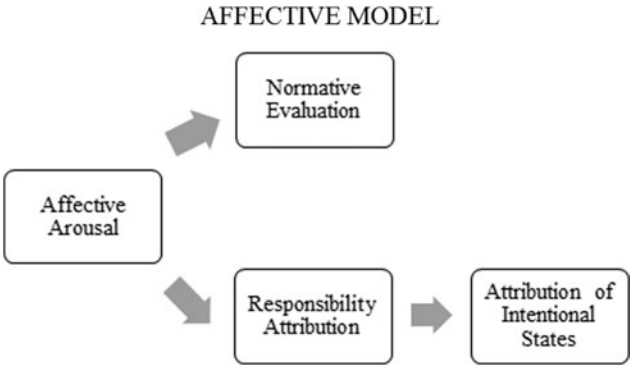

Figure 5. The Affective Model.

anger. However, in study 3, there is no indication that Tim allows his friend to die, and therefore no emotions of anger. This could plausibly explain why the effect disappears in study 3.[5] Additionally, most of Strandberg and Björklund's (2012) cases are low affect and result in low attributions of responsibility, and therefore, midline attributions of moral belief. In the Nichols (2004) cases, affect is low for the mathematician and high for the psychopath (he actually killed people), leading to differences in belief attribution (and presumably differences in moral evaluation, if it were measured).

The second hypothesis to improve on NFH is a "cognitive" hypothesis, where normative evaluation *does* play a causal role in intentional attributions, but mediated by a third factor. One candidate for this third factor is something like Davidson's principle of charity (Davidson, 1973, 1974), which Kauppinen (unpublished manuscript) has applied to some of Knobe's cases. The principle of charity is a hypothesis about a strategy people use to interpret the behavior of others; the basic idea is that "we make maximum sense of the words and thoughts of others when we interpret them in a way that optimizes agreement," while leaving room for "explicable error" (Davidson, 1974, p. 19). Under this strategy, we try to project true beliefs (true according to our own belief system) onto others as much as possible. Assuming that this includes moral beliefs, this means that a person's normative evaluation of a situation will cause her to try to attribute similar moral beliefs to others about that situation as much as possible (Figure 6).

How would the principle of charity help to explain the data? We have seen that in the supererogatory (Anna) case, most people do not view giving to charity as morally required, so according to the principle of charity, they are likely to project a similar belief onto Anna. Yet in the required (Tim) case, most people do view saving his friend as morally required, so they are likely to project a similar belief on to him, even in the second experiment where his actions appear inconsistent with this belief. The strength of normative force may have a strong influence on the principle of charity, as we see in the Nichols (2004) cases, where people are much more likely to attribute genuine belief to the psychopath who actually harmed people rather than the mathematician who did not.

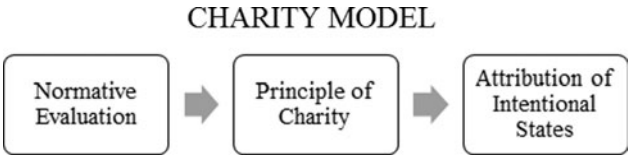

Figure 6. The Charity Model.

Finally, in our third experiment, because the Tim character has already endorsed the moral belief that participants want to attribute to him, they are relatively indifferent about whether he has any corresponding motivation (just as with the supererogatory condition).

A third hypothesis is also "cognitive," presenting a mediating factor between moral evaluation and belief attribution, where this factor is a construction of what features are in the agent's "deep self." While Davidson's principle of charity holds that people simply project their own beliefs onto others, the Deep Self model considered here proposes that people use their own expectations along with other behavioral cues to try and construct a representation of that agent's self, and then determine whether intentional states are consistent with that self (Figure 7). Thus, a person's attribution of an intentional state to an agent relies crucially on whether that state is consistent with their "more stable, enduring, and fundamental attitudes" rather than those attitudes which are more fleeting and temporary (Sripada, 2010 , p. 165). The Deep Self Model originally proposed by Sripada (2010) and Sripada and Konrath (2011) is "unidirectional" because it predicts that moral evaluation is only the outcome of deep self attributions, never the cause. However, Cova and Naar (2012), along with Newman et al. (unpublished manuscript), have proposed bidirectional versions of the Deep Self Model where moral evaluation also plays a role in determining which features are in an agent's deep self. We will not get into the details of this dispute here; rather, we will describe how a general deep self model would interpret our results.

