Forthcoming in *Mind & Language*.

Intuitions and individual differences: The Knobe effect revisited[*]

SHAUN NICHOLS
JOSEPH ULATOWSKI

Abstract: Recent work by Joshua Knobe indicates that people's intuitions about whether an action was *intentional* depends on whether the outcome is good or bad. This paper argues that part of the explanation for this effect is that there are stable individual differences in how 'intentional' gets interpreted. That is, in Knobe's cases, different people interpret the term in different ways. This *interpretive diversity* of 'intentional' opens up a new avenue to help explain Knobe's results. Furthermore, the paper argues that the use of intuitions in philosophy is complicated by fact that there are robust individual differences in intuitions about matters of philosophical concern.

Joshua Knobe has produced perhaps the most intriguing set of data in experimental philosophy. He finds that people's intuitions about whether an outcome was *intentionally* produced seem to vary depending on the moral status of the outcome itself (Knobe, 2003a, 2003b, forthcoming). These striking results have attracted attention from numerous different fields of inquiry, including action theory (Mele, 2003; McCann, forthcoming), social psychology (Malle, forthcoming), moral psychology (Hauser, 2006), philosophy of law (Nadelhoffer, forthcoming a), philosophy of language (Adams & Steadman, 2004a, 2004b, forthcoming), and developmental psychology (Leslie et al., forthcoming). What has not yet emerged, though, is a satisfying explanation of the phenomenon. In our view, the literature makes apparent that all of the prominent explanations have serious shortcomings. The puzzle persists.

---

[*] We'd like to thank Fred Adams, Kent Bach, Anne Bezuidenhout, Steve Downes, Michael Gill, Edouard Machery, Elijah Millgram, Ron Mallon, Thomas Nadelhoffer, Paulo Sousa, Jason Turner, Jonathan Weinberg, an anonymous referee, and an editor for *Mind & Language* for comments and discussion on a previous draft. An earlier version of this paper was presented at the Social Cognitive Development group at Harvard University, and we'd like to thank the audience for helpful feedback. Finally and especially, we are extremely grateful to Joshua Knobe for numerous discussions about the material presented here.
Address for correspondence: Shaun Nichols, Department of Philosophy, University of Arizona, Tucson, AZ, USA.
Email: sbn@email.arizona.edu

In this paper, we have two main goals. First, we want to present and defend a new proposal intended to explain part of the phenomenon that Knobe has uncovered.[1] We will argue that part of what drives Knobe's finding is a stable pattern of individual differences in intuitions. In particular, there seems to be diversity in how 'intentional' gets interpreted, and this contributes to a novel explanation of Knobe's results. Our second goal is to argue that the existence of stable individual differences in intuitions has significant implications about the use of intuitions in philosophy.

1. Philosophy, Intuitions and Diversity

Many philosophical problems and projects find their source in intuitions. For instance, when philosophers try to answer questions like 'What is free will?', 'What is knowledge?' or 'What is it for an action to be intentional?', a common strategy is to appeal to our intuitions. We consult our intuitions about various cases to develop the proper account of these philosophically important notions.

This approach in philosophical inquiry plausibly stretches back to Plato, but it is especially prominent in recent work in analytic philosophy. In much work in analytic philosophy, it is quite explicit that the goal is to give an analysis of *folk* concepts (e.g. Jackson, 1998; Gibbard, 1990; Lewis, 1972). If the goal is to analyze folk concepts, one might expect philosophers to ask the folk their opinions. Jackson is admirably clear in acknowledging this:

> I am sometimes asked… why, if conceptual analysis is concerned to elucidate what governs our classificatory practice, don't I advocate doing serious opinion polls on people's responses to various cases? My answer is that I do – when it is necessary (Jackson, 1998, 36-37).

So Jackson maintains that it is appropriate to poll the folk for their intuitions. However, he goes on to claim that such polls are often unnecessary because 'often we know that our own case is typical and so can generalize from it to others.' (Jackson, 1998, 37).

This latter assumption – that we know our own case is typical – has been challenged by recent work on cultural differences in intuitions of philosophical interest. Preliminary evidence indicates that East Asians and Westerners have different intuitions about knowledge (Weinberg et al., 2001) and reference (Machery et al., 2004). These results on cultural diversity in philosophical intuitions have been recruited to pose a problem for traditional armchair conceptual analysis (see also Stich & Weinberg, 2001).

[1] Folk intuitions on intentional actions depend on a complex set of psychological processes, and it would be rash to think our account is a complete story. We will be well satisfied if our proposal captures an important part of the Knobe effect.

One option, one way to save armchair analysis from the threat of cultural diversity, has been to embrace a kind of ethnographic approach. One might say that one is using the armchair methodology to discern the nature of the concepts within one's own culture (cf. Jackson, 1998, 2001). The fact that there is cultural variation in intuitions, then, would not be a problem. For one is merely relying on *intra*-cultural intuitions. This reply is perfectly sensible. And even if our concept of free will or knowledge or intentional action is local to our own culture, it is still an interesting and worthy project to seek an analysis of those culturally local concepts. However, as we'll argue, a further problem looms. For it's possible that there are significant *intra*-cultural differences in philosophically relevant intuitions. The problem takes a more concrete shape in recent work on folk intuitions about intentional action, to which we now turn.

2. The Knobe Effect

In an experiment that has now been replicated several times, Joshua Knobe presented lay subjects with one of these two closely matched scenarios:

> **Harm:** 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.

> **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.

A large majority of subjects given the Harm scenario said that the CEO intentionally harmed the environment; a large majority of subjects given the Help scenario denied that the CEO intentionally helped the environment. What can explain this surprising asymmetry in folk intuition?[2]

[2] The result certainly came as a surprise to us, along with many others, but it is exactly what some philosophers would have predicted (e.g. Harman 1976).

Perhaps we can consult the folk themselves to get an account of 'intentional' on which the asymmetric responses are appropriate. Maybe their explanations will reveal why 'intentional' applies asymmetrically to Harm and Help. We gave the Knobe cases to 85 undergraduates at the University of Utah and asked them to explain why they answered as they did. Here are typical examples of answers from subjects who said that the CEO intentionally harmed the environment:

'He knew the consequences of his actions before he began.'

'He knew that implementing the new program would hurt the environment'

'Because he knew he was going to hurt the environment, but chose to do it anyway'

'The chairman knew that this program would harm the environment.'

'The chairman intentionally harmed the environment because he was forewarned by the Vice President that the new program would harm the environment'

'He knowingly made a decision to start the new program after being told that it would harm the environment'

Most of these examples appeal to the fact that the CEO had foreknowledge of the consequences of his action. Next we turned to people who said that the CEO did not intentionally help the environment. They tended to give much different explanations. Here, the explanations appeal to the CEO's intentions:

'He didn't care. It was an unintended consequence.'

'Because his intention was to make money whether or not it will help the environment'

'He didn't INTEND on helping the environment, he INTENDED on making a profit'

'The chairman's intent was to increase profit'

'His motive was to earn profits'

'His motive was to get as much money as possible'

In the Help case, then, subjects say that the outcome wasn't intentional because it wasn't his *motive*. The problem is, of course, these two types explanations don't present a consistent explanation for the asymmetric responses. The explanations for why the CEO in Harm intentionally harmed the environment would suggest that the CEO in Help *did* intentionally help the environment. And the explanations for why the CEO didn't intentionally help in Help would also indicate that the CEO *didn't* intentionally harm in Harm. Folk explanations seem to be of little use for our problem.

Clearly these explanations do not point to an account of the folk notion of intentional action that will explain why people give asymmetric responses to Harm and Help. The goal of the recent literature has been to give a unified account of the underlying competence and then, if necessary, to explain away the problematic responses. Knobe's approach maintains that the asymmetric responses appropriately

reveal the subjects' competency with the concept of *intentional*.[3] That is, the concept *intentional* really does apply differently to Harm and Help (see Knobe, forthcoming a, b). Alternatively (and more popularly), theorists have maintained that the asymmetric responses do not reflect the correct application of the concept *intentional.* Rather, on these views, the side effects of harming and helping are both unintentional consequence of the actions Subjects' responses to the harm case are somehow a product of a distortion, distraction or other extraneous factor in cognition or language. So, broadly speaking, the two approaches have been to (i) embrace the asymmetry as reflecting the competency with *intentional* or (ii) maintain that one of the asymmetric responses does not reflect the competency with the concept. Unfortunately, none of these accounts seems adequate to the data, as we'll see in the following brief review of the recent literature.

3. Psychological Bias Accounts

One prominent and attractive explanation of the Knobe effect maintains that the majority responses in the Harm condition are the product of a bias. Some feature of the Harm case triggers processes that distort the true expression of our competence with the concept *intentional.* It's useful here to distinguish two different bias theories. According to a *blame-driven bias* account (e.g. Malle, forthcoming), our judgments of blame in Harm distort normal judgment and lead people to judge incorrectly that the agent intentionally harmed the environment. According to an *affect-driven bias* account (e.g. Nadelhoffer, forthcoming a), the emotions triggered by the case lead people to judge incorrectly that the CEO intentionally harmed the environment. In these cases, according to Nadelhoffer, 'affective or emotional responses …. *inappropriately* bias our otherwise rational judgments' (Nadelhoffer, forthcoming a).[4]

Both the blame-driven and the affect-driven accounts meet a serious problem with an important variation on the CEO case. Knobe (forthcoming b) gave subjects the following case:

> In Nazi Germany, there was a law called the 'racial identification law.' The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps.

[3] More precisely, it is the concept of *intentional action* that Knobe is interested in, but for ease of exposition, we will typically use the short form.

[4] Obviously the blame-driven bias account and the affect-driven bias account need not be in opposition. For one might plausibly maintain that blame works its magic through the emotions. However, one might also try to treat blame as having its effects through a non-affective channel. And the problem posed below by Knobe (forthcoming b) applies regardless of whether the views are fused.

> Shortly after this law was passed, the CEO of a small corporation decided to make certain organizational changes.
>
> The Vice-President of the corporation said: 'By making those changes, you'll definitely be increasing our profits. But you'll also be violating the requirements of the racial identification law.'
>
> The CEO said: 'Look, I know that I'll be violating the requirements of the law, but I don't care one bit about that. All I care about is making as much profit as I can. Let's make those organizational changes!'
>
> As soon as the CEO gave this order, the corporation began making the organizational changes.

Subjects were asked whether the CEO *intentionally* violated the requirements of the law. 81% of subjects said that he did do this intentionally. The law he violated was manifestly horrific, and subjects presumably think it's a *good thing* that he violated the law. So it doesn't seem like we can explain the responses by appealing to a bias generated by judgments of blame or feelings of anger. Thus, the bias accounts apparently fail to explain the range of Knobe-style effects.

4. Gricean Accounts

An equally attractive explanation for the Knobe effect maintains that it is the product of the pragmatics of intentional language. The most prominent pragmatic account comes from Fred Adams & Annie Steadman (2004a, 2004b, forthcoming), who rely on ideas about pragmatics developed by Paul Grice (1989).[5] A Gricean explanation of the Knobe effect might be cast as follows: people say that the CEO intentionally harmed the environment because to *deny* that he did it intentionally would have unwanted pragmatic implicatures. For instance, to deny that the CEO intentionally harmed the environment might have the conversational implicature that he is not responsible or not to blame for the bad outcome.

Adams & Steadman suggest that when subjects claim that the CEO intentionally harmed the environment, this does not reflect the 'semantic core' of the concept of intentional action (Adams & Steadman, 2004a). On their view, respondents say that the chairman harmed the environment intentionally because they want to blame him for hurting the environment, and 'judging the actions to be "intentional" in the harm condition pragmatically implies strengthened blame' (Adams & Steadman, forthcoming, 11). The respondents want to blame the chairman, so they say that the chairman harmed the

[5] It's worth noting that more recent accounts of pragmatics (e.g., Sperber & Wilson, 1996) diverge in important ways from traditional Gricean pragmatics.

environment intentionally, even though 'they are likely not doing a mental check for consistency' (Adams & Steadman, forthcoming, 11). Thus, according to Adams & Steadman, the conversational implicature prompts respondents to say that the chairman acted intentionally. This response results from social training in the pragmatic use of intentional language.

One way to test the Gricean account of the Knobe effect is to explore whether subjects' responses will change when the pragmatic implicatures are altered appropriately. To their credit, Adams & Steadman pursue this course by giving subjects options that were unavailable in Knobe's experiment (Adams & Steadman, forthcoming). Knobe's experiments only offer respondents the opportunity to choose between whether the chairman acted intentionally or unintentionally. Adams & Steadman predicted that 'if given the option of saying the chairman harmed the environment knowingly versus harmed the environment intentionally, a significant number would opt for the former' (Adams & Steadman, forthcoming, 17). The idea, which seems quite plausible, is that if respondents could answer that the chairman acted knowingly but not intentionally, then this would relieve the pragmatic pressures that lead to the asymmetric responses; for subjects could maintain that the CEO is to *blame* for the harm (since he *knowingly* harmed the environment) and still acknowledge that the CEO didn't *intentionally* harm the environment.

In their experiment, subjects received Knobe's harm vignette and they were asked to select the best answer: either 'The chairman harmed the environment knowingly, but not intentionally' or 'The chairman harmed the environment knowingly and intentionally.' 20% of respondents answered 'knowingly but not intentionally' and 80% of respondents said that the action was done 'knowingly and intentionally.' Adams & Steadman point out that this shows that some people can distinguish *knowingly* from *intentionally*. But we think the more striking fact here is that the results almost exactly duplicate Knobe's original results. Adams & Steadman recognize this, of course, and they suggest that the reason the pragmatic manipulation doesn't alter the Knobe results is because 'the life long habit and social training of pragmatic use of intentional language to assign blame is hard, if not impossible, to override' (Adams & Steadman, forthcoming, 21). Perhaps this is the case, but we take the most important point to be that, since Adams & Steadman's results reproduce Knobe's results, the pragmatic prediction did not pan out. Thus, the experiment does not support the Gricean account of the Knobe effect.[6]

[6]Adams & Steadman report another experiment in which the outcome is morally neutral, and in this experiment, they find that most subjects do distinguish 'intentionally' from 'knowingly'. Further, they report that the Delaware legal code discriminates between actions done knowingly and those done intentionally. But it's not clear how this changes the dialectic since most participants in the debate

In a bout of optimism about the Gricean account, one of us (Nichols) conducted a related experiment designed to demonstrate that the Knobe effect derives from conversational implicatures. Subjects were randomly assigned to one of two conditions, each of which was given Knobe's Harm case. In condition A (modeled on the original Knobe experiment), subjects were simply asked to indicate which claim best describes the chairman's action (i) 'The chairman *intentionally* harmed the environment' or (ii) 'The chairman didn't *intentionally* harm the environment'. In condition B (inspired by the Gricean account), the choice was between (i) 'The chairman *intentionally* harmed the environment, and he is responsible for it' or (ii) 'The chairman didn't *intentionally* harm the environment, but he is responsible for it.' The prediction was that in Condition B, subjects would be more willing to say that the chairman didn't *intentionally* harm the environment, since the unwanted implicature (lack of responsibility) is explicitly denied. As with the Adams and Steadman study, the prediction of the Gricean account was not borne out. There was no significant difference between conditions (χ^2 (1, N=58) = .009, p=.923, n.s.). Indeed, the raw percentages of those who said the chairman intentionally harmed the environment were almost identical in both conditions (68% in condition A and 67% in condition B). Once again, it seems that manipulating the pragmatic features of the case has no effect on the outcome. Thus, although it has immediate appeal, the Gricean account of the Knobe effect lacks the empirical confirmation that one would have wanted. There is no evidence that the asymmetry between Harm and Help is primarily due to Gricean factors.[7]

5. The Competence Account and Individual Differences

Knobe's own explanation of the effect embraces the asymmetry as telling us something fairly amazing about our concept of intentional action. On Knobe's view, the majority responses in Harm and Help are *both right.* In the Harm case, if we properly apply our concept of *intentional*, we will recognize that the CEO did intentionally harm the environment; in the Help case, if we properly apply the concept, we will recognize that the CEO did not intentionally help the environment. Our basic competence with the concept of *intentional* draws on the moral status of an outcome in order to determine whether the outcome

---

(including Knobe and Nadelhoffer) would happily agree that the folk can distinguish 'intentionally' from 'knowingly'.

[7] Nadelhoffer (forthcoming b) raises a further problem for Gricean accounts. He finds that even in cases in which blame is obviously inappropriate, a majority of people still say that an agent intentionally produced an outcome that wasn't intended.

was intentionally produced.[8] This account is quite radical. Knobe maintains that it requires us to overhaul our view of the function of folk psychology as restricted to prediction and explanation. Rather, on his view, 'moral considerations truly do play a role in the fundamental competence underlying people's theory-of-mind capacities' (Knobe, 2005). So the view is quite controversial, but it has the signal virtue that it explains the robustness of the effect -- the reason manipulating the pragmatics and eliminating blame don't make the effect disappear is simply that the majority response is *correct*.

Despite it's empirical virtues, Knobe's proposal has important empirical drawbacks too. The central problem that interests us here is that, while the effects on the CEO case are statistically quite large, the experiments consistently reveal nontrivial minority responses, and those responses require some explanation too. For example, while most people say that the CEO intentionally harmed the environment, in each experiment a significant portion of the sample (from 18% in Knobe, 2003a to 36% in McCann, forthcoming) says that the CEO did not intentionally harm the environment. If the concept of *intentional* really is as Knobe suggests, why is there always a significant minority that maintains that the CEO didn't harm the environment intentionally?

Often when we find a robust majority response to a task, we rightly treat the minority response as noise. Perhaps the subjects weren't paying attention, perhaps they were deliberately misrepresenting their views, perhaps they overthought the task. If you think that the majority response reflects the genuine competence underlying our judgments, then you are likely to treat the minority response as noise or as a kind of performance error. As Stanovich puts it, 'Mean or modal performance might be viewed as centered on the normative response – the response all people are trying to approximate. However, scores will vary around this central tendency due to random performance factors' (Stanovich, 1999, p. 33). Thus, for Knobe's competence account, the natural explanation of the minority responses is that they arise from 'random performance factors'. But given the recurring presence of minority responses with this range of magnitudes, it's far from obvious that we can dismiss the minority responses in this way.

---

[8] Some of Knobe's cases have side effects that are more transparently morally bad. For instance, Knobe (2003) has a case in which a lieutenant says "I don't care at all about what happens to our soldiers"; in one condition, the soldiers are put in the line of fire (and some are killed), and in the other condition, the soldiers are removed from the line of fire (and thus escape death). In this case, people were more likely to say that the lieutenant intentionally put the soldiers in the line of fire than that he intentionally removed them from the line of fire. More generally, Knobe's view is that while the (perceived) moral status of the side effect is one factor that can affect judgments of intentional action, it isn't the only factor that can have such influence; rather, moral evaluations form a subset of a broader set of factors that influence judgments of intentional action.

Indeed, as we will see, the minority responses provide the basis for a new perspective on the Knobe effect.

6. An Experiment on the Knobe Effect

In the interests of intellectual honesty, we're going to present our experiment as we originally envisioned it. We had hoped to find instability in the Knobe results by showing that they were susceptible to order effects. As a result, we gave all subjects ($N = 45$) both the Harm and Help versions of Knobe's CEO cases. Half of the subjects got Harm first and half got Help first. The surveys were completed through an on-line website, designed so that subjects couldn't go back and change their answers. Our prediction was that subjects who got Help first would be more likely to say that the outcome in Harm was unintentional.

To our (initial) dismay, we found no order effects at all. When we compared subjects who got Harm first to subjects who got Help first, there was no significant difference in their responses to either question. Subjects who got Help first were no less likely to say that the CEO intentionally harmed the environment than subjects who got Harm first.[9] And subjects who got Harm first were no more likely to say that the CEO intentionally helped the environment than those who got Help first.[10] Thus our prediction was not confirmed. Indeed, the raw percentages for each case were very similar across conditions. (See chart 1) The Knobe effect continues to be remarkably robust!

Insert chart 1 about here

Before we closed the stats program on another failed attempt to understand the Knobe effect, we noticed a surprising pattern. There was a strong correlation between responses. That is, people tended to answer the same way on the harming and helping questions (r=.516, N=44, p<001, two tailed). Responses split roughly into thirds. One third said neither was intentional; another third said both were, and another third responded asymmetrically. The asymmetric responses were all of the same variety. No one said that the helping was intentional but not the harming.[11] (See table 1)

---

[9] χ^2 (1, N=44) = .043, p=.837, n.s. The number of subjects for this comparison is 44 rather than 45 because one subject neglected to answer the Harm question.

[10] χ^2 (1, N=45) = .085, p=.771, n.s.

[11] Our percentages here differ somewhat from Knobe's original studies on which 82% said 'intentional' on Harm and 23% said 'intentional' on Help. But our numbers in this study are in line with other studies run on undergraduates (e.g., McCann forthcoming) and children (Leslie et al. forthcoming).

	Harm intentional	Harm not intentional
Help intentional	16	0
Help not intentional	14	14

Table 1: Participant responses to both Harm & Help cases (N=44).

To appreciate the character of the correlation, it's important to emphasize the fact that there weren't any order effects. One might have expected (as we did) that subjects would try to make their answer to the second question consistent with the answer they gave to the first question. But that didn't seem to happen. If people had behaved in this way, we should have seen differences between the group who got Harm first and the group who got Help first. But as figure 1 reveals, there were no such differences. Since subjects apparently don't change their answers significantly depending on the order in which they receive the questions, the correlation in their responses can't be dismissed as the consequence of which question they received first. The correlation seems to be a stable feature of the individuals at the time.

7. Individual Differences and Interpretive Diversity

How are we to explain the minority responses in Harm and Help? In section 5, we noted that one option is to treat the minority responses as the product of random performance factors. But in light of the correlation we found in our experiment, this seems implausible. Far from random, the minority responses turn out to be strikingly systematic.

Before we continue, we'd like to pause a moment to consider Keith Stanovich's program of research on individual differences in reasoning. Famously, people do very badly on a battery of reasoning tasks. In the Wason-selection task, people make apparently simple errors in evaluating conditionals (Wason 1966); in syllogistic reasoning, people have difficulty setting aside the believability of a conclusion when evaluating the validity of an argument (Evans et al. 1983); and in statistical reasoning, people tend to neglect representative statistical evidence in favor of the testimony of a single friend (e.g. Fong, Krantz, & Nisbett, 1986). But in each of these experiments, there is typically a smallish minority that answers correctly (i.e., as logicians or statisticians recommend). Should the minority responses be regarded as the product of 'random performance factors'? Stanovich undertook to explore this question,

and he found that the minority responses were quite clearly *not* random noise. On the contrary, he found a strong correlation between success on one reasoning task and success on other reasoning tasks (see, e.g., Stanovich, 1999, p. 35). For instance, people who perform well at evaluating syllogisms also tend to perform well on statistical reasoning and the Wason-selection task. This seems to be a stable and rather interesting feature of the individuals.[12]

In the present case, obviously the project of exploring individual differences in intuitions is only in its infancy. But Stanovich's program provides an important model and precedent. In our experiment, we found pronounced systematic individual differences. The minority who said that the CEO didn't intentionally harm the environment in Harm *also* said that the CEO didn't intentionally help the environment in Help. And the minority who said that the CEO intentionally helped the environment in Help also said that the CEO intentionally harmed the environment in Harm. What do we say in light of the consistent responses of these two minorities? Our hypothesis is that 'intentional' exhibits *interpretive diversity,* i.e., it admits of different interpretations. Part of the population, when given these sorts of cases, interpret 'intentional' one way; and part of the population interpret it in another way. On one interpretation both cases are intentional and on the other interpretation, neither is. Linguists and philosophers of language distinguish several ways in which a term can admit of different interpretations: the term might be ambiguous, polysemous, or exhibit some other form of semantic underspecification. We mean for the interpretive diversity hypothesis to be neutral about which form of interpretive diversity holds for 'intentional'. Settling those matters is important, but it's also difficult and exceeds our present ambitions. What is crucial for us is the claim that 'intentional' has multiple interpretations.[13]

The primary evidence for interpretive diversity is the discovery that the minorities are adopting a consistent (but systematically different) pattern of responses for 'intentional'. As noted in section 5, the minority responses are a central problem for Knobe's proposal. Further, although it's gone unremarked in the literature, minority responses also pose a problem for the psychological bias and Gricean accounts. For the bias and Gricean explanations, the problem arises primarily from the minority responses to the Help scenario. The general strategy those accounts take is to maintain that the correct response to both Harm and Help is that the the CEO did not intentionally produce the outcomes, and the majority

[12] Stanovich also found that these subjects tend to have higher SAT scores (Stanovich, 1999, p. 41).

[13] In Stanovich's work, the standard picture is that some subjects get the right answer and the rest just get it wrong. We do not want to extend this feature into our framework. Rather than say that some subjects get the 'right' interpretation and the other are mistaken, we are inclined to be charitable and say that different subjects interpret the word differently, but that even minority subjects are not systematically mistaken in their interpretations.

responses to Harm are somehow a performance error. But even if we leave that claim unchallenged, those accounts owe some explanation for the minority response to Help. Why, on those accounts, does anyone say that the CEO intentionally helped the environment? There's no unwanted implicature to avoid, and there's no blame or affective force pushing in favor of the judgment. Yet, in each experiment, a significant portion of the sample says that the agent intentionally helped the environment in Help. On the bias and Gricean accounts, the most obvious explanation for the minority response to Help is that it's a result of random performance factors. But in light of our experiment, this response looks less plausible. For the minority who say that the CEO intentionally helped the environment consistently apply this usage to the harm case as well. Thus, out of the accounts that we have considered here, only the interpretive diversity hypothesis can easily accommodate the minority responses.[14]

8. Listen to the People

Now that we have the interpretive diversity hypothesis on the table, we think it's worth looking back to the lay explanations for the responses that we recounted in section 2. As noted there, the lay explanations don't yield a consistent interpretation of 'intentional' that accounts for the asymmetric responses to Harm and Help. Hence people have tended to assume that at least one class of the lay explanations is wrong. Now that the interpretive diversity option is on the table, however, a much more charitable view emerges. People are in fact giving the right explanations for their answers. It's just that 'intentional' admits of different interpretations, so their explanations reflect the interpretation that they have assigned to the term in the context.

Consider again the kinds of explanations we get from the majorities in Harm and Help. Most people given the Harm case explain that they said that the CEO intentionally harmed the environment because he *knew* what would happen. Most people given Help explain that they said that the CEO didn't intentionally help the environment because his *motive* was to make money. If we take the majorities at their word, the most natural hypothesis is precisely the interpretive diversity hypothesis. 'Intentional' gets interpreted differently in Harm and Help. We will not attempt to give precise characterizations of the two interpretations. But the lay explanations suggest that something like *foreknowledge* and *motive* are at

[14] There are different ways the interpretive diversity hypothesis might play out. One possibility is that part of the population always interprets the word one way and another part always interprets it another way. Another possibility, though, is that virtually everyone has both interpretations available to them, but that some people tend, in certain circumstances, to interpret it one way or another. We know of no evidence to decide the matter.

the heart of the two different interpretations of 'intentional'. Henceforth we will refer to these two interpretations as 'foreknowledge' and 'motive', though of course that is a very rough approximation.[15]

Let's now listen to what the minorities in Help and Harm have to say for themselves. Notice that the interpretive diversity hypothesis makes a prediction about explanations from the minorities. The explanations given by people responding with the minority to Harm should tend to fit with the explanations given by the majority to Help, and vice versa. So what do people in the minority say when they explain their answers? Of the 39 undergraduates who received the harm case in the explanation task (section 2), 10 said that the CEO did not intentionally harm the environment. As predicted by the interpretive diversity hypothesis, the minority who said that the chairman *didn't* intentionally harm the environment typically invoked 'motive' or 'intention' in their explanations for why the CEO didn't intentionally harm the environment. In fact, all 10 of these subjects invoked motive or intention in explaining why they said the CEO didn't intentionally harm the environment:

> 'The chairman did not intentionally harm the environment because his motivating desire was to make money'
>
> 'He wasn't trying to harm the environment, he was trying to be profitable'
>
> 'His initial intention was to make money not to harm the environment'
>
> 'He did not set out to harm the environment, he set out to gain a profit'
>
> 'His original intent is not to harm the environment'
>
> 'He didn't intentionally harm the environment because his intention was to make as much profit as he could'
>
> 'The chairman's goal was not to harm the environment but to make a profit'
>
> 'His motivation was not to harm the environment but to make a profit'
>
> 'The chairman just wanted to make a profit'
>
> 'His goal was not to harm the environment but rather to make more money'

When we turn to the other minority, of the 46 undergraduates who received the help case, 17 said that the CEO *did* intentionally help. For this minority most of the explanations were consistent with the hypothesis that the subjects interpreted 'intentional' as foreknowledge. And nearly half of the explanations (8 out of the 17) explicitly appealed to foreknowledge:

> 'while … it was not his intention to help the environment he knowingly pursued a course which did'

[15] It's also possible that there are many different interpretations of 'intentional'. We are content to argue for two.

'he knew that if he implemented the program that both profits would increase and the environment…'

'he knew a certain action would help the environment, and he acted in a way which brought about that end result. It may not have been the manifest function of his actions but they were a result of his deliberate action'

'because he was aware…'

'he was initially informed about how the new business plan would help the environment'

'even though he didn't care he knew it would help… so it is intentional'

'He knew that the new program would help the environment, he intentionally chose the program knowing it would help the environment, whether he cared or not is another matter'

'he knowingly decided to do the program even when he knew it would help the environment'

Thus, the minority explanations conform to the prediction of the interpretive diversity hypothesis. The minority in Harm give the same kind of explanations as the majority in Help, and most of the minority in Help give the same kind of explanation as the majority in Harm. The minority explanations also help to shore up and make plausible the correlation we found in our experiment. The reason the minorities respond consistently is because one minority interprets 'intentional' as *motive*, the other interprets it as *foreknowledge*.[16]

9. Discussion

The data provide support for the view that there are two different interpretations of 'intentional action' available, one based on foreknowledge and the other on motive. In the CEO cases, we found that one minority seems to consistently interpret 'intentional' as *foreknowledge*, and another minority seems to consistently interpret it as *motive*. But the alert reader might have noticed that we've said nothing about the other subjects -- the minority that reported asymmetric intuitions. Does this group have yet a third interpretation for 'intentional'? Perhaps, but in light of the interpretive diversity hypothesis another possibility opens up: some people are *flexible* in how they interpret 'intentional' in these contexts.

[16] We should note that as far as the individual differences go, this contributes only a tiny step. We've found that the minorities tend to be systematic in their interpretations of Harm and Help. What we don't know is why some people tend to be systematic in one direction ('motive') and others in the other direction ('foreknowledge'). There are several interesting avenues to pursue. For example, the individual differences might correlate with religious upbringing, academic affiliation (humanities vs. sciences), analytic skills, or having cranky grammar teachers. We hope that future research will illuminate the matter.

If some people are flexible in interpreting 'intentional', what leads them to flex one way rather than another? On this point we are flexible ourselves. Perhaps the blame-driven hypothesis or the Gricean proposal could be recruited to explain the behavior of flexible subjects. Or perhaps something somewhat closer to Knobe's view can be recruited here. One of the important ideas in Knobe's work is that attributions of 'intentional' are affected by the different ways in which we evaluate praise and blame (Knobe, forthcoming a). In particular, if we are trying to decide whether someone is accountable for some outcome, either foreknowledge or motive is sufficient for the person to be accountable. Thus, when flexible subjects are trying to determine whether someone is accountable, if only one feature (motive or foreknowledge) is present, flexible subjects might tend to fix that as the most relevant interpretation of 'intentional'. By contrast, if they are trying to decide whether someone deserves praise, foreknowledge is not sufficient for the person to be creditable, but motive is. As a result, flexible subjects might tend to interpret "intentional" as "had the motive" in these cases.[17]

Although our account can incorporate some aspects of Knobe's view, there remains a key difference. Knobe maintains that the role of outcomes is built into the concept *intentional* itself, as a result of which the concept applies differently to Harm and Help. By contrast, we think that considerations of outcome might influence *which interpretation* the term is given, at least for some people. The difference can perhaps be made more perspicuous if we rephrase our proposal in terms of a mapping between words and concepts, since this is how Knobe presents his own view. We have argued that the interpretive diversity hypothesis provides the best explanation of the minority responses to Harm and Help. If that's right, then we already have good reason to think that there are two interpretations of 'intentional'. One way to render this idea is to maintain that 'intentional' gets mapped to two different concepts, (again very roughly) *foreknowledge* and *motive*. There is independent reason to think it's a common phenomenon that a natural language word gets mapped to multiple concepts, depending on the context (see e.g., Machery, forthcoming; Sperber & Wilson, 1998). In light of this, it strikes us that the most parsimonious explanation of the responses of subjects who respond asymmetrically to Harm and Help is that they are mapping the word 'intentional' onto different concepts in the different contexts. We are open-minded about exactly why the flexible subjects respond asymmetrically, but we find it more plausible to explain the phenomenon by appealing to a complex relation between words and concepts than

[17] There are important and interesting questions about the processes that might be implicated in this kind of interpretive practice (for one approach, see Wilson, 2003; Wilson & Sperber, 2004). There are also important questions about the nature of the concepts that get elicited under these different interpretations (see e.g. Carston, 2002). But discussing those questions would take us too far afield given our central aim for this paper.

by accepting Knobe's idea that the asymmetric responses flow from the complexity built into the concept of *intentional* itself.[18]

There is one more technical point to chart. Although the preceding paragraph speaks of a mapping between one word, 'intentional', and two concepts, we do not want to suggest that we have locked on to a general theory of concept individuation. So, while we find it easiest to think of the proposal by invoking two separate concepts, we are not strongly opposed to the idea that *intentional* is a single concept with dissociable components corresponding to *foreknowledge* and *motive*. In some contexts, subjects map 'intentional' onto the *foreknowledge* component of the concept, and in other contexts, subjects map 'intentional' onto the *motive* component of the concept. We regard such a dissociable-component proposal as a friendly variant of our own proposal. However, the dissociable-component account would no longer be a friendly variant of our account if the theorist goes on to claim that the minority responses *mistakenly* focus on the wrong components. If a dissociable-component theorist claims that the minorities in Harm and Help are making a mistake, then that really is a different proposal from ours. But such a proposal would face the challenging task of arguing that the minorities don't know the meanings of their own words, despite their consistent responses and explanations. This is not a challenge that our view, or its dissociable-component variant, faces.

10. Implications

If the interpretive diversity hypothesis is right, it helps to explain the puzzling pattern of results on intentional action. We can acknowledge the robustness of Knobe's findings without adopting the radical theory that the very concept of *intentional* implicates moral considerations. However, we think that the individual differences revealed by Knobe's vignettes also have broader implications, which we will briefly discuss here.

As noted in section 1, cultural differences in intuitions about philosophical cases pose an important objection to *a priori* philosophical analysis. However, one plausible reply to this objection is to admit that the pursuit of a philosophical analysis is culturally local. The existence of individual differences poses a rather different, perhaps more trenchant problem. In the literature on cross-cultural differences in intuitions, it has been remarked that there are also *intra*-cultural differences in intuitions about philosophically important concepts (e.g. Machery et al. 2004, p. B8). However, these previous studies did not investigate directly the individual differences, and so it's impossible to know whether the

[18] There are, we think, interesting considerations on both sides of this issue. But this is yet another issue that will have to wait for another occasion.

intra-cultural differences were merely the result of random performance factors.[19] In this paper, we have sought to explore the individual differences more systematically, and the findings raise a significant worry about even the more modest project of using *a priori* methods to discern the concepts that we have in our culture. For the results from work on the Knobe effect suggest that within our own culture, indeed within the halls of Western academia, there are robust and stable individual differences in intuitions about philosophically important concepts.

The problem for *a priori* approaches is that the Knobe effect (on our interpretation) reveals individual differences in intuitions that are of considerable importance to assessing philosophical disputes, but these differences cannot be detected from the armchair. Consider (i) the (large) minority that says both CEOs intentionally produced the side effect and (ii) the other (large) minority that says that neither CEO intentionally produced the side effect. Obviously there is a major obstacle to productive discussion between these parties on the issue of whether the CEO *intentionally* harmed the environment. For (on our account) the groups are interpreting the term in systematically different ways. But this fact could not have been made apparent without actually doing the experiments. Indeed, it seemed so clear to us that the right interpretation of 'intentional' required 'motive' that it seemed perfectly obvious to us that the minority who said that the CEO intentionally helped the environment were just making a mistake. It was only after we consulted the results of our experiment that it occurred to us that each minority might be responding in a consistent way to Knobe's scenarios. We couldn't tell from our armchairs that in cases like Harm and Help, the word 'intentional' gets interpreted differently by different people. And without knowing such basic facts about how people interpret the word, it's unlikely that the discussion between the two groups will be fully satisfactory. It's more likely that the two groups, when pitted in philosophical dispute over the category of *intentional action,* will simply talk past each other.

To make this more concrete, consider the philosophical dispute over the 'simple view' of intentional action. According to the simple view, if S intentionally did A, then S intended to do A. Some philosophers (e.g. Adams, 1986; McCann, 1986) defend the simple view, which commits them to denying that the CEO intentionally harmed the environment; other philosophers (e.g. Harman, 1976) reject the simple view and would presumably allow that the CEO intentionally harmed the environment. Now, in this dispute over the proper analysis of 'intentional' who is right? One conciliatory answer is: Both are right since each is applying a different interpretation to the term. Simple view enthusiasts are interpreting 'intentional' as *intended to*; simple view opponents are interpreting 'intentional' as *foreknowledge*. Hence on these interpretations the one group rightly denies that the CEO intentionally harmed the

[19] Of course, the existence of random performance factors need not undermine the cross-cultural effects. For in these experiments there are reliable differences in the responses given in different cultures.

environment and the other group right affirms that he did intentionally harm it. A less conciliatory answer is that both groups are wrong to be fighting over it. For there's no single unified interpretation that will be revealed by the fight. They are talking past each other.[20]

More generally, the fact that there are robust individual differences in intuitions about philosophically important concepts provides further evidence of the importance of empirical work on intuitions. If a philosophical dispute is driven partly by systematic individual differences in intuitions, proper evaluation of the dispute demands that we know this. The Knobe effect provides an existence proof that interpretive diversity can be at the root of philosophical troubles. It's an open question whether traditional philosophical disputes about matters such as free will and personal identity grow out of systematic individual differences in intuitions. Frankly, we hope not. But the Knobe effect has chastened us against blithely assuming that our own confident intuitions are widely shared by others in our culture or even in our profession.

Department of Philosophy
University of Arizona

Department of Political Science and Philosophy
Weber State University

[20] Several people have suggested that the philosophical dispute over the "simple view" concerns how to attain reflective equilibrium about what it is for an action to be *intentional*. But this is quite consistent with our hypothesis that individual differences in first-blush intuitions play an important role in generating and sustaining the dispute. Moreover, we think that the goal of reflective equilibrium will be advanced if disputants recognize the individual differences in first-blush intuitions.

References:

Adams, F. 1986: Intention and intentional action: the simple view. *Mind & Language*, 281-301.

Adams, F. and Steadman, A. 2004a: Intentional action in ordinary language: core concept or pragmatic understanding? *Analysis*, 64: 173-181.

Adams, F. and Steadman, A. 2004b: Intentional action and moral considerations: still pragmatic. *Analysis*, 64: 264-267.

Adams, F. and Steadman, A. forthcoming: Folk concepts, surveys, and intentional action.

Carston, R. 2002: *Thoughts and Utterances: The Pragmatics of Explicit Communication.* Oxford: Blackwell.

Evans, J. St. B. T., Barston, J., & Pollard, P. 1983: On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11, 295-306.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. 1986: The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, 18, 253-292.

Grice, P. 1989: *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Harman, G. 1976: Practical reasoning. *Review of Metaphysics* 79: 431-463.

Hauser, M. 2006: *Moral Minds: The Unconscious Voice of Right and Wrong.* NY: Harper Collins.

Jackson, F. 1998: *From Metaphysics to Ethics: a Defence of Conceptual Analysis.* Oxford: Oxford University Press.

Jackson, F. 2001: 'Responses.' *Philosophy and Phenomenological Research* 62(3): 653-664.

Kahneman, D., Slovic, P. and Tversky, A. (eds.) (1982): *Judgment under Uncertainty*. Cambridge: Cambridge University Press.

Knobe, J. forthcoming a: The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies.*

Knobe, J. forthcoming b: Reason explanation in folk psychology. *Midwest Studies in Philosophy.*

Knobe, J. 2005: Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, 9, 357-9.

Knobe, J. 2003a: Intentional action and side effects in ordinary language. *Analysis*. 63, 190-193.

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

Knobe, J. 2004: Intention, intentional action and moral considerations. *Analysis* 64: 181-187.

Knobe, J. and Burra, A. forthcoming: Intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition.*

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

McCann, H. 1986: Rationality and the range of intention. *Midwest Studies in Philosophy*, 191-211.

McCann, H. forthcoming: Intentional Action and Intending: Recent Empirical Studies. *Philosophical Psychology,* 5.

Machery, E. forthcoming. Concepts are not a natural kind. *Philosophy of Science.*

Machery, E., Mallon, R., Nichols, S., and Stich, S. 2004. Semantics, Cross-Cultural Style. *Cognition,* 92, B1-B12.

Malle, B. F. forthcoming: The moral dimension of people's intentionality judgments. *Journal of Culture and Cognition*.

Malle, B. and Knobe, J. 1997. The folk concept of intentionality. *Journal of Experimental Social Psychology* 33: 101-121.

Malle, B. F. and Nelson, S. E. 2003: Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law*. 21, 563-580.

Mele, A. 2003: Intentional action: Controversies, data, and core hypotheses. *Philosophical Psychology,* 16, 325-340.

Nadelhoffer, T. 2003. The Butler Problem revisited. *Analysis.*

Nadelhoffer, T. forthcoming a: Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations.*

Nadelhoffer, T. forthcoming b: On saving the simple view. *Mind & Language.*

Sousa, P. 2004: Unpublished data. University of Michigan.

Sperber, D. & Wilson, D. 1998. The mapping between the mental and the public lexicon. In P. Carruthers & J. Boucher (eds) *Language and Thought*. Cambridge University Press, Cambridge: 184-200.

Stanovich, K. 1999: *Who Is Rational?* Hillsdale, NJ: LEA.

Stich, S. 1992. What is a theory of mental representation? *Mind*, 101, pp. 243-61

Stich, S. and J. Weinberg 2001. Jackson's Empirical Assumptions, *Philosophy & Phenomenological Research*, 62.

Wason, P. 1966: Reasoning. In B. Foss (ed.), *New Horizons in Psychology*. Harmonsworth, England: Penguin, 135-151.

Weinberg, J. Nichols, S. and Stich, S. 2001: Normativity and epistemic intuitions, *Philosophical Topics*, 29, 429-460.

Wilson, D. 2003. Relevance theory and lexical pragmatics. *Italian Journal of Linguistics/Rivista di Linguistica*, 15(2): 273-291. Special Issue on Pragmatics and the Lexicon.

Wilson, D. & Sperber, D. 2004. Relevance theory. In G. Ward and L. Horn (eds.) *Handbook of Pragmatics*. Oxford: Blackwell, 607-632.

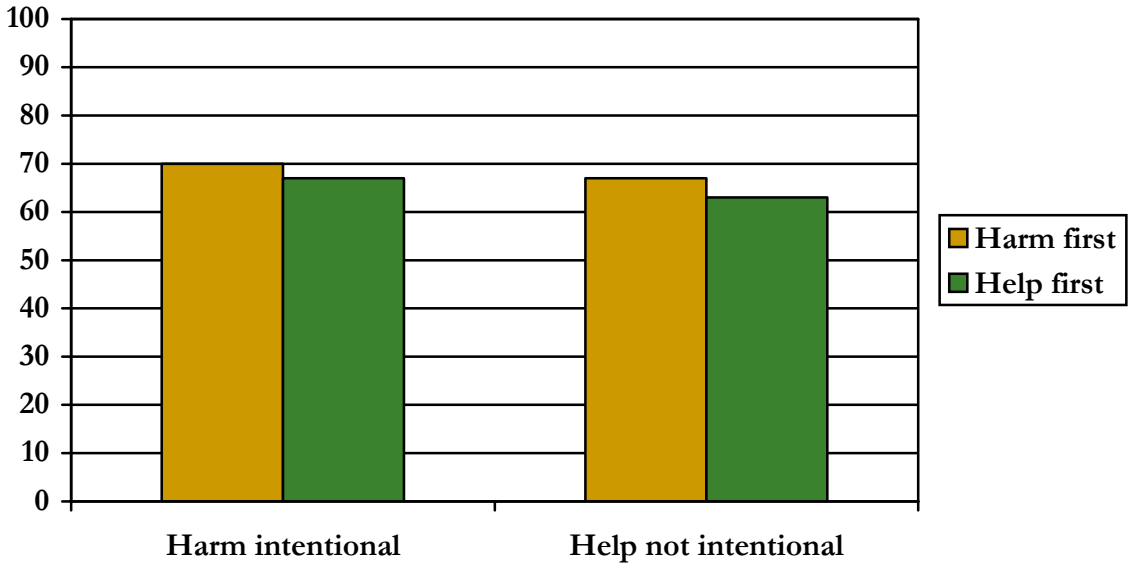

Chart 1: The absence of order effects