In the Anna (supererogatory) cases, using cues from Anna's lack of motivation and their own expectations about giving to charity (whether descriptive or normative),

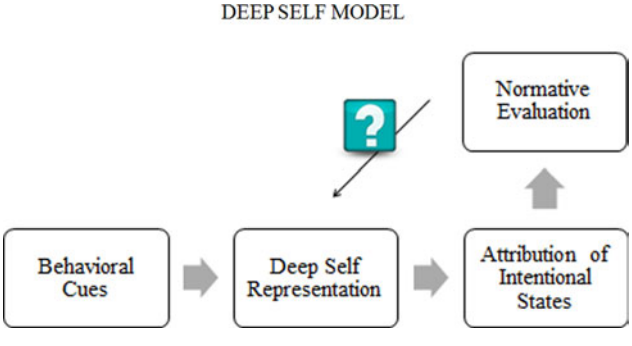

Figure 7. The Deep Self Model.

participants might construct a deep self representation of Anna which is indeterminate about believing that she is required to give to charity. Most additional information given in Strandberg and Björklund's (2012) study, such as being depressed, apathetic, or a member of an apathetic society, all lead to deep-self constructions just as indeterminate about this moral belief. Yet discovering the person is a psychopath might be good enough reason to think that her deep self is inconsistent with the belief that giving to charity is morally required, since stereotypes about psychopaths include them being violent and unhelpful.[6]

In the Tim (required) cases, the information used to construct an agent's deep self representation is quite different. Tim used to be a lifeguard, he is friends with John, and there are no other reasons why he wouldn't believe that saving John is required. Apparently, having a lack of motivation does not prevent constructing such a deep self attribution (and thus still undermines Internalism). The unidirectional deep self model might even argue that in study 2 all of this information is so overwhelming that it even trumps the fact that Tim does not save John. The bidirectional model would suspect that the evaluator's expectations play a significant role here. For example, in studies on *akrasia* or weakness of will, May and Holton (2012) used a character with a belief that X is wrong as well as a desire to do X. They found that which property is attributed to the character's true self (the desire or the belief) is influenced by the evaluator's moral judgments. Here, both Anna and Tim have conflicting belief and motivation, but the bidirectional deep self model predicts that people's evaluation of Tim's situation as morally required leads them to include the belief in his deep self more than in the Anna case. It is less clear how the Deep Self model can account for not attributing more motivation to Tim than to Anna in the third study. Perhaps it is because the model is only designed to explain the attribution of intentional states, and motivation is not properly an intentional state.

Future work will address which of these hypotheses best explains the available data. The central claim of our paper has been that normative evaluation is a key factor in Internalist judgments, and in all of the interpretations of our data, that is indeed the case. Thus, the normative force hypothesis (NFH−) is supported, but it is not the whole story. In addition to normative evaluation, people are also attributing moral beliefs to others based on either affective arousal, a principle of charity, or a deep self representation gathered from behavioral cues.

On a final note, we will add that this work also has consequences for the evidential status of intuitions about Internalism in arguments for non-cognitivism. If either normative evaluation or affective arousal is causally responsible in any significant way for Internalist judgments, this undermines their evidential status. This is because Internalism is a logically distinct evaluation from either normative evaluation or affective arousal. Internalism is a question about mental states: can an agent have a moral judgment without a corresponding motivation? If judgments about mental states are caused by evaluations not related to mental states, then the intuitions behind Internalism are unreliable for the task of forming true beliefs. In other words, any attempts to defend the truth of Internalism on strictly a priori grounds should be abandoned entirely. This is because, as Leben (forthcoming) argues, intuitions are

unreliable evidence for forming true beliefs whenever they are insensitive to their target domain, and there are no independent grounds for their reliability. One way of establishing this insensitivity is by showing that an intuition (and underlying psychological mechanisms) is influenced by irrelevant factors. For example, if a thermometer is influenced by the gender of the person taking the temperature, we have good grounds for dismissing any beliefs using that thermometer as evidence, since gender is unrelated to temperature. Similarly, if intuitions about the mental states of agents are influenced by unrelated factors (normative evaluations of an agent's situation or affective arousal of the evaluator), these intuitions should not be treated as reliable sources of evidence. Of the possible interpretations discussed here, only the unidirectional Deep Self Model would avoid this result.

The Internalist may object by saying that we should therefore dismiss conceptual versions of Moral Externalism as well. We concur, and concede that neither Internalism nor Externalism can be settled on purely a priori grounds. This is the double-meaning behind the title, "pushing" the intuitions (as in jettisoning them for purposes of justification). However, there are empirical methods for evaluating the question of whether moral judgments are necessarily linked to corresponding motivations. These involve real case-studies of those who hold moral beliefs yet do not have the corresponding motivation (e.g., those with depression and psychopathy) (Roskies, 2003). One might still maintain that such putative amoralists do not have genuine moral judgments (Cholbi, 2006; Kennett & Fine, 2008), but as an empirical question, the issue must be resolved by investigating whether these individuals do indeed display the same kind of moral judgments as normal individuals.

Acknowledgements

We would like to thank Caj Strandberg, Joshua Knobe, Jeff Maynes, Jonathon Hricko, John Waterman, and two anonymous reviewers for their helpful comments and suggestions.

Notes

[1] Although this appears to be an uncontroversial assumption, the studies will also rate and compare evaluations of supererogatory with required scenarios just to confirm this empirically.

[2] This does indeed target Moral Internalism, since Internalism is the claim that, necessarily, if one believes that X is morally required, then one is at least a little motivated to do X. It follows from modus tollens that if Internalism is true and someone has no motivation to do X, then they would not have the relevant belief.

[3] The title of the stories is intended to describe what information is provided about the protagonist. In this case, the story is titled "Motivation (Required)" because the protagonist's motivation is given, but not her belief. If only her belief was given, but not her motivation, the story would be called "Belief (Required)."

[4] One common objection to this conclusion is that philosophers are somehow "experts" in the types of issues under discussion, and thus would have more expert intuitions. Since we have not intentionally included professional philosophers as a group, it is possible that their

intuitions may not be susceptible to the effects demonstrated here. However, the small amount of research that has been done comparing the intuitions of philosophers with the general public has not been supportive of this idea (Tobia, Buckwalter, & Stich, 2013; Weinberg, Gonnerman, Buckner, & Alexander, 2010).

[5] Thanks to an anonymous reviewer for this part of the affective model's interpretation.

[6] Although, we think the unidirectional deep self model has difficulty explaining why genuine belief is attributed to the psychopath more than the mathematician in Nichols' cases. The bidirectional model can explain this.

References

Björklund, F., Björnsson, G., Eriksson, J., Francen Olinder, R., & Strandberg, C. (2012). Recent work on moral internalism. *Analysis Reviews*, *72*, 124–137.

Cholbi, M. (2006). Belief attribution and the falsification of motive internalism. *Philosophical Psychology*, *19*, 607–616.

Cova, F., & Naar, M. (2012). Testing Sripada's deep self model. *Philosophical Psychology*, *25*, 647–659.

Davidson, D. (1973). Radical interpretation. *Dialectica*, *27*, 314–328.

Davidson, D. (1974). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, *47*, 5–20.

Greene, J. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (vol. 3, pp. 35–80). Cambridge, MA: MIT Press.

Hare, R. M. (1952). *The language of morals*. Oxford: Oxford University Press.

Hume, D. (1739). *A treatise of human nature*. Oxford: Clarendon.

Kauppinen, A. (unpublished manuscript). *Lovers of the good: Comments on Knobe and Roedder*. Retrieved from http://philosophycommons.typepad.com/opc1/2006/05/joshua-knobe-and-erica-roedder.html

Kennett, J., & Fine, C. (2008). Internalism and the evidence from psychopaths and "acquired sociopaths." In W. Sinnott-Armstrong (Ed.), *Moral psychology* (vol. 3, pp. 173–190). Cambridge, MA: MIT Press.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190–193.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*, 309–324.

Knobe, J., & Burra, A. (2006). Intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition*, *6*, 113–132.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (vol. 2, pp. 441–448). Cambridge, MA: MIT Press.

Leben, D. (forthcoming). When psychology undermines beliefs. *Philosophical Psychology*.

May, J., & Holton, R. (2012). What in the world is weakness of will? *Philosophical Studies*, *157*, 341–360.

Newman, G., Knobe, J., & Bloom, P. (unpublished manuscript). Value judgments and the true self.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. Oxford: Oxford University Press.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, *41*, 663–685.

Phillips, J., Misenheimer, L., & knobe, J. (2011). The ordinary concept of happiness (and others like it). *Emotion Review*, *3*, 1–3.

Preacher, K., & Hayes, A. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Prinz, J. (2007). *The emotional construction of morals*. Oxford: Oxford University Press.

Roskies, A. (2003). Are ethical judgments intrinsically motivational? Lessons from "acquired sociopathy". *Philosophical Psychology*, *16*, 51–66.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs*, *1*, 229–243.

Sripada, C. S. (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*, 159–176.

Sripada, C. S., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind & Language*, *26*, 353–380.

Strandberg, C., & Björklund, F. (2012). Is moral internalism supported by folk intuitions? *Philosophical Psychology*, *26*, 319–335.

Svavarsdottir, S. (1999). Moral cognitivism and motivation. *The Philosophical Review*, *108*, 161–219.

Tobia, K., Buckwalter, W., & Stich, S. (2013). Moral intuitions: Are philosophers experts? *Philosophical Psychology*, *26*, 629–638.

Weinberg, J., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? *Philosophical Psychology*, *23*, 331–335